2004-01-30

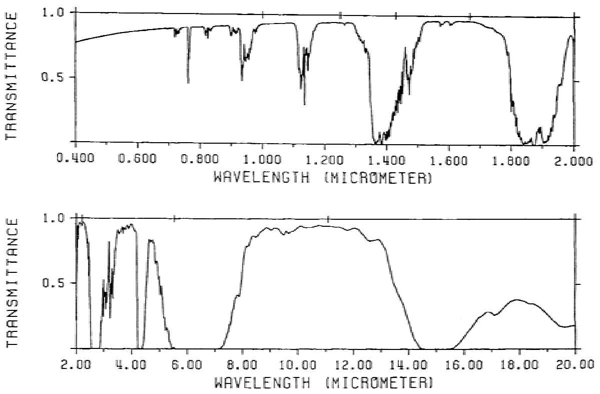
COMBAT SYSTEMS Volume 1. Sensor Elements Part I. Sensor Functional Characteristics

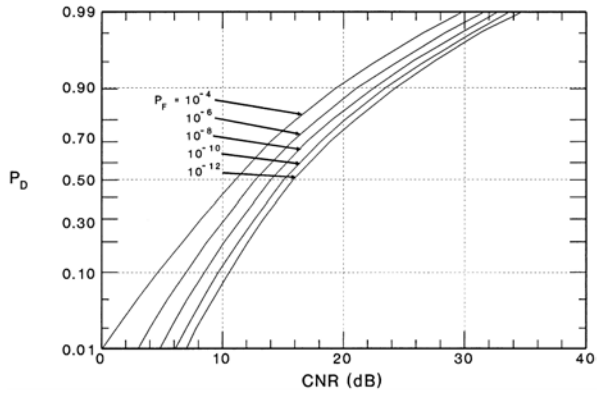
Harney, Robert C.

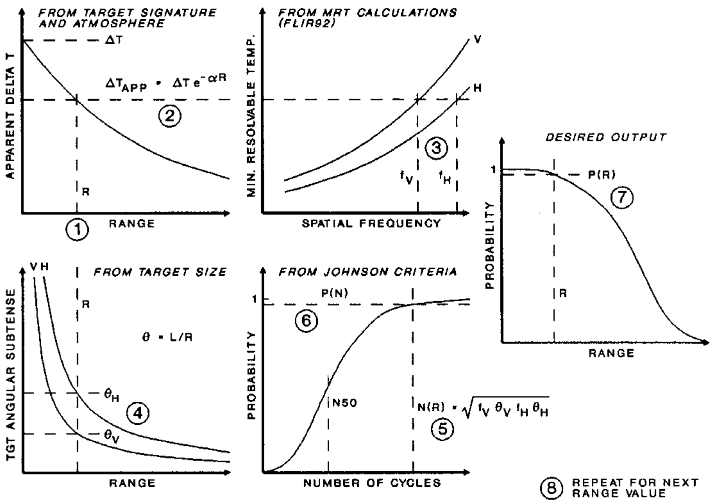
https://hdl.handle.net/10945/69347

COMBAT SYSTEMS

Volume 1. Sensor Elements
Part I. Sensor Functional Characteristics







Robert C. Harney, Ph.D.

30 January 2004
Robert C. Harney
Monterey, California

VOLUME 1. SENSOR ELEMENTS.
PART I. SENSOR FUNCTIONAL CHARACTERISTICS

Note: Boldface entries on the following pages are not yet completed.

VOLUME 2. SENSOR ELEMENTS.
PART II. SENSOR TECHNOLOGIES

VOLUME 3. ENGAGEMENT ELEMENTS.
PARTS I & II. CONVENTIONAL WEAPONS

VOLUME 4. ENGAGEMENT ELEMENTS.
PART III. UNCONVENTIONAL WEAPONS

VOLUME 5. CONTROL ELEMENTS & PLATFORM INTEGRATION

PART III. TOPSIDE DESIGN

FOREWORD

INTRODUCTION TO "COMBAT SYSTEMS"

Combat systems are the integrated systems that give modern military units their enormous warfighting (and peacekeeping) potential. In the current context, combat systems include almost everything on board modern platforms (ships, submarines, ground vehicles, aircraft, or spacecraft are platforms) except those related to structure, mobility, and habitability (if the platform is manned). The elements that make up modern combat systems include the sensor systems that detect and track the targets, the weapons that engage the targets, the command & control systems that authorize or assist the engagement of the targets, and anything else needed to integrate the elements into a platform and permit the operator to use and control them.

This set of books evolved from notes used in two one-quarter courses to teach the elements of combat systems and systems integration to military officers in the Total Ship Systems Engineering (TSSE) program at the Naval Postgraduate School (NPS) in Monterey, California. Students in this multidisciplinary program are drawn from three curricula at NPS: Naval & Mechanical Engineering, Electrical Engineering, and Combat Systems Science & Technology. The students take six elective courses (1 on systems engineering, 1 on shipboard power systems, 1 on naval architecture, 1 on hull, mechanical, & electrical integration, and 2 on combat systems elements and integration) and perform a six-month capstone ship design project in addition to fulfilling all requirement for M. S. degrees in Mechanical Engineering, Electrical Engineering, or Applied Physics, respectively. The students are typically junior naval officers (mostly O-3 – Lieutenants) who have served at least two tours aboard ship and have been given the opportunity to obtain graduate technical educations prior to their next promotions. The majority of these are Engineering Duty Officers, although a number have been Unrestricted Line Officers. However, the classes have also had a number of Coast Guard officers, and even a few Marine Corps and Army officers and Department of Defense civilians. Although most of the students were interested in Naval surface ships and came from the Surface Warfare community, a number were aviators, submariners, space systems operators, and even armor officers. This fact coupled with the author's belief in the necessity of joint warfare led him to tailor the course to avoid restriction of the discussions to "Surface Navy only" systems. A Systems Engineering and Analysis (SEA) program was developed a few years ago to teach technology to future requirements setters in the Unrestricted Line Community. The two TSSE courses on combat systems were expanded to four courses in the SEA curriculum, necessitating the incorporation of more material into the texts.

In recent years a small number of foreign military officers have been allowed to enter the program. This required that the course materials be unclassified and available for unlimited distribution. This was facilitated by taking a first principles approach to combat systems rather than dissecting existing systems. The laws of nature are open to all who observe them and there is little in any generic sensor or weapon that can be classified. By being forced to avoid discussing details of existing hardware, we gained the advantage of avoiding a focus on the present. We can address the basic principles that govern the performance of all systems: old, new, and yet-to-be-developed.

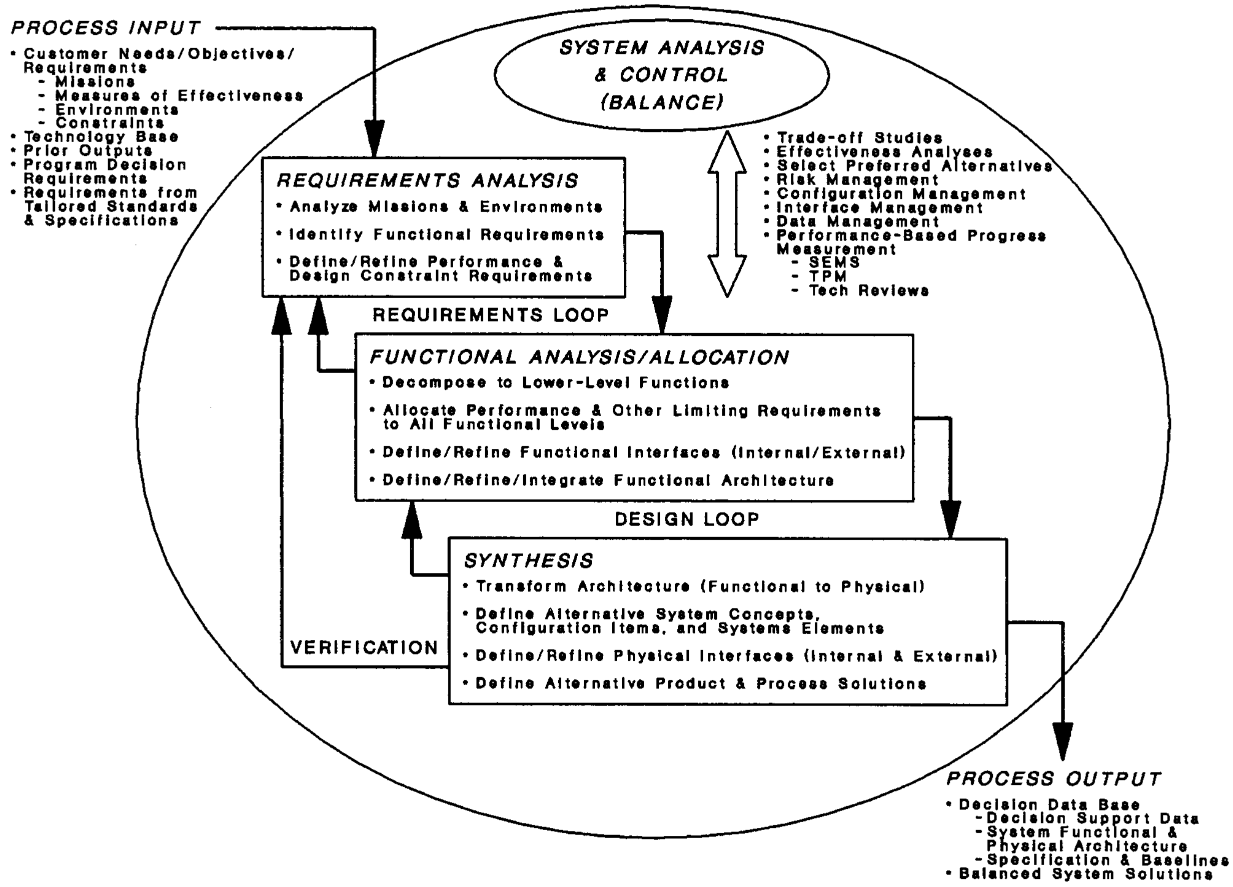
Another consideration in the design of the courses was the diversity of the students. Because the students come from vastly different undergraduate backgrounds, the only commonality that could be assumed was the curriculum requirements of having taken one-year of calculus and one-year of calculus-based physics at the undergraduate level. However, these courses have typically been taken by the students at least five to eight years prior to their beginning the TSSE program and the skills taught in those classes have seldom been required to perform their intervening military jobs. This meant that all necessary technical background had to be contained within the courses themselves. Another consideration was the fact that the courses were intended to educate systems engineers in sufficient detail to be able to interact with element designers, not to educate element designers in the intricacies of their craft. However, based on the author's fifteen years as an element designer followed by fifteen more years as a systems engineer (element integrator), it was decided that excess design-level information was preferable to insufficient design-level information. Shortcomings in the author's own breadth of knowledge will guarantee less than complete detail in more a than a few of the topics to be covered. Nevertheless, even experienced element designers may find organizing principles and a philosophical framework connecting their specialty to other fields in ways that they may not have consciously recognized from their own experience.

This material is taught as part of a broader systems engineering-oriented curriculum. Although this is not a firm requirement – the material contained in these volumes is relatively self-contained – the author feels that a systems engineering approach to the subject of combat systems is essential for successful utilization of the principles presented here. In both of the curricula currently using this material, the students take at least one course in systems engineering principles and processes. This course may be taken prior to or concurrent with the course that covers the material in the first two volumes of this text. The systems engineering process that is taught is captured in Figure F-1. To facilitate internalization of this overall process, the combat systems courses attempt to emphasize those aspects that might impact requirements, functional capabilities, concept synthesis, and performance analysis.

The TSSE program was established because the Navy realized that combat systems and ship hull, mechanical, and electrical (HM&E) design had become stovepiped with combat systems and HM&E engineering being in separate stovepipes. Requirements entered the bottom of the stovepipes and systems came out the top. However, there was no interconnections along the length of the stovepipes. The top was the first time that combat systems interacted with HM&E. As a result, there were interface problems and inefficiencies. TSSE was intended to provide cross channels between the stovepipes to prevent the standard problems that arose. TSSE appears to have been very successful in doing this.

However, the author has long recognized that it is not just combat systems and HM&E that suffer from suffer from stovepipe problems. Sensors, weapons, and command & control have there own stovepipes. Each of these is composed of other stovepipes and few have any significant interactions. In sensors, radar is taught as a completely different subject area from sonar. Radar engineers seldom know much of anything about sonar and vice versa. And both radar and sonar are taught as a completely different subject areas from electro-optical sensors and field sensors. The author recognizes that this is not only unnecessary, but also counter-productive, at any level, but

Figure F-1. The systems engineering process.


PROCESS INPUT
• Customer Needs/Objectives/ Requirements
- Missions
- Measures of Effectiveness
- Environments
- Constraints
• Technology Base
• Prior Outputs
• Program Decision Requirements
• Requirements from Tailored Standards & Specifications
SYSTEM ANALYSIS & CONTROL (BALANCE)
• Trade-off Studies
• Effectiveness Analyses
• Select Preferred Alternatives
• Risk Management
• Configuration Management
• Interface Management
• Data Management
• Performance-Based Progress Measurement
- SEMS
- TPM
- Tech Reviews
REQUIREMENTS ANALYSIS
• Analyze Missions & Environments
• Identify Functional Requirements
• Define/Refine Performance & Design Constraint Requirements
REQUIREMENTS LOOP
FUNCTIONAL ANALYSIS/ALLOCATION
• Decompose to Lower-Level Functions
• Allocate Performance & Other Limiting Requirements to All Functional Levels
• Define/Refine Functional Interfaces (Internal/External)
• Define/Refine/Integrate Functional Architecture
DESIGN LOOP
SYNTHESIS
• Transform Architecture (Functional to Physical)
• Define Alternative System Concepts, Configuration Items, and Systems Elements
• Define/Refine Physical Interfaces (Internal & External)
• Define Alternative Product & Process Solutions
VERIFICATION
PROCESS OUTPUT
• Decision Data Base
-Decision Support Data
-System Functional & Physical Architecture
-Specification & Baselines
• Balanced System Solutions

especially at the systems engineering level. Combat systems should not be taught or practiced as a set of stovepipes, but as an integrated discipline. The key is in recognizing the many common aspects of these systems elements and teaching them once, and then pointing out the individual areas where these common elements may be applied, and providing the necessary language to describe and use these common elements in the same fashion as taught individually by the stovepiped disciplines. This will be illustrated further as we proceed through the foreward.

Breadth and depth are required to be an outstanding systems engineer – this is also true of a combat systems engineer. Systems engineers need to be able to perform trades between complex alternatives. This requires at a minimum the ability to perform "back of the envelope" calculations of sensor, weapon, and communication performance. Since few teams can afford to have dedicated engineers on-hand to cover every possible specialty discipline, systems engineers are often called to "fill-in" and address whatever tasks cannot be accommodated by the rest of the team. In his career as a systems engineer for a defense contractor, the author was at different times required to perform reliability predictions, nuclear hardness analysis, maintainability assessments, threat assessments, operations analyses, fault analyses, safety assessments, development of pre-planned product improvement strategies, manpower planning, establishment of embedded training requirements, and conceptual designs of weapon components outside his area of expertise. The author's original specialty was lasers and their applications – yet he became involved in designing fuzes, propulsion schemes, signal processing techniques, seekers, warheads, launchers, and power supplies, among others, in addition to all of the normal activities of systems engineering. The requirement of breadth and depth is further supported both by the need of systems engineers to know the value and potential contributions of every specialty discipline involved in complex design tasks (to insure that they are considered equally in the overall design) as well as the need to know enough of the jargons of those specialty disciplines to be able to communicate as near peers with the specialty experts. The system engineer is the final technical arbiter of disputes between specialty disciplines. Specialty engineers are much more likely to accept such arbitration if they feel the arbiter understands their point of view and their problems and is therefore capable of making enlightened technical decisions. It is the author's belief that one cannot be a systems engineer without broad and deep knowledge of the domain in which that systems engineering is being carried out.

Performing trade studies of widely differing technical solutions and synthesizing truly innovative concepts requires the systems engineer to collect and evaluate data from highly diverse sources. It is quite common for the multiple authors of that data to have used widely differing terminology to describe basically the same thing. For example, in the collection of microwave attenuation data one group may quote their results in terms of loss tangents, another in terms of complex refractive indices, and a third in terms of extinction coefficients. A "**Rosetta Stone**" is needed to help translation of the different terminology into a common language. Good systems engineers have created, inherited, or otherwise acquired a number of such Rosetta Stones during their career. In the present text, the author makes every attempt to present the important Rosetta stones that he has acquired or developed.

Recognizing that there is a common foundation to diverse technologies and that **common formalisms** can be used to analyze seemingly different problems aids in recognizing the need for using a Rosetta Stone when available. In the sensor sections, the author has explicitly tried to define and present the minimum number of methodologies needed to address all known and hopefully knowable sensor technologies. Once the methodologies are mastered and applied to one technology, the student is 90% of the way to understanding all "**analogs**" of that technology. For example, once one has learned basic detection theory, estimation theory, and the essentials of propagation (attenuation, diffraction, refraction, scattering, and interference), and has applied them to developing an understanding of radar, there is no longer any need to spend comparable effort in developing an understanding of sonar. Virtually every phenomenon affecting radar has a counterpart affecting sonar. An understanding of "radar" coupled with learning the domain specific terminology of sonar, permits an equivalent understanding of "sonar". The converse is also true. Since argument by analogy is a major skill that successful systems engineers possess and use, we present the student with analogy after analogy. This is done not only to demonstrate the skill, but also to condense an otherwise enormous volume of seemingly unrelated material (and having considerable duplication in the author's view) into a manageable set of courses.

At points in the presentation the author makes strong points concerning problems that system procurers often encounter. A certain amount of cynicism may be detected here. The author has been given too many reasons not to be slightly cynical. He seen (and even worse, experienced) the cancellation of too many programs prior to their completion. In some instances, the cancellations came about because of inadequacy of the technology; in others, due to inadequacy of the program management. In still others, there were attempts to go against political expediency. The systems engineer on a program is seldom the highest authority on that program and cannot control or even influence all of these factors. However, he can help guarantee that cancellation is not due to engineering incompetence. Where there are lessons to be learned from past mistakes (and since the author considers himself to be an expert – experts are defined as those who have already made every mistake that novices are likely to make – he has made many himself) those lessons should be shared. In studying "**lessons learned**", a little cynicism is often justified.

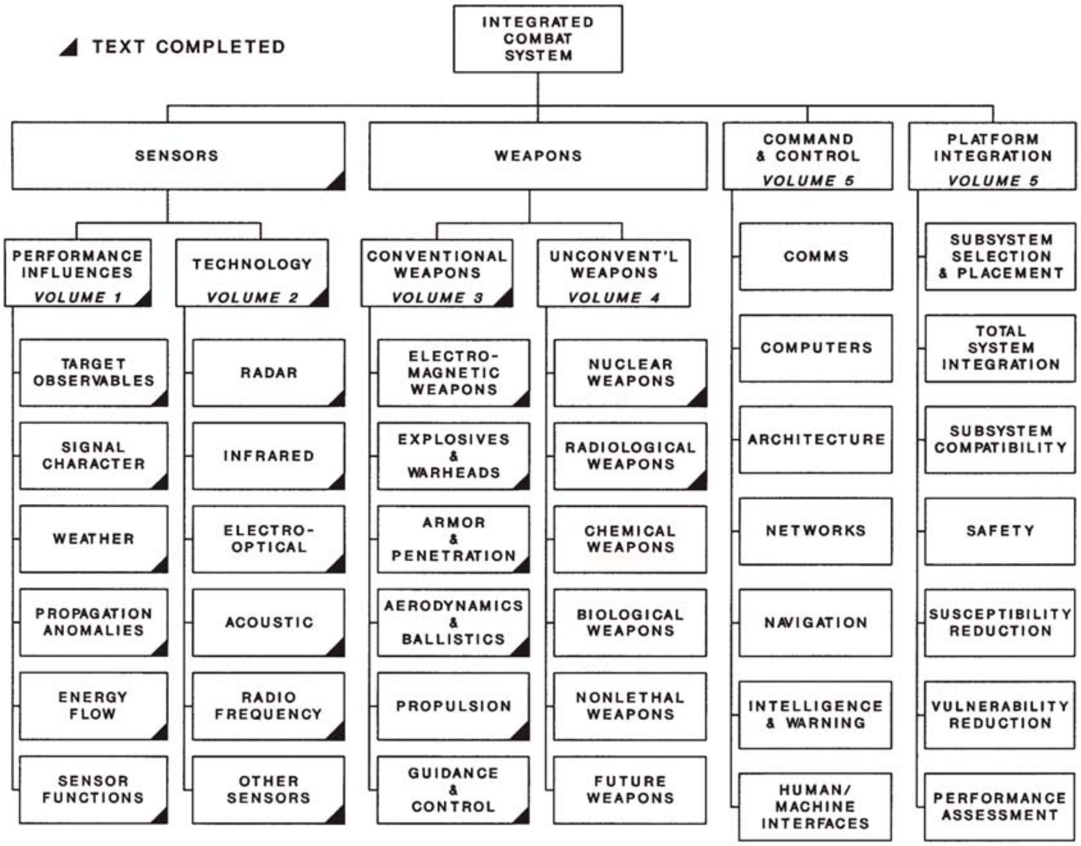
Another aspect of philosophical outlook that shows up in this book stems from the author's intermittent involvement with the "electronic warfare" community. He has been consistently impressed by the ingenuity with which countermeasures experts can find ways to negate the performance of "flawlessly designed" military systems. This is not universally true of systems designers. Many are blissfully unaware of the existence of a countermeasures community. There is also a strong tendency for design engineers to be enamored of the technologies with which they work. This leads to an unwillingness to closely examine the shortcomings of those technologies. To prevent certain disaster, the systems engineer must not also share this trait. At the same time as he is providing counsel and guidance to making a system work, the system engineer should also be thinking "**How can I break it?**". Someone working for our adversaries is certainly asking that same question. Countermeasures, vulnerability, and survivability find frequent and serious discussion throughout these volumes.

Lastly, this is not merely a textbook that condenses a well-established body of knowledge. Nor is it a handbook that collects the varied works and thoughts of others. Although it is a textbook that presents a condensed body of knowledge and it was designed to provide much of the utility of a handbook, it also contains considerable **original material** that the author has not found the time or place to publish before. Many derivations are the author's (references to the original source are given when the author is merely following a standard derivation) as are many of the tables and figures presented. Finally, the author has made an attempt to present a **coherent philosophy** and comprehensive presentation **of combat systems**. Some of the original material stems from this effort. He hopes the following pages reflect at least partial success in that attempt.

In summary, this work attempts to present a coherent approach to combat systems by making maximum use of "universal" tools and techniques. These include Rosetta stones, analogies, common formalisms, lessons learned, and "How can I break it?", combined with considerable original material to unify prior work into a comprehensive presentation of the subject.

Figure F-2 portrays the big picture of combat systems that this work is attempting to document. A fully functioning, **integrated combat system** will consist of **sensors** (often referred to as **detection elements**), **weapons** (often referred to as **engagement elements**), and **command & control systems** (often referred to as **control elements**) skillfully integrated into a **platform** in

Figure F-2. Integrated combat systems – the "Big Picture".



such a way as to optimize the performance of the whole. The term sensors covers a wide variety of devices ranging from the simple and minute (such as a simple telescope) to the enormous and complex (such as a phased array microwave radar). The systems engineer needs to be able to select from this variety, compare the performance of diverse sensors on a logical and quantifiable basis, and combine the outputs from multiple sensors to perform complex warfighting functions. Experience has taught that performing this task requires two kinds of knowledge: the factors that influence sensor performance and the technologies relevant to specific sensor systems.

Performance influences include the characteristics of targets that provide observables capable of being sensed, the fundamental characteristics of the signal (e.g., the unique characteristics of electromagnetic radiation), weather and anomalies in the propagation of the signal, the explicit flow of energy from the target to sensor taking losses and diffusion into account, and the specific functions being performed (e.g., detection or tracking). Many of these influences affect many different sensors in the same way. For example, ducting (a propagation anomaly) can occur for most sensor types including both electromagnetic and acoustic sensors. The results of detection theory can be applied equally to all sensors with a detection function. Some affect different sensors quite differently. For example, weather minimally affects many radars, dramatically affects electro-optical sensors, and only indirectly affects sonar systems. Understanding these **performance**

influences is essential to effective design and use of sensor systems. It is helpful to have this understanding before studying the sensors themselves.

Nevertheless, of equal importance are the **technologies** that are used to implement the sensor systems. One cannot make decisions about the utility or lack of utility of a radar if one doesn't know what a radar is. In military applications we must be concerned about sensors ranging from radars (microwave radars, laser radars, synthetic aperture radars, bistatic radars, over-the-horizon radars, etc.) to infrared sensors (thermal imagers, infrared search and track, reticle seekers) to electro-optical sensors (television, image intensifiers, direct vision) to acoustic sensors to radio frequency intelligence sensors to other diverse kinds of sensors (radiation monitors, star trackers, magnetic anomaly detectors, etc.). Many of the concepts used in one kind of sensor find analogs in other kinds of sensor. For example, there is no fundamental difference between phased array antennas in radar systems and phased array transducers in sonar systems. It is surprising how many instances of similar analogies can be identified.

Sensors are an essential part of a total combat system. It is difficult to employ weapons if the targets cannot be located. On the other hand, without weapons, located targets cannot be attacked. Sensors not only locate targets, they are integral to the performance of the weapons themselves. Without seekers or fuzes or navigation systems, most weapons would not be effective. This is why sensors are emphasized first. However, the systems engineer must understand all aspects of the problem. Understanding of **weapons** also requires two kinds of knowledge: the means by which weapons accomplish their function (**effectors**) and the means by which the effectors are delivered to the targets (**delivery systems**). Effectors include directed energy and the means to produce and direct it, explosives and their use in warheads, kinetic energy weapons, weapons of mass destruction (WMD – nuclear, chemical, and biological weapons), and even nonlethal weapons. Delivery systems include guns, bombs, and missiles. Delivery systems of all kinds can be understood by understanding a few basic elements. These elements include basic aerodynamics (the forces that act on bodies in motion), control systems (the way in which feedback systems respond to commands and external disturbances), guidance (determination of desired motional characteristics), actuation (transformation of desired motion into actual motional characteristics), propulsion (the means of acceleration objects to high speeds and/or maintaining those speeds), and ballistics (the motions of projectiles). Special characteristics of WMD delivery systems also need discussion.

Data must be exchanged between sensors and weapons in order to perform the desired missions. This is one of the functions of the **control elements**. **Communications** systems permit data to be exchanged between separate system elements as well as between platforms. In one sense communications links are very similar to sensor systems and their analysis shares many things in common with sensor analysis. Communication systems are often organized into **networks** to permit easy sharing of important data. **Computers** store and process data and convert data into elements of knowledge that may be used to guide future actions. Modern computer systems use a variety of specialized processing elements. These elements perform special kinds of arithmetic and memory & recall functions and are assembled and organized in many different ways to form what is commonly referred to as computers. The organization of these elements is called **computer architecture**. **Architecture** is also used to describe the way in which communications systems are organized as well as the way in which complete weapons systems are organized and interconnected. It is a critical topic to be assimilated and internalized. Complete weapons systems must also accommodate the specialized systems used for **navigation** and for **intelligence** gathering, dissemination, and analysis. Since all systems still require human beings to perform critical decision functions, it is important to understand aspects of **human/machine interfaces**. This field addresses subjects such as displays, symbology, input/output devices, and controls (switches, knobs, etc.) and how to design such interfaces to maximize information transfer and minimize operator error.

Sensors, weapons, and control elements do not of themselves constitute complete and effective combat systems. A combination of a thermal imager, a laser-designated missile, and a radio command guidance link can perform all of the functions desired of a weapon system, except that it won't work. The pieces cannot properly interact with each other. **Integration** is the key to functioning systems. Key to this task is the rational **selection of subsystems** and the intelligent **placement** of these subsystems on the platform destined to carry them. Integration is a task that must be iterative. The initial selection and placement will invariably create integration problems that must be addressed. For example, it may be necessary to physically align the reference axis of a thermal imager with that of a laser designator. Initial placement of the thermal imager high on a mast and the laser designator with the missile launcher on a deck poses **alignment** problems. The mast is a flexible structure that can move relative to the deck, the relative motion makes alignment a difficult if not insurmountable problem. It would be better to co-locate the imager and the designator. Selected systems may have electromagnetic **compatibility** problems. It is common for radars to set off electronic warfare systems and for those electronic warfare systems to attempt to jam the radars. Virtually any sensor or weapon element may cause compatibility problems or be susceptible to compatibility problems with other elements. **Safety** issues are rampant with sensors, weapons, and communications systems. Radar emissions can kill and lasers can blind; heat and blast from rocket motors are clear hazards; even communications antennas carry voltages that can shock individuals who may contact them. The **survivability** of the platform is one of the primary concerns of the combat systems engineer. Survivability is usually considered to be the sum of two parts: **susceptibility** (the ability of threat weapons to be delivered against the platform) and **vulnerability** (the ability of threat weapons to damage or destroy the platform once those weapons are delivered). Both selection and placement of combat systems elements can affect susceptibility and vulnerability. In addition, specific platform design functions that relate to survivability are usually delegated to combat systems. For example, stealth (low observables) and armor are the province of the combat systems engineer not the naval architect (ship designer). Lastly, it is essential that the entire system be capable of being evaluated before being sent into combat. **Performance assessment** is used to evaluate all of the integration decisions made by the combat systems engineer. Any decision found wanting needs to be revisited (necessitating the iterative character of integration). Development of effective integrated combat systems requires consideration of all of the elements of this "Big Picture".

This series of books is intended to capture the "Big Picture" as described above. At the time of writing this foreword to the first volume, the complete work, Combat Systems, is expected to be five volumes in length. This approach was suggested by the practical need to use the text as it was being written. The lectures notes from which the text is being generated are used for at least eight different courses at NPS (the two TSSE courses and six other sensors and weapons courses in other curricula). A single volume would have been too massive (over 2000 pages) and would have delayed availability of all of the included information for several years. The five volumes:

* Sensor Functional Characteristics
* Sensor Technologies
* Conventional Weapons
* Unconventional Weapons
* Control Elements and Platform Integration

are intended to be used in sequence, but have been made as self-contained as possible.

For example, Volume 1 discusses the nature of sensor observables, the effects of propagation through the environment, the functions that sensors perform, and the ways in which the performance of those functions are analyzed. Volume 1 presents a minimal set of tools needed to perform analyses and a theoretical framework within which any and every kind of sensor can be analyzed in a fashion that permits direct comparisons and understanding through analogies. Appendices augment the material and reduce the necessity for reference to other works in order to solve realistic problems. A student who thoroughly understood the material in Volume 1 would know almost everything needed to be able to understand a totally new and generic sensor system. However, he would know very little about any specific sensor technology, such as radar or sonar.

Volume 2 discusses specific sensor technologies applicable to military systems. It treats every major kind of radar, infrared, electro-optical, acoustic, nuclear sensor system, as well as a few others. Each sensor is discussed in terms of sensor architecture, critical components and their function, key mission uses, and useful design relations. Analogy is used extensively. Once a technology or phenomenon has been discussed with respect to one sensor type, discussions of the same technology or phenomenon in other sensor types is done by analogy. The author has attempted to discuss every major sensor class in Volume 2. Both realized and potential future performance are evaluated. Sensor designers with a good grasp of one sensor type will often find that study of Volume 2 alone is enough to gain a good understanding of other sensor types. Novices will need to study Volume 1 first.

Volume 3 discusses conventional weapons and weapons systems. It breaks the discussion into two major divisions of weapons: electromagnetic weapons (information warfare, directed energy weapons, and electronic warfare) and projectile weapons (guns, missiles, bombs, torpedoes, etc.). Common aspects of each weapon type are pointed out as are differential aspects of individual weapons. Aspects of key components such as warheads, guidance, and propulsion are discussed in separate chapters. Volume 3 covers the gamut of conventional weapons useful for military purposes.

Volume 4 addresses unconventional weapons. In this broad category we have placed weapons of mass destruction (nuclear, chemical, biological, and radiological weapons), nonlethal weapons, and potential weapons of the future (weather warfare, tectonic weapons, gravitic weapons, advanced nuclear weapons, nanoweapons, cyborgs, clones, engineering warriors, etc.). Design characteristics, means of employment, and effects are presented for each type of unconventional weapon.

Volume 5 discusses control elements (communications and processing systems) and the overall integration of sensors, weapons, and control elements with the platforms to achieve usable and complete combat systems. Elements of survivability, observables reduction, structural hardening, operations analysis, electromagnetic compatibility, and system safety are necessarily covered here. Volume 5 is a candidate for possible division into two separate volumes in the future, but given the limited time available to cover these last topics in the courses for which the books are being developed, this is not likely to occur before a second or later edition.

For the most part, Volumes 3 and 4 can be used and studied independent of Volumes 1 and 2. However, as sensors play a role in many practical weapons systems, prior study of Volumes 1 and 2 will proved beneficial in appreciating Volumes 3 and 4. Volume 5 can be studied without prior study of Volumes 1-4. The principles are understandable and applicable in their own right. However, the presentation of these principles is geared to match the information and presentation in the earlier volumes

Each chapter of each volume is augmented by references to sources of specific results being quoted or of further information on key topics. A limited set of problems is also presented. These may be used for homework. However, their primary use is to indicate the kinds of problems that a combat systems engineer may encounter and to provide a rationale for the topics being covered.

The author welcomes comments, corrections, suggested improvements, interesting problems, anecdotes, examples of the application of various principles, sample data from real sensor systems (if publishable), and additional topics for inclusion. This work has been evolving for nearly a decade. It should continue to evolve in the future. The author can be contacted via regular mail at:

Robert C. Harney, Ph.D.
Department of Systems Engineering
Naval Postgraduate School
Monterey California, 93943

or via electronic mail at:

harney@nps.edu.

Robert C. Harney
February 2005

PART I

SENSOR FUNCTIONAL CHARACTERISTICS

PREFACE

This section addresses those attributes which are common to broad classes of sensor system. For example, the external environment dramatically affects the performance of all sensor systems. Similarly, the ability to perform critical functions such as detection or tracking, can be related to certain universal sensor characteristics in an unambiguous fashion. Among these common attributes are the characteristics of observable signals, the effects of propagation through the environment, and the functions desired to be performed. Other than their use as examples, specific sensor technologies are not discussed until Part II.

Specifically, this section describes a set of tools that can be used by the analyst to predict and compare the performance of any kinds of sensor systems. Use of these tools permits "apples to apples" trade studies to be made between sensors with little or no physical similarity.

In addition, it will become apparent that even after a detailed discussion of radar or sonar components, integration, and operation, the student will be unable to predict the performance of any radar or sonar until after he has learned how to obtain and incorporate the results of detection theory, estimation theory, and propagation theory, among others.

This material is presented first in the hopes of instilling a recognition that sensors can be studied as a unified topic. In this context, attempts to impart "special" status to certain sensors such as "radar" or "sonar" will be seen to be not only superficial and but also potential causes of less-than-optimal designs.

CHAPTER 1

SIGNATURES, OBSERVABLES, & PROPAGATORS

Sensors, Transducers, and Detectors

This volume deals with sensors and their performance. According to the dictionary, a **sensor** is:

> *"a device that responds to a physical stimulus (as heat, light, sound, pressure, magnetism, or a particular motion) and transmits a resulting impulse (as for measurement or operating a control)."* [1]

For purposes of this text, the author prefers his own slightly more pragmatic definition. A sensor is:

> *"a device that measures, observes, detects, or otherwise characterizes one or more attributes of an object, a collection of objects, or the external environment."*

In order to perform these tasks, a sensor must take information about the target that is expressed in one form (emitted or scattered waves, fields, or particles) and convert them into another form (currents in a wire or nerve impulses transmitted to our brains). The dictionary defines a **transducer** as:

> *"a device that is actuated by power from one system and supplies power usually in another form to a second system (as a telephone receiver is activated by electric power and supplies acoustic power to the surrounding air."* [1]

The author's broader definition is:

> *"a device which converts information-carrying signals from one physical form into another physical form."*

The process of conversion is called transduction. An example of a transducer is a silicon photodiode which converts light intensity into electrical current. All sensor systems contain transducers. In many sensor systems we refer to the transducer element as a "**detector**", especially if it converts the incident signal into an electrical current.

Observables

Physical objects possess a myriad of properties: mass, volume, shape, motion, composition, ability to emit radiation, ability to absorb or reflect radiation, etc. Most of these properties can be detected and measured with appropriate sensors. Any property of an object that is measured or otherwise observed is called an **observable** of that object. For example, mass can be measured (in principle) by a scale. Mass is a potential observable of the object. Mass also creates a gravitational field, and the resulting field can be measured (in principle) by a gravitometer. Thus the gravitational field of an object is a potential observable. If an observer actually possesses and uses a sensor capable of measuring a potential observable, it becomes a real observable. Any object possesses hundreds of potential observables (too many to attempt to list), however, any practical sensor system or suite of sensors will only exploit a few real observables.

If there is a direct physical relationship between two observables, then measurement of one can be used to infer the value of the other. The relationship between mass and gravitational field strength is well established, so measurement of one allows us to unequivocally predict the value of the other. If this is true, then we assume that both quantities are observables of the system.

Observables can have measurable values that range from very small to very large. Usually, the ability of a sensor system to measure an observable depends on the magnitude of its physical value. The property of being able to measure values of an observable other than zero is called **observability**. Objects with limited observability are hard to detect and measure and are said to be **low observable**. **Stealth** is the technology of designing objects to possess low observability and/or reducing the value of currently possessed observables. Stealth and observables reduction will be discussed in Volume 3 of this series.

Signatures and Propagators

The measured value of any observable is called a signature of the object. The dictionary [2] defines a **signature** as a noun:

> *"a characteristic trace or sign that indicates the presence of a substance or the occurrence of a physical event."*

and as an adjective:

> *"serving to identify or distinguish a person or group."*

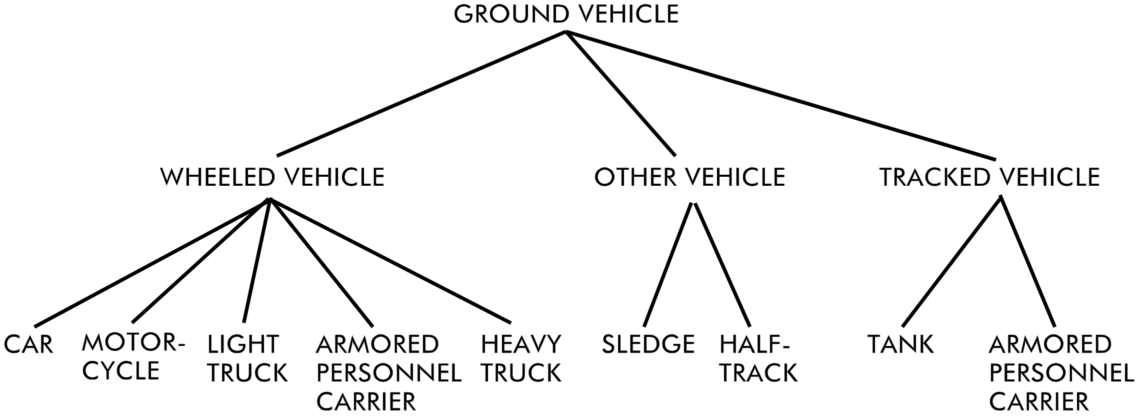
As in the earlier section, we shall broaden the definition of the noun to:

> *"a characteristic value of an observable that identifies the presence of an object, the occurrence of an event, and/or aids in identifying or distinguishing an object from other objects."*

Although the dictionary definitions imply uniqueness to signatures, just as one's signature is supposed to identify you sufficiently uniquely that when it is placed on a check, your bank will give your money to someone else. However, the mark "X" is still used as a valid signature for illiterate persons. Clearly, X does not convey much uniqueness. Although the author believes that the term signature should be used for the measured value(s) of an observable or observables, especially those which discriminate the object from other objects, the terms observable and signature are often used interchangeably with little resulting confusion.

As we shall see later in this volume, all information is useful. However, some information may not seem so at first glance. If all we can do is measure mass, and we measure a mass of 1000 kg, this does not tell us whether it is a small car or a large bomb. However, it can tell us that the object is not a man and it is not an armored vehicle. This may or may not seem to be immediately useful. However, consider execution of a search of a classification tree (this topic will be discussed in more detail in Chapter 12). Basically, a classification tree is a diagram that breaks objects into finer and finer classes giving progressively more information about the object. A simple classification tree is illustrated in Figure 1-1. Any information that either directs a search onto a particular branch of the tree or discourages search of any branch or branches is useful information.

Figure 1-1. A classification tree for ground vehicles.



People who work with or design sensors are familiar with a limited number of observables. However, as we discussed in the previous section, there is an enormous variety of observables and a correspondingly large variety of sensors. To gain a better appreciation of this variety we have attempted to classify all known observables in several different ways. Such classifications serve not only to provide a semblance of order to a chaotic situation, but also sets the stage for possible future identification of novel (at least to most people) observables.

One distinct mode of classification is by the propagator of the observable information. A **propagator** is the physical entity that carries the observable information. If the observable is the radar cross section of a target, the propagator is electromagnetic radiation. If the observable is the sounds emitted by a target, the propagator is acoustic radiation. Based on his analysis and observations, the author has identified six distinctly different classes of propagator. These are listed along with examples of observables associated with each class of propagator and significant subclasses of propagator in Table 1-1.

We have already identified **electromagnetic radiation** as a major propagator of observable information. As we shall see in the next chapter there are several major and numerous minor subdivisions of the electromagnetic spectrum (for example, radio frequency, microwave, millimeter-wave, infrared, visible, and ultraviolet). Individual sensors seldom work over more than one of these subdivisions. **Acoustic radiation** is another major propagator. Sound waves carry information through the air and through water. Seismic waves carry acoustic information through the ground. Acoustic and electromagnetic radiation are the propagators associated with most of the sensors with which military personnel are familiar. Radars, sonars, thermal imagers, television imagers, image intensifiers, laser rangefinders, etc., all use either acoustic or electromagnetic radiation. **Nuclear radiation** is another significant propagator of information. The author has included x-rays and gamma rays in the class of nuclear radiation propagators, even though they are forms of electromagnetic radiation. The rationale for this involves the observation that x-rays and gamma rays tend to be associated with the same kinds of physical processes with which neutrons or alpha particles are associated. Furthermore, x-ray and gamma ray detectors tend to have different designs than electromagnetic detectors operating at lower energies such as the ultraviolet. In this case, the author decided that practical aspects outweighed theoretical ones.

Radiation is not the only means by which an observable may be propagated. **Atoms and molecules** can be propagators. Internal combustion engines emit distinct chemical pollutants (e.g., ethylene and carbon monoxide) in their exhausts. The water vapor produced in jet engines may form contrails (condensation trails). One of the major functions of the Air Weather Service is to calculate probabilities of contrail formation to permit military flights to be routed to minimize the production of unmistakably observable contrails. Human beings emit detectable odors. There were multiple sensor systems developed during the Vietnam War that were dropped through the jungle canopy along the Ho Chi Minh Trail [3]. These sensors were designed to detect the ammonia released from urine. When the sensors detected the presence of urine in an area, air strikes would be called in to bomb the suspected areas. In this instance we probably killed more water buffalo than North Vietnamese soldiers, but the sensors worked. Chemical warfare agents often propagate their own

Table 1-1. Classification schemes of observables.

BY PROPAGATOR	**RELATED OBSERVABLES**
ELECTROMAGNETIC	* ULTRAVIOLET, VISIBLE, INFRARED, MILLIMETER-WAVE, RADIATION MICROWAVE, OR RADIO FREQUENCY RADIATION
ACOUSTIC RADIATION	* SOUND WAVES, SEISMIC WAVES
NUCLEAR RADIATION	* X-RAYS, GAMMA RAYS, NEUTRONS, ALPHA PARTICLES, BETA PARTICLES
ATOMS & MOLECULES	* ENGINE EXHAUSTS, PERSONNEL ODORS, CHEMICAL AGENTS, CONTRAILS
ENERGY & MOMENTUM	* WAKE TURBULENCE, SHOCK WAVES, ENVIRONMENTAL HEATING, TURF/FOLIAGE DAMAGE
FIELDS	* MAGNETIC FIELDS, ELECTRIC FIELDS, GRAVITY, ELECTRIC CONDUCTIVITY, MAGNETIC PERMEABILITY

BY SOURCE	**RELATED OBSERVABLES**
EMISSION	* INTENTIONAL (ON-BOARD RADAR, SONAR, RADIO, OR LASER RANGER) * INADVERTENT (EMI, ACOUSTIC NOISE, THERMAL RADIATION, RADIOACTIVE DECAY)
REFLECTION/SCATTERING/ ABSORPTION	* NATURAL ILLUMINATION (VISION, TELEVISION) * ACTIVE ILLUMINATION (RADAR, SONAR, LASER)
OCCULTATION	* SHADOWS, HOLES IN BACKGROUND NOISE
INELASTIC SCATTERING	* FREQUENCY-SHIFTED RADIATION * INDUCED EMISSIONS * KINEMATIC PERTURBATIONS
RESIDUE	* MOLECULAR (EXHAUST, CONTRAIL, ODORS) * ENERGY (WAKE TURBULENCE, TURF/FOLIAGE DAMAGE ENVIRONMENTAL HEATING)
ANOMALY	* FIELD STRENGTH, FIELD DIRECTION, INTERACTION STRENGTH

observable signatures in the form of their vapors. The atoms or molecules may be detected directly (e.g., with a mass spectrometer or a wet chemical sensor) or they may be detected indirectly by using electromagnetic radiation or some other propagator. Such hybrids involving more than one propagator often even more possibilities for sensor design.

Energy and momentum can act as propagators. Many people have felt the shock wave produced by a supersonic aircraft. This is an example of an energy/momentum propagator. Hearing the sonic boom is a hybrid propagator as it involves acoustic waves as well. Swimmers have felt the turbulence in the wake of a passing speedboat. Those who see the wake experience another hybrid propagator. Environmental heating produced by the presence of a warm object is another energy propagator example. Lastly, the expenditure of energy and momentum in a location can cause physical alteration of structures in that environment. Simple examples, like the foliage damage or turf displacement that is produced when an animal moves through the brush have been exploited by hunters and trackers for millennia. Finally, **fields** can be propagators. Many objects generate electric fields, magnetic fields, or gravitational fields. A magnetohydrodynamic propulsion unit may create a magnetic field that is measurable at considerable distances. Other objects may modify the characteristics of external fields. Different objects or conditions can alter the interactions characteristics of fields. Magnetic permeability, or electric conductivity are examples of interaction strengths that can carry information. For example, the conductivity of soil is a function of its moisture content and density. Both moisture content and density are different in an area of disturbed soil (such as that produced by digging a hole to bury a landmine) compared to undisturbed soil. Sensor technologies exist that are capable of measuring fields and interaction strengths of all kinds.

Observables can also be categorized by their observable source. Figure 1-1 also illustrates this second categorization. **Emission** is a prominent source of information. Emission may be intentional such as that produced by sensors (radar, sonar, laser ranger, radios, etc.) on board the platform or it may be inadvertent, produced by natural processes (thermal emission or radioactive decay) or by artificial processes (such as electromagnetic interference or acoustic noise). **Reflection, scattering, and/or absorption** of radiation is another prominent source of observable characteristics. The source of the radiation may be natural illumination such as sunlight or earthshine (thermal emission from the surroundings) or it may be produced by active illumination by sensors such as radars, sonars, or laser sensors. Television systems and image intensifier systems may used either natural or active illumination. Related to absorption but qualitatively different is **occultation**, the loss of a background signal due to blockage of that signal by nearby objects. A shadow is a form of occultation. So is the loss of background acoustic noise when it is blocked by a submarine coated with acoustic coatings or the loss of blue sky (or night sky) background when an air target passes through the sensor line of sight. Lack of signal is a useful form of signal in its own right.

Inelastic scattering is the source of three other kinds of observable. Inelastic scattering is any process in which one object interacts with a second object and loses (or gains) energy from the interaction. One or both objects may change their fundamental character in the interaction. Radiation may interact with atoms or molecules and emerge with characteristic shifts in frequency. For example, Raman scattering is a process in which electromagnetic radiation with a fixed input frequency scatters from molecules and emerges with different frequencies which have been shifted by the vibrational and/or rotational frequencies of the molecule. Induced emissions are different form of inelastic scattering. Nuclear particles incident on certain materials can be absorbed with the resultant emission of different particles. Neutrons incident on fissionable materials will react producing very distinctive emissions that can identify the material as being fissionable. Kinematic perturbations can be a source of observable characteristics. Every hunter or infantryman is familiar with the change in motion of a target when it is hit by a rifle bullet. In a more bizarre application, there have been serious proposals to discriminate reentry vehicles from decoys by bombarding them with sand (or possibly particle beams). Reentry vehicles are heavy and will not have their velocity changed significantly by a small change in momentum. Light decoys such as balloons will slow down considerably when they are hit with a small amount of high-velocity sand.

Residues are another source of observables. Molecular residues are vapors and odors produced by equipment or personnel. They also include fuel leakage, oil slicks, contrails, smoke from engine exhausts or cooking fires, etc. Energy residues include environmental damage, environmental heating, wakes, shock waves, turbulence, etc. Residues may yield the knowledge that certain kinds of system are (or were) present in the vicinity, without requiring direct observation of those systems. The last major source of observables is **anomalies.** An anomaly is an abnormal condition observed in a typically static quantity. Anomalies may exist in field strength, field direction, or interaction strength. For example, an iron object will modify the earth's magnetic field in a characteristic fashion. At large distances, the change in the Earth's field (the magnetic anomaly) will have a characteristic dipole shape.

A final way of categorizing observables is on the basis of their relative strength. **Signal-dominant** observables are characterized by a well-defined signal corrupted by a small amount of noise usually produced in the sensor itself. Radars and infrared sensors looking at sky backgrounds are examples in which the signal dominates. **Clutter-dominant** observables are characterized by signals that do not differ very much in strength from the background. The signal must be extracted based on small contrasts against the background. Variations in the background (clutter) may exceed the contrast of the target against the background. Imaging sensors, sonars, and radars looking at terrestrial backgrounds are often dominated by clutter. Typically, different kinds of signal processing must be employed when dealing with signal-dominant observables than when dealing with clutter-dominant observables. These distinctions will be addressed further when we discuss sensor functions and sensor implementations.

The primary reason for this chapter is the indication to the reader that there is much more to military sensor systems than initially meets the eye. Weapons systems have seriously exploited only two of the six classes of propagators and only two of the six classes of observable sources. As threats evolve to counter the capabilities of existing sensor systems we should be looking to novel ways of performing our mission. This will involve looking at sensors that exploit observables that are not being considered today. Today's sensors use obsolete technology. The sensors under development today will be obsolete by the time they are fielded. Technology will continue to evolve at an exponential rate. As we progress through this text, every attempt will be made to provide all of the conceptual tools that will permit the reader to address non-traditional or even “science fiction” sensor systems as they are developed, to an equivalent degree of sophistication as more traditional sensors are treated today.

References

[1] Mish, Frederick C. (Editor in Chief), Merriam Webster's Collegiate Dictionary 10th Ed. (Merriam-Webster, Springfield MA, 1993).

[2] Anonymous, Webster's New Universal Unabridged Dictionary (Barnes & Noble Books, New York NY, 1996).

[3] Dickson, Paul, The Electronic Battlefield (Indiana University Press, Bloomington IN, 1976).

Problems

1-1. For each “propagator” of observables in Table 1-1 give at least one concrete example of a sensor using that observable. List as many examples as you can, but avoid duplication of basic principles.

1-2. For each “source” of observables in Table 1-1 give at least one concrete example of a sensor using that source. List as many examples as you can, but avoid duplication of basic principles. Every example listed in Problem 1-1 should be categorized in Problem 1-2.

1-3. For some military mission with which you are familiar, hypothesize a sensor to perform key aspects of that mission. However, you must choose a sensor whose propagator and source are not currently used or proposed to perform that mission.

CHAPTER 2

PROPAGATION OF ELECTROMAGNETIC RADIATION. I. FUNDAMENTAL EFFECTS

What is Electromagnetic Radiation?

Electromagnetic radiation consists of simultaneously oscillating electric and magnetic fields that can move from one point in space to another preserving certain fundamental characteristics of the oscillations. Of primary interest is the fact that these oscillating fields can transport energy from one point in space to another. No physical medium is required.

An old joke among physicists is based on Genesis 1:1 through 1:3 of the Bible [1]. In the language of the "King James version" it goes:

> *"In the beginning, God created the heaven and the earth. And the earth*
> *was without form, and void; and darkness was upon the face of the deep.*
> *And the Spirit of God moved upon the face of the waters.*
> *And God said:*
>
> $$\nabla \cdot \vec{D} = 4\pi\rho, \tag{2.1}$$
>
> $$\nabla \cdot \vec{B} = 0, \tag{2.2}$$
>
> $$\nabla \times \vec{E} + \frac{1}{c}\frac{\partial \vec{B}}{\partial t} = 0, \tag{2.3}$$
>
> $$\nabla \times \vec{H} - \frac{1}{c}\frac{\partial \vec{D}}{\partial t} = \frac{4\pi \vec{J}}{c}: \tag{2.4}$$
>
> *And there was light!*

The humor comes from the fact that the existence of electromagnetic radiation (i.e., light) is a direct and inevitable consequence of Maxwell's Equations – equations (2.1) through (2.4) above. Had "God" uttered Maxwell's equations, instead of the Scriptural "*Let there be light!*", there would have been light without further ado. Equations (2.1)-(2.4) plus the constitutive relations:

$$\vec{D} = \vec{E} + 4\pi\vec{P} = \vec{E} + 4\pi\underline{\chi_e}\vec{E} = (1 + 4\pi\underline{\chi_e})\vec{E} = \underline{\varepsilon}\vec{E} \tag{2.5}$$

and

$$\vec{B} = \vec{H} + 4\pi\vec{M} = \vec{H} + 4\pi\underline{\chi_m}\vec{H} = (1+4\pi\underline{\chi_m})\vec{H} = \underline{\mu}\vec{H} \qquad (2.6)$$

can together be easily solved to yield wave equations.

Before performing this mathematical *piece de resistance* the multiplicity of terms used above needs to be defined. Specifically:

$\vec{E}$ = Electric Field

$\vec{D}$ = Electric Displacement

$\vec{P}$ = Electric Polarization

$\underline{\chi_e}$ = Electric Susceptibility

$\underline{\varepsilon}$ = Dielectric Constant

$\vec{B}$ = Magnetic Induction

$\vec{H}$ = Magnetic Field

$\vec{M}$ = Magnetization

$\underline{\chi_m}$ = Magnetic Susceptibility

$\underline{\mu}$ = Magnetic Permeability

c = Speed of Light in Vacuum

ρ = Charge Density

$\vec{J}$ = Current Density

The underlined quantities are complex constants. Now assume that the region of space of interest is empty (i.e., charge density and current density are zero). In empty space, Maxwell's equations may be written as

$$\nabla \cdot \vec{E} = 0, \qquad (2.1a)$$

$$\nabla \cdot \vec{B} = 0, \qquad (2.2a)$$

$$\nabla \times \vec{E} + \frac{1}{c}\frac{\partial \vec{B}}{\partial t} = 0, \qquad (2.3a)$$

$$\nabla \times \vec{B} - \frac{\underline{\mu}\underline{\varepsilon}}{c}\frac{\partial \vec{E}}{\partial t} = 0. \qquad (2.4a)$$

If we take the curl (▽×) of equations (2.3a) and (2.4a) we obtain

$$\nabla \times \nabla \times \vec{E} + \frac{1}{c}\frac{\partial \nabla \times \vec{B}}{\partial t} = 0 \tag{2.7}$$

and

$$\nabla \times \nabla \times \vec{B} - \frac{\underline{\mu}\underline{\varepsilon}}{c}\frac{\partial \nabla \times \vec{E}}{\partial t} = 0 \tag{2.8}$$

If we now use an identity relationship from multidimensional calculus

$$\nabla \times \nabla \times \vec{Y} \equiv \nabla\left(\nabla \cdot \vec{Y}\right) - \nabla^2 \vec{Y} \tag{2.9}$$

we obtain

$$\nabla\left(\nabla \cdot \vec{E}\right) - \nabla^2 \vec{E} + \frac{1}{c}\frac{\partial \nabla \times \vec{B}}{\partial t} = 0 \tag{2.10}$$

and

$$\nabla\left(\nabla \cdot \vec{B}\right) - \nabla^2 \vec{B} - \frac{1}{c}\frac{\partial \nabla \times \vec{E}}{\partial t} = 0 \tag{2.11}$$

Upon substitution of the original curl and divergence equations (2.1a)-(2.4a) back into Eq. (2.10) and (2.11) we obtain the wave equation solutions

$$\nabla^2 \vec{E} - \frac{\underline{\mu}\underline{\varepsilon}}{c^2}\frac{\partial^2 \vec{E}}{\partial t^2} = 0 \tag{2.12}$$

and

$$\nabla^2 \vec{B} - \frac{\underline{\mu}\underline{\varepsilon}}{c^2}\frac{\partial^2 \vec{B}}{\partial t^2} = 0\,. \tag{2.13}$$

Equations (2.12) and (2.13) admit solutions of the form

$$\vec{E}(\vec{r},t) = \vec{E}_0\, e^{i\omega t \mp i\underline{n}\left(\vec{k}_0 \cdot \vec{r}\right)} \tag{2.14}$$

and

$$\vec{B}(\vec{r},t) = \vec{B}_0\, e^{i\omega t \mp i\underline{n}\left(\vec{k}_0 \cdot \vec{r}\right)} \tag{2.15}$$

where

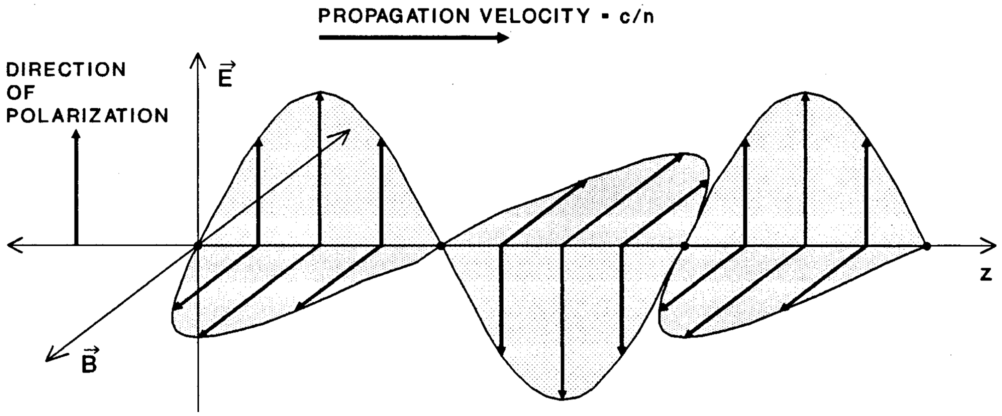
$$\underline{n} \equiv \sqrt{\underline{\mu}\underline{\varepsilon}} \cong \sqrt{\underline{\varepsilon}} \quad \text{for non - magnetic media} \tag{2.16}$$

and

$$\omega = c\left|\vec{k}_0\right| = ck_0 = 2\pi c / \lambda\,. \tag{2.17}$$

These solutions oscillate sinusoidally in both space and time with temporal frequency ω/2π and free-space wavelength λ, as illustrated in Figure 2-1.

Figure 2-1. Graphical representation of electromagnetic radiation as simultaneous in-phase oscillations of electric and magnetic fields oriented at right angles to each other and the direction of propagation.



Equations (2.14) and (2.15) are plane waves in that the amplitude of the oscillations varies only sinusoidally along the direction of propagation and there is no variation in amplitude perpendicular to the direction of propagation. The wave equations admit another kind of solution, that is, spherical waves. The spherical wave may be represented as

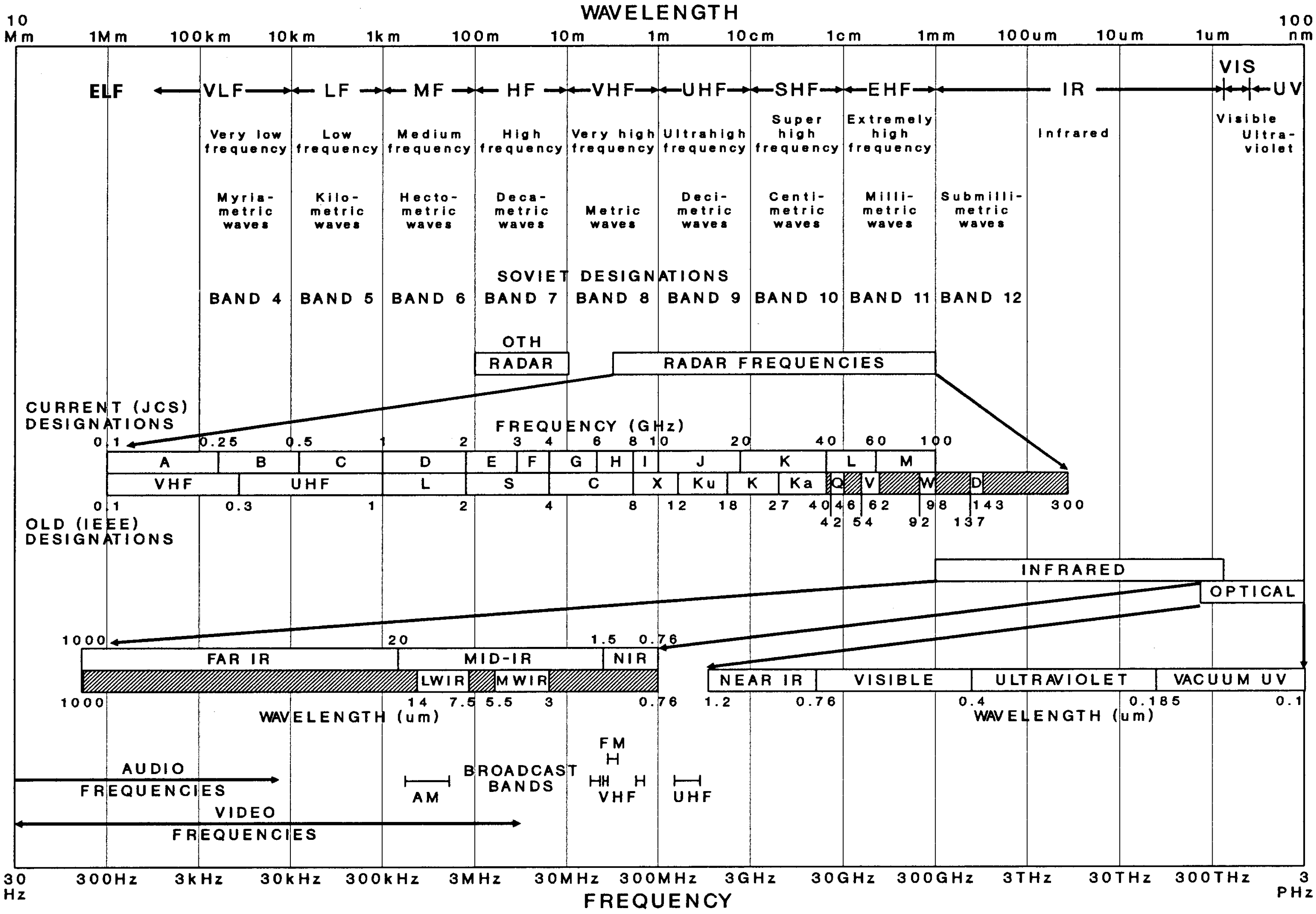
$$\vec{E}(r,t) = \vec{E}_0 \frac{e^{i\omega t \mp ikr}}{r} \tag{2.18}$$

$$\vec{B}(r,t) = \vec{B}_0 \frac{e^{i\omega t \mp ikr}}{r} \tag{2.19}$$

Spherical waves either expand radially outward from a point or collapse radially inward on a point. The field strength falls off inversely with the radial distance *r* from the point of expansion or collapse. Both plane and spherical waves find use in describing phenomena associated with electromagnetic radiation.

Before continuing with discussions of these phenomena, it is instructive to discuss some of the nomenclature associated electromagnetic radiation of different frequencies. Figure 2-2 shows

Figure 2-2. The electromagnetic spectrum from radio frequencies to the ultraviolet.



the electromagnetic spectrum from radio frequencies to the ultraviolet. In truth, the electromagnetic spectrum extends well beyond the ultraviolet into the x-ray and gamma-ray regions. We have not included the very high frequency x-ray and gamma-ray regions because these are most commonly encountered as by-products of nuclear reactions. We will discuss these and other nuclear radiations in a separate chapter. Electromagnetic radiation with frequencies below 30 Hz is detected by sensors which are essentially capable of detecting static fields. We will consider them separately when we discuss "static" fields and anomalies.

Within the remaining electromagnetic spectrum, many different groups of technologists have applied their own terminologies to different regions of the spectrum. In most cases, the nomenclature was merely descriptive, although in a few it was intentionally confusing. In the early days of radio research, transmissions tended to occur in the band from 30 kHz to 30 MHz without much discrimination. As people noticed differences in propagation characteristics, they started talking about low frequency, medium frequency, and high frequency radio waves, each covering roughly a decade of frequency. With the advent of radar systems, a great push to higher and higher frequencies began. Radars first moved into the very high frequency (VHF) band and later into the ultrahigh frequency (UHF) band. Again it was convenient to maintain roughly a decade between band centers. The development of the magnetron permitted radars to push beyond UHF frequencies. Some people referred to the very, very high frequency radiation as microwaves. Others continued using stronger comparatives such as superhigh frequency (SHF) and ultimately extremely high frequency (EHF). Fortunately, at this point people working on ever higher frequency radio waves started to transgress into the realm of those people working on ever longer wavelength infrared radiation and ultimately dropped the race for stronger comparatives. Otherwise we might be describing the near infrared as very-super-ultra-extremely-high frequency (VSUEHF) radio waves or some other ridiculous term. Interest in lower frequency radio waves led to the use of the terms very low frequency (VLF) and extremely low (ELF) radiation, to be consistent with existing radio frequency nomenclature.

Given that this terminology conveys very little information, other researchers started describing the frequency regions using an explicit wavelength terminology. Waves with wavelengths centered around one meter (i.e, 0.3 m to 3 m) were called metric waves. The next longer wavelength region was called dekametric waves; the next short wavelength region was called decimetric waves, and so on using the SI (International System of Units) prefixes. The EHF region is most commonly referred to as the millimetric-wave or millimeter-wave region of the spectrum. Researchers in the former Soviet Union adopted a terminology based on the exponent of the frequency at the center of each decade-wide band. Thus, Band 6 had a center frequency of 10^6 Hz (= 1 MHz).

During World War II, radar was a highly classified field. To make it easier to conduct long-distance development programs while retaining a degree of secrecy, the region of the spectrum in which radars operated was broken into a large number of narrow bands (typically no wider than an octave) and assigned a random letter (L, S, C, X, K, etc.). Researchers could describe a system as operating at C-band (i.e, at 4-8 GHz) and maintain a significant level of security. Only those with clearances knew the band codes. After the war, when radar proliferated beyond military applications, this terminology was retained. This situation lasted until the 1980's when the Joint Chiefs of

Staff promulgated a directive that simplified the alphabet soup into a simple A, B, C, D, etc. code sequence as shown in Figure 2-2.

The very high frequency end of the spectrum is subdivided into two partially overlapping regions: the infrared and the optical regions. The optical region spans from the ultraviolet around 0.1 μm to the near infrared around 1.2 μm. The cutoff at 1.2 μm is somewhat arbitrary. Photons (quanta of electromagnetic radiation) at wavelengths longer than about 1.2 μm cannot cause external photoelectrons in material. Thus photomultipliers and image intensifiers can only be implemented at wavelengths shorter than about 1.2 μm. The author has used this photoelectric cutoff to define the edge of the optical region of the spectrum. The optical spectrum is further subdivided into the visible spectrum (0.4 μm to 0.76 μm – the range of wavelengths most human beings can see), the ultraviolet (0.4 μm to 0.185 μm), and the vacuum ultraviolet (0.185 μm to 0.1 μm – all known materials absorb ultraviolet radiation below 0.185 μm). The infrared region can be subdivided in several ways. People working with thermal imaging systems tend to ignore infrared radiation with wavelengths longer than 14 μm and shorter than 3 μm. They talk about long-wave infrared (roughly 7.5 μm to 14 μm) and mid-wave infrared (roughly 3 μm to 5.5 μm). Other researchers have divided the infrared spectrum into far infrared (1000 μm to 20 μm), mid-infrared (20 μm to 1.5 μm), and the near infrared (1.5 μm to 0.76 μm). This last set of divisions is not universally accepted. One has a high probability of finding definitions that differ from any of the above..

In Figure 2-2 we have also listed the regions in which commercial radio and television systems operate, as well as ranges commonly ascribed to audio and video technologies. The primary purpose of Figure 2-2 is to serve as a reference that may serve to make it easier for the system engineer to collect and compare data from different sources and communities. Having a common set of definitions or at least a “Rosetta stone” (such as Figure 2-2) is essential to cross-cultural communication.

Attenuation of Electromagnetic Radiation in Bulk Materials

The solutions of the wave equations depend (among other things) on the dielectric constant of the medium in which the waves are propagating. In general, the dielectric constants of real materials are complex functions. They are also generally tensor quantities but such detail is the province of specialists and usually not of importance to systems engineers. If the dielectric function is complex, then the refractive index must also be complex. Let us define the complex dielectric function as having real (ε_1) and imaginary (ε_2) parts, as shown below.

$$\underline{\varepsilon} = \varepsilon_1 - i\varepsilon_2 \tag{2.20}$$

Then the refractive index must be complex

$$\underline{n} = \sqrt{\varepsilon} = n - i\kappa \tag{2.21}$$

with real part (n) and imaginary part (κ). The components of the refractive index are related to the real and imaginary components of the dielectric function by the relations

$$\varepsilon_1 = n^2 - \kappa^2 \tag{2.22}$$

$$\varepsilon_2 = 2n\kappa \tag{2.23}$$

$$n = \left[\frac{\varepsilon_1}{2} \left[1 + \left[1 + \frac{\varepsilon_2^2}{\varepsilon_1^2} \right]^{1/2} \right] \right]^{1/2} \tag{2.24}$$

$$\kappa = \left[\frac{\varepsilon_1}{2} \left[\left[1 + \frac{\varepsilon_2^2}{\varepsilon_1^2} \right]^{1/2} - 1 \right] \right]^{1/2} \tag{2.25}$$

At this point the many readers will be asking themselves what does it mean for a material to have a complex refractive index. As we shall soon see, the impact of a complex refractive index is that the material absorbs electromagnetic energy. The intensity (power flowing through a unit area) of electromagnetic radiation is proportional to the square of the electric field

$$\begin{aligned} I(\vec{r}) \propto \left|\vec{E}(\vec{r})\right|^2 &= \vec{E}_0^* \, e^{-i\omega t + in\left(\vec{k}_0 \cdot \vec{r}\right) - \kappa\left(\vec{k}_0 \cdot \vec{r}\right)} \vec{E}_0 \, e^{i\omega t - in\left(\vec{k}_0 \cdot \vec{r}\right) - \kappa\left(\vec{k}_0 \cdot \vec{r}\right)} \\ &= \left|\vec{E}_0\right|^2 e^{-2\kappa\left(\vec{k}_0 \cdot \vec{r}\right)} \approx I_0 \, e^{-2\kappa k_0 r} = I_0 \, e^{-\alpha r} \end{aligned} \tag{2.26}$$

where

$$\alpha = 2\kappa k_0 = 4\pi\kappa / \lambda \tag{2.27}$$

is the absorption coefficient. There is an enormous body of technical literature which has tabulated the real and imaginary parts of the refractive index, the real and imaginary parts of the dielectric coefficient, or the absorption coefficients of many materials at a number of wavelengths of interest. Use of the relations above permits conversion of data obtained in one form to other potentially more useful forms. As a concrete example, we reproduce the complex refractive index data for water in Figure 2-3 and compare this with the absorption coefficient calculated using the same data in Figure 2-4.

In the microwave region of the spectrum, some technologists have adopted a different nomenclature for absorption. A quantity called the loss tangent (*tan δ*) is defined such that

$$\tan\delta \equiv \varepsilon_2 / \varepsilon_1 . \tag{2.28}$$

Many references give tabulated data on loss tangent and real dielectric coefficient versus frequency for a number of materials. In terms of the loss tangent the other dielectric quantities can be determined from the relations

$$\underline{\varepsilon} = \varepsilon_1 \left(1 - i \tan\delta\right) \tag{2.29}$$

and

$$\alpha = \frac{4\pi}{\lambda}\left[\frac{\varepsilon_1}{2}\left(\left(1 + \tan^2\delta\right)^{1/2} - 1\right)\right]^{1/2} . \tag{2.30}$$

In the limit that the loss tangent is small compared to unity, Eq. (2.30) may be approximated by

$$\alpha \approx \frac{4\pi}{\lambda}\left[\frac{\varepsilon_1}{4}\tan^2\delta\right]^{1/2} \approx \frac{2\pi n}{\lambda}\tan\delta \tag{2.31}$$

with inverse

$$\tan\delta = \lambda\alpha / 2\pi n . \tag{2.32}$$

There are also occasions in which it is necessary to calculate the attenuation of electromagnetic radiation in conductive media. If σ is the conductivity (equal to 1/resistivity) of a medium then the attenuation coefficient is

$$\alpha = \sqrt{2\mu\omega\sigma} \approx 4\pi\sqrt{10^{-7}\,\nu\sigma}\ \text{(for non - magnetic materials)} . \tag{2.33}$$

Figure 2-3. The complex refractive index of pure liquid water at 298 K.[3]

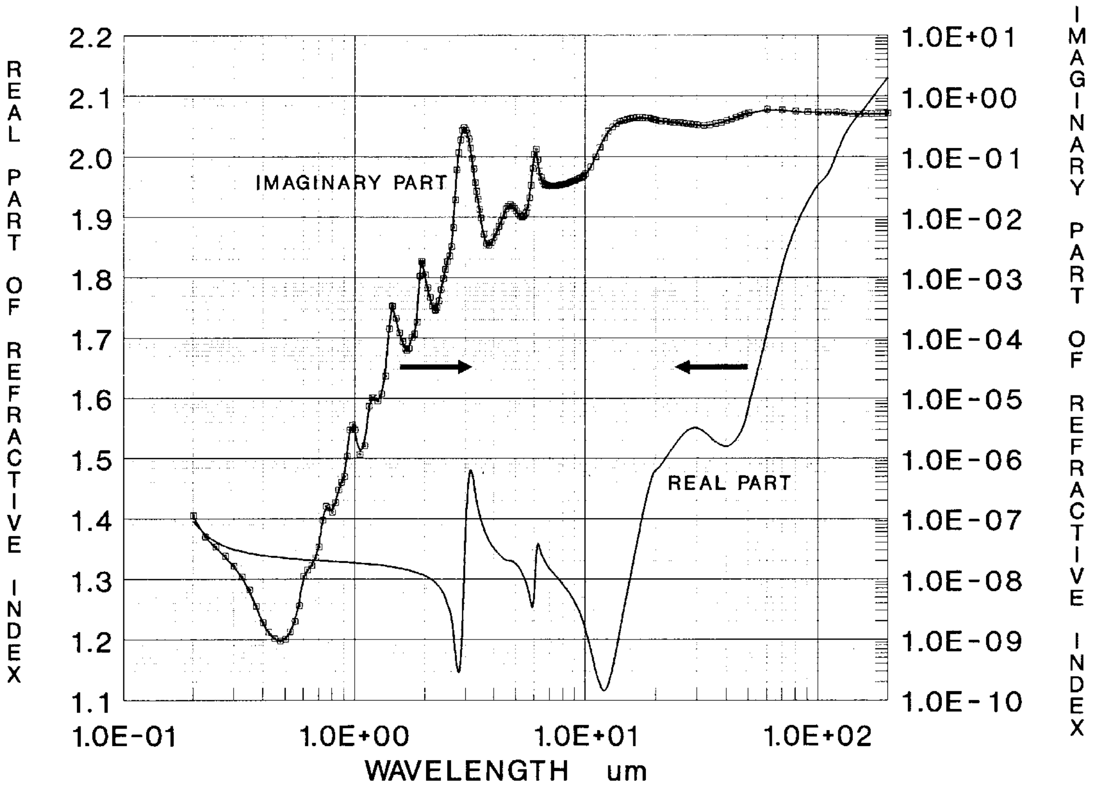
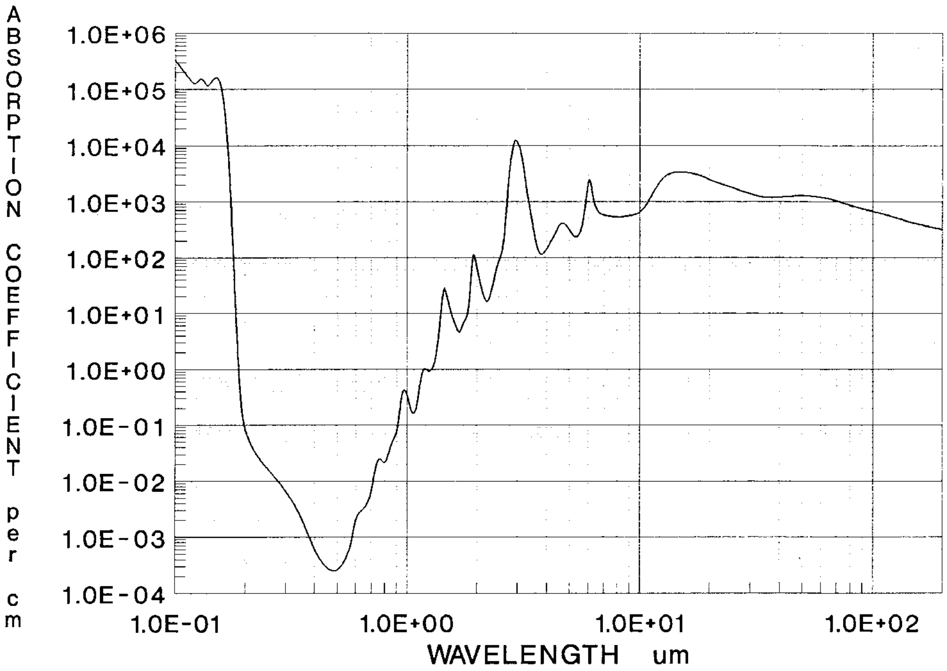


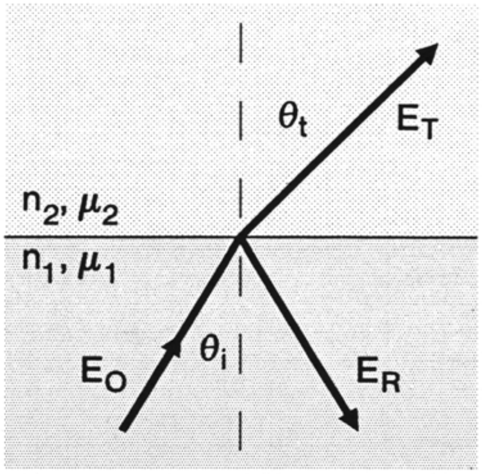
Figure 2-4. The absorption coefficient of liquid water calculated using the data from Fig. 2-3.



Reflection and Refraction

When an electromagnetic wave is incident on an interface between two propagation media with different characteristics, several things happen. Some of the radiation is reflected by the interface, while some of the radiation is transmitted. However, the propagation direction of the transmitted radiation is altered in a process known as refraction. Consider the situation outlined in Figure 2-5.

Figure 2-5. Geometry of reflection and refraction at an interface.



The incident ray with amplitude E_0 strikes the interface with an angle θ_i relative to the normal to the surface. The reflected ray with amplitude E_R leaves the interface at an angle θ_i relative to the normal and lies in the plane formed by the incident ray and the surface normal on the opposite side of the surface normal from the incident ray. The transmitted ray with amplitude E_T leaves the interface at an angle θ_t relative to the surface normal and lies in the plane formed by the incident ray and the surface normal.

The angle of the transmitted ray is related to the angle of the incident ray by Snell's Law of refraction

$$\sin\theta_t = \left(n_1 / n_2\right)\sin\theta_i \tag{2.34}$$

where n_1 and n_2 are the refractive indices of the two media as shown in the Figure. Snell's law of refraction coupled with the law of reflection (angle of reflection equals angle of incidence) form the basis of almost all optical design.

The reflected component is of significance for a number of reasons that will be discussed later in this section. Fresnel developed expressions that relate the reflected component amplitude to the incident amplitude. For this reason we commonly refer to reflection at an interface between dielectric media as Fresnel reflection [2]. Specifically, Fresnel obtained

$$\left.\frac{E_R}{E_0}\right|_\perp = \frac{n_1\cos\theta_i - \left[n_2^2 - n_1^2\sin^2\theta_i\right]^{1/2}}{n_1\cos\theta_i + \left[n_2^2 - n_1^2\sin^2\theta_i\right]^{1/2}} \tag{2.35}$$

for the component of incident radiation polarized perpendicular to the plane formed by the incident ray and the surface normal (that is, the component of the electric field vector that is perpendicular to this plane), and

$$\left.\frac{E_R}{E_0}\right|_{\parallel} = \frac{n_2^2 \cos\theta_i - n_1 \left[n_2^2 - n_1^2 \sin^2 \theta_i\right]^{1/2}}{n_2^2 \cos\theta_i + n_1 \left[n_2^2 - n_1^2 \sin^2 \theta_i\right]^{1/2}} \tag{2.36}$$

for the component of incident radiation polarized parallel to the plane formed by the incident ray and the surface normal. Because these two expressions are different, reflection from a dielectric surface may alter the polarization of the radiation. This is a major design concern in systems that attempt to measure the polarization of the radiation as an integral part of their function.

Equations (2.35) and (2.36) may be used to calculate the strength and degree of polarization alteration of reflected radiation at any angle of incidence. However, the equations are complicated and are usually not used for back-of-the-envelope analyses. However, at normal incidence, the expressions simplify considerably to

$$\left.\frac{E_R}{E_0}\right|_{\parallel \text{ or } \perp} = \frac{\mp(n_1 - n_2)}{n_1 + n_2} \tag{2.37}$$

Since the intensity (or power) in an electromagnetic field is proportional to the square of the field strength, Eq. (2.37) may be squared to yield

$$\left.\frac{P_R}{P_0}\right|_{\text{normal}} = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2 \tag{2.38}$$

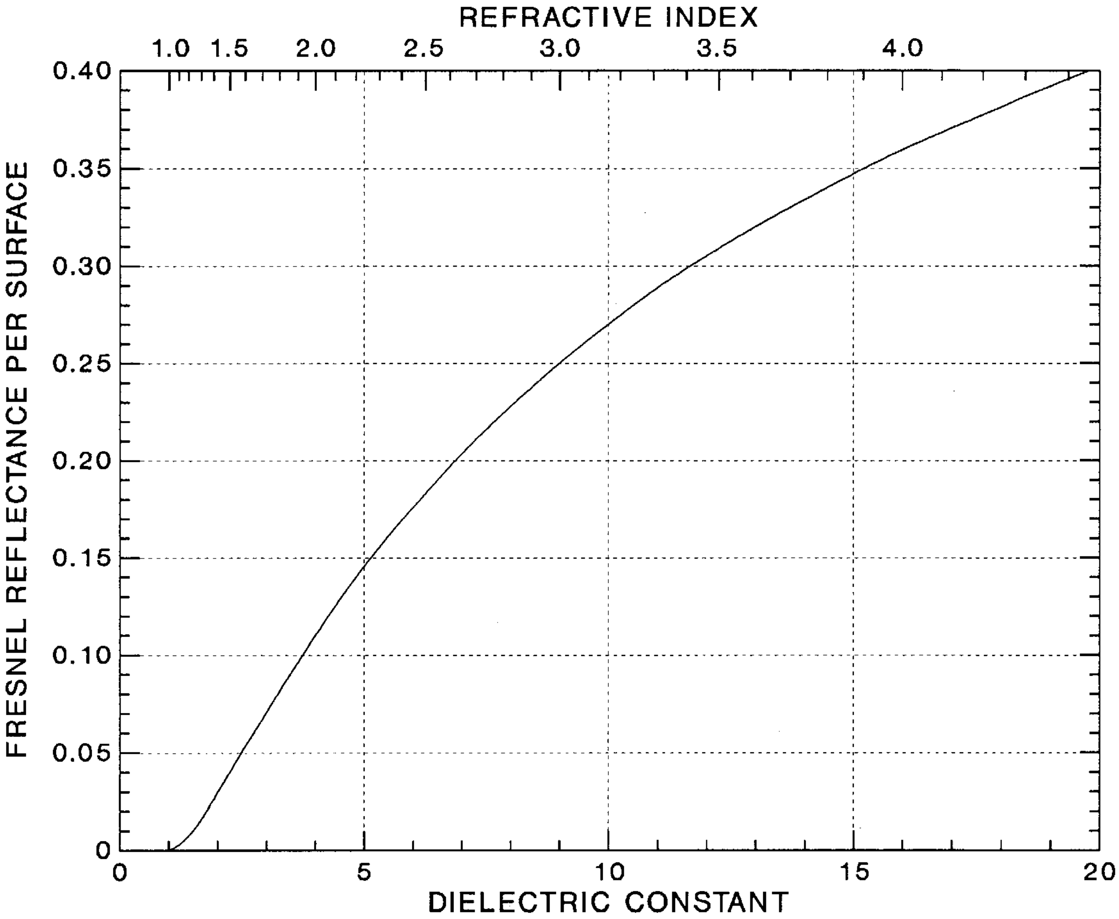
If one of the materials is air ($n_1 = 1$), then we get an even simpler expression

$$\left.\frac{P_R}{P_0}\right|_{\text{normal}} = \left(\frac{n_2 - 1}{n_2 + 1}\right)^2 \tag{2.39}$$

Equation (2.39) is plotted for various values refractive index (and dielectric constant) in Figure 2-6. Glass has a refractive index of roughly 1.5 relative to air. From the figure we find that the Fresnel reflectance is 0.04 per surface. A thin pane of glass will reflect 8% (4% per surface) of incident light. In another example, germanium has a refractive index 4.0 in the infrared. The infrared Fresnel reflectance of germanium is 0.36. Thus, each surface of a plate of germanium will reflect 36% of the incident infrared radiation. The reflectance is so high that multiple internal reflections must be considered when determining the total reflectance which corresponds to about 53%.

In general, the Fresnel reflectance for perpendicular polarized radiation will increase monotonically with increasing angle of incidence. However, for parallel polarized radiation, a quite

Figure 2-6. Fresnel reflectance of a surface as a function of relative refractive index or dielectric coefficient.



different behavior is observed. The reflected component will decrease to zero as the angle of incidence approaches an angle called Brewster's angle. Brewster's angle is defined by

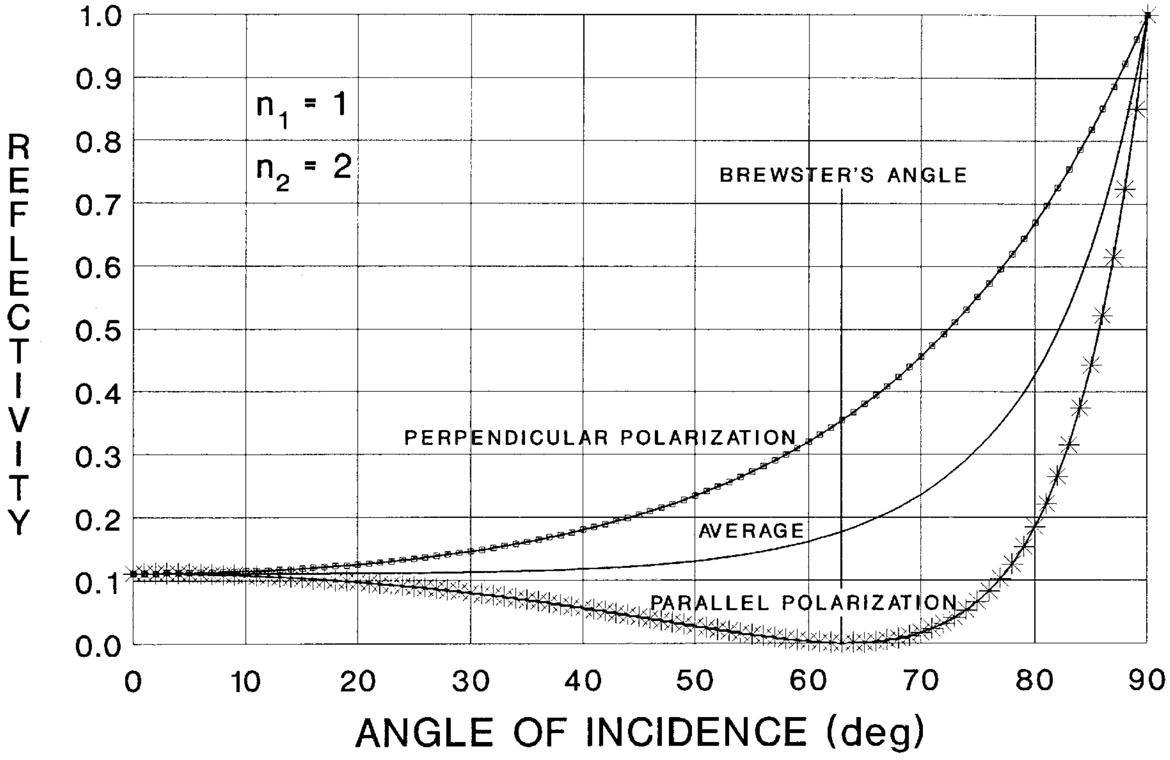
$$\theta_B = \tan^{-1}\left(n_2 / n_1\right). \tag{2.40}$$

At Brewster's angle the reflectance of the perpendicular polarization is given by

$$\left.\frac{E_R}{E_0}\right|_{\perp} = \frac{n_1 \cos\theta_B - \left[n_2^2 - n_1^2 \sin^2\theta_B\right]^{1/2}}{n_1 \cos\theta_B + \left[n_2^2 - n_1^2 \sin^2\theta_B\right]^{1/2}} \neq 0 . \tag{2.41}$$

Thus, unpolarized light that is reflected off of a surface at Brewster's angle will be transformed into light which is perfectly polarized in the perpendicular polarization. The different behavior of the two polarizations is shown graphically in Figure 2-7. Note that the average reflectivity for unpolarized light (light containing equal amounts of parallel and perpendicular polarization increases with increasing angle of incidence. However, the increase is very small until the incident angle exceeds 40°.

Figure 2-7. Angular dependence of Fresnel reflection.

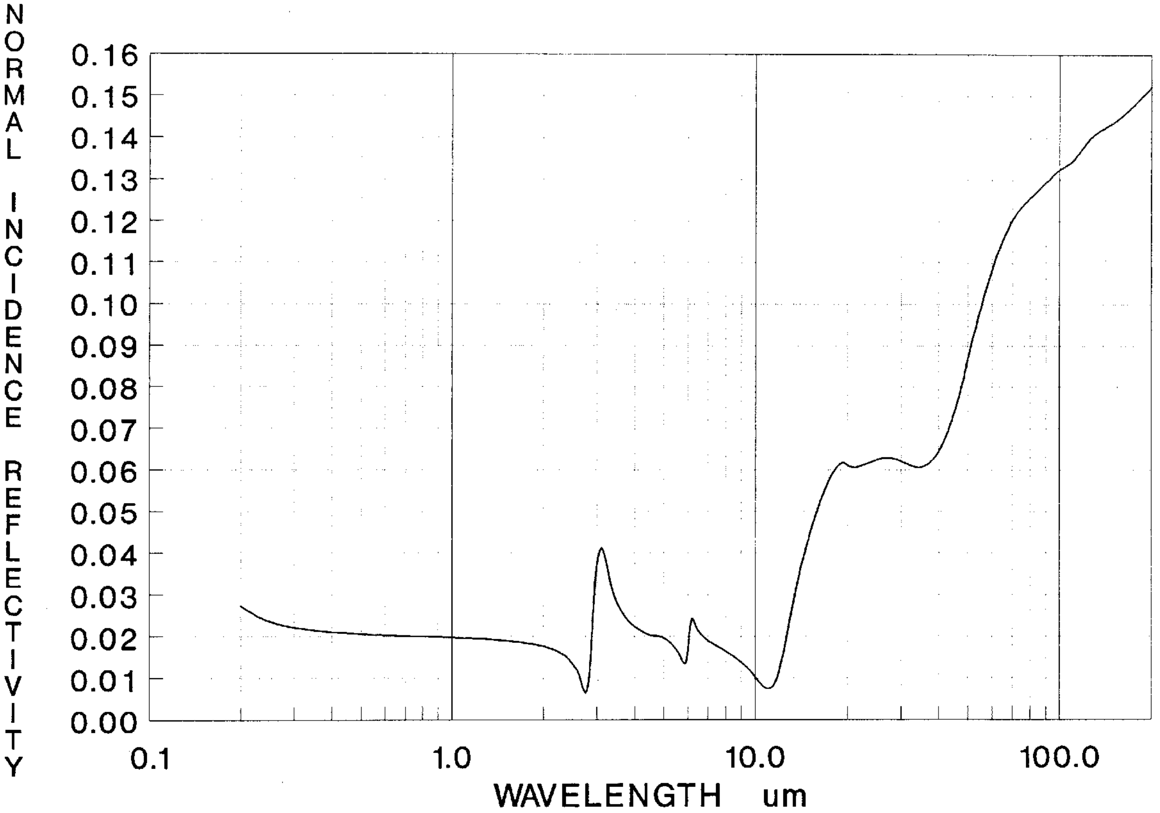


For completeness, Figure 2-8 shows the calculated reflectivity at normal incidence of liquid water using the data from Figure 2-3. This was calculated using the expression

$$r = \left[(n-1)^2 + \kappa^2\right] \Big/ \left[(n+1)^2 + \kappa^2\right] \tag{2.42}$$

which can be obtained directly from Eq. (2.39) by expressing n_2 explicitly as the sum of its real (n) and imaginary (κ) parts. The reader may question why such a large quantity of supplemental material is being included. Part of the answer lies in attempting to bring esoteric mathematical results to a more concrete form. Another part of the answer is that the information may be useful to the reader in the future and is often hard to locate. To a significant extent, the author hopes that this work will not only serve a textbook for combat systems engineering, but also as a handbook to assist the engineer in what the author acknowledges as an extremely difficult task.

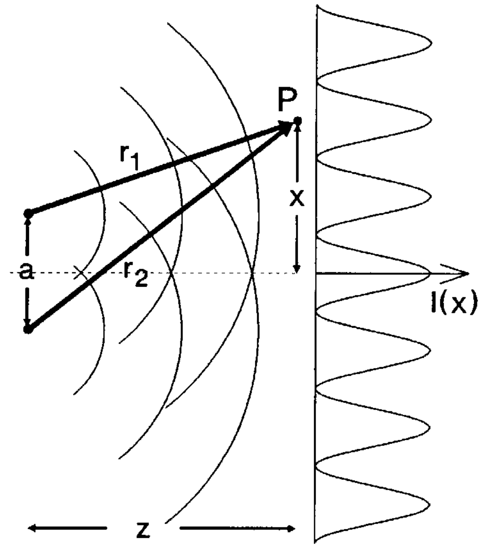
Figure 2-8. Reflectivity of liquid water at normal incidence.



Interference

Two electromagnetic waves can interfere with each other. The intensity at any point in space is equal to the absolute value squared of the local electric field. The local field is the vector sum of the electric fields originating from all contributing sources. If the frequencies of the two wave differ, then the interference pattern comes and goes at a frequency equal to the difference in frequency between the two waves. The effect is not often directly observed because human beings typically have sensor frequency responses which are too low to see the difference signal. Nevertheless, this is exploited in heterodyne receivers. Heterodyne reception will be discussed in more detail in Chapter 14 on coherent laser radars. However, if the two waves have the same frequency, the interference pattern is stationary and easily observed.

Figure 2-9. Interference between waves from two point sources.



Consider spherical waves emanating from two point sources separated by a distance *a* (as illustrated in Figure 2-9. Let us consider the field at a point P at some considerable distance from both sources, such that its location is a distance *z* along the bisector between the two sources and distance *x* perpendicular to the bisector. The distance *z* is assumed to be much larger than the wavelength (this will let us ignore the differences between r_1 and r_2 in the *1/r* dependence of the amplitude terms but not in the phase (cosine) terms). The local field at P is

$$\begin{aligned} E(\text{at P}) &\approx (E_1/r_1)e^{i(\omega t - kr_1)} + (E_2/r_2)e^{i(\omega t - kr_2)} \\ &\approx (E_1/z)e^{i(\omega t - kr_1)} + (E_2/z)e^{i(\omega t - kr_2)} \end{aligned} \tag{2.43}$$

and obviously possesses two components. The intensity at point P is the square of this field and is seen to be

$$\begin{aligned} I(\text{at P}) &= \left|E(\text{at P})\right|^2 \\ &\approx (E_1/z)^2 + (E_2/z)^2 + \left(E_1E_2/z^2\right)e^{-ikr_1+ikr_2} \\ &\quad + \left(E_1E_2/z^2\right)e^{+ikr_1-ikr_2} \end{aligned} \tag{2.44}$$

which reduces to

$$I(\text{at P}) \approx \frac{1}{z^2}\left[E_1^2 + E_2^2 + E_1E_2\left(e^{-ik(r_1-r_2)} + e^{+ik(r_1-r_2)}\right)\right] \tag{2.45}$$

or

$$I(\text{at P}) \approx \frac{1}{z^2}\left[E_1^2 + E_2^2 + 2E_1E_2\cos\bigl(k(r_1 - r_2)\bigr)\right] \tag{2.46}$$

where we have use the complex function identity relations

$$e^{\pm iy} = \cos y \pm i\sin y \tag{2.47}$$

and

$$e^{iy} + e^{-iy} = (\cos y + i\sin y) + (\cos y - i\sin y) = 2\cos y\,. \tag{2.48}$$

If E_1 and E_2 are roughly equal in strength, then

$$I(\text{at P}) \propto 4E_1^2\cos^2\bigl(k(r_1 - r_2)/2\bigr) \tag{2.49}$$

where we have used the trigonometric identity

$$\cos 2y = 2\cos^2 y - 1\,. \tag{2.50}$$

If z is much greater than both the transverse distance x and the source spacing a then

$$\begin{aligned} r_1 - r_2 &= \left\{z^2 + \left[x - (a/2)\right]^2\right\}^{1/2} - \left\{z^2 + \left[x + (a/2)\right]^2\right\}^{1/2} \\ &\approx -ax/z \end{aligned}\,. \tag{2.51}$$

Therefore,

$$I(\text{at P}) \propto 4E_1^2\cos^2\left[\frac{\pi ax}{\lambda z}\right] \tag{2.52}$$

with maxima at $x = m\lambda z/a$ where m is an integer. We see that two-wave interference produces a sinusoidal spatial intensity pattern with peaks that are 4 times the average value of the intensity of a single source and valleys that drop to zero intensity. Interference between waves from more than three sources can produce complex patterns with the highest peaks being even larger than those from two sources.

Huygens' Principle and Diffraction

At any point in space, the electric field of an electromagnetic wave can be defined in terms of its vector field strength $\vec{E}$, angular frequency ω, propagation direction $\vec{k}$, and a phase ϕ

$$\vec{E}\left(\vec{r},t\right) = \vec{E}_0\, e^{i\left(\vec{k}\cdot\vec{r} - \omega t + \phi\right)} . \tag{2.53}$$

A wavefront of a monochromatic (single frequency ω) electromagnetic wave is any surface on which the phase is constant and equal.

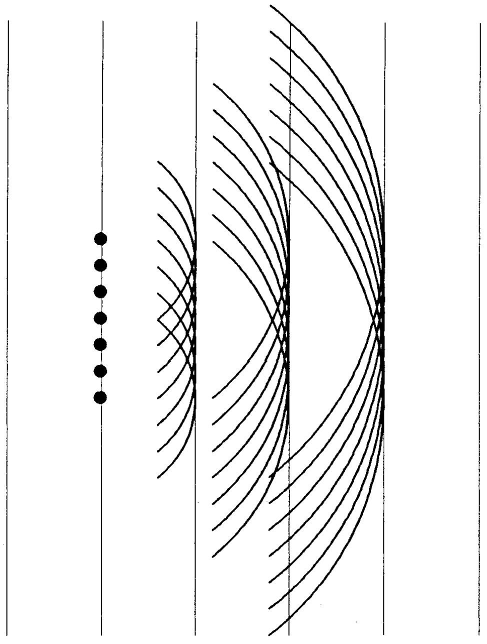
Huygens postulated that the propagation of electromagnetic waves could be analyzed by assuming that each point on a wavefront acts as a source of a spherical "wavelets" (i.e. a tiny spherical wave). The coherent summation (and interference) of the wavelets at some distance from the original wavefront generates a new wavefront. This wavelet generation followed by interference followed by wavelet generation can be continued to determine the wavefront at any distance given the wavefront at some plane. This is the essence of Huygens' principle.

Consider Figure 2-10. The parallel lines represent wavefronts of a plane wave separated by one wavelength. The dots represent some of an infinite number

of a plane wave.

of points on one of the wavefronts used to generate Huygens' wavelets. The wavefront of each wavelet is shown as it expands away from the source point. Each of these wavefronts is assumed to have the same phase as the others. Note further that the wavelets appear to add in phase along a line which coincides with one of the original plane wavefronts. If we look at the first point on either side of the center point, the angle away from the plane wave propagation direction is symmetric. Any tendency for one point to cause a change in direction is exactly counteracted by the opposite tendency resulting from the symmetric point. Since the line of points is infinite, for any point that would cause deviation from the plane wave character, there is a symmetric point that prevents that deviation.

Figure 2-10. Addition of Huygens' wavelets to reconstruct the wavefront



Mathematically, Huygens' principle can be represented as an integral over the input plane of a large number of spherical waves generated by the points of that plane. Specifically, we will represent an arbitrary electromagnetic wave as the real part of a complex function

$$\Xi\left(\vec{r},t\right) = \mathrm{Re}\left[\xi\left(\vec{r}\right)e^{i\omega t}\right] \tag{2.54}$$

For this input function, Huygens' integral is

$$\underset{\substack{\text{Output}\\ \text{Field}}}{\xi\left(\vec{r}\right)} \;=\; \underset{\substack{\text{Surface}\\ \text{Integral}}}{\frac{i}{\lambda}\iint\limits_{\substack{\text{input}\\ \text{plane}}} dx_0 dy_0} \quad \underset{\substack{\text{Input}\\ \text{Field}}}{\xi\left(\vec{r}_0\right)} \quad \underset{\substack{\text{Obliquity}\\ \text{Factor}}}{\frac{1+\cos\alpha}{2}} \quad \underset{\substack{\text{Spherical}\\ \text{Wavelet}}}{\frac{e^{ik\left|\vec{r}-\vec{r}_0\right|}}{\left|\vec{r}-\vec{r}_0\right|}} \tag{2.55}$$

where the individual terms have been specifically identified with α (in the obliquity factor) being the angle between the vector $\vec{r}-\vec{r}_0$ and the z-axis and k being the wavenumber

$$k \;=\; 2\pi/\lambda \;=\; 2\pi\omega/c\,. \tag{2.56}$$

The Huygens' integral correctly predicts many observed propagation phenomena. However, it is difficult to evaluate. It is common to use an approximation called the Fresnel approximation. Specifically we define $L = z - z_0$ and we assume $\cos\,\alpha \sim 1$. In this approximation,

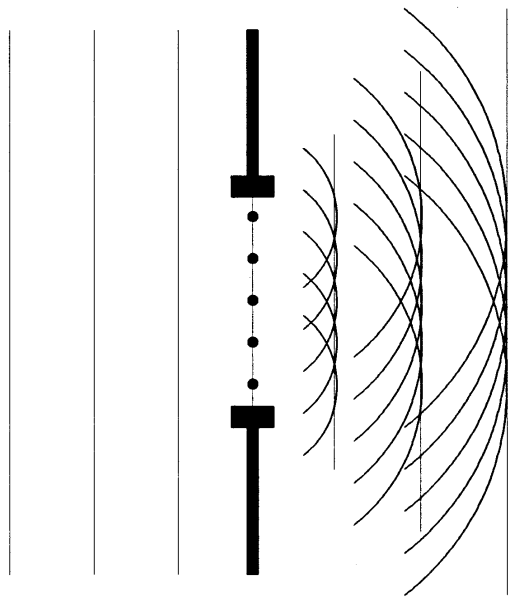
$$\begin{aligned} \left|\vec{r}-\vec{r}_0\right| \;&=\; \left[\left(x-x_0\right)^2+\left(y-y_0\right)^2+\left(z-z_0\right)^2\right]^{1/2} \\ &\approx\; L\left[1 \;+\; \frac{\left(x-x_0\right)^2}{2L^2} \;+\; \frac{\left(y-y_0\right)^2}{2L^2}\right] \end{aligned} \tag{2.57}$$

Substituting the Fresnel approximation into Huygens' integral leads to the Fresnel diffraction equation

$$\begin{aligned} \xi\left(x,y,z_0+L\right) = \frac{ie^{-ikL}}{L\lambda}\iint\limits_{\substack{\text{input}\\ \text{plane}}} dx_0 dy_0 \;\Big\{\xi\left(x_0,y_0,z_0\right) \\ \times\exp\left[\frac{-ik}{2L}\left[\left(x-x_0\right)^2+\left(y-y_0\right)^2\right]\right]\Big\} \end{aligned} \tag{2.58}$$

This equation can be explicitly calculated to determine how the field at any one plane will transform when it propagates to a distant plane.

Figure 2-11. Huygens' wavelets as an explanation of the origin of diffraction.



In general, diffraction calculations are the province of sensor design specialists. However, knowing that Huygens' principle leads to a mathematically correct diffraction integral makes it easier for us to visualize why diffraction occurs. Consider Figure 2-11. In this figure an aperture blocks all of the wavefront except for a small amount. Only inside the aperture is there any field to produce Huygens' wavelets. Wavelets produced near the center of the aperture behave much like there was no aperture present. They see a roughly symmetric situation with wavelets on one side of center seeing corresponding wavelets on the opposite side to cancel out their spherical expansion characteristics. Radiation corresponding to those wavelets continues to propagate in the general direction of the original wavefront. However, wavelets produced at the edge of the aperture see an asymmetric situation. Clearly, the wavelets which would cancel the spherical character are missing. Lacking the symmetric wavelets, radiation near the edges of the aperture will maintain its spherical character, allowing the wave to bend around the corner of the aperture. Those with "true faith" can even draw upon the fact that interference is the source of the success of Huygens' principle, to argue that there should be some sort of interference pattern superimposed on the general spreading. Indeed, there are interference effects, although they are not the simple two-source pattern calculated earlier.

A few cases of diffraction can be evaluated explicitly. One of these is diffraction by a uniformly illuminated circular aperture. Let *a* represent the radius of the aperture. The result of calculations which will not be repeated here, is that the field pattern at a considerable distance from the aperture is given by the Airy function [2]. That is,

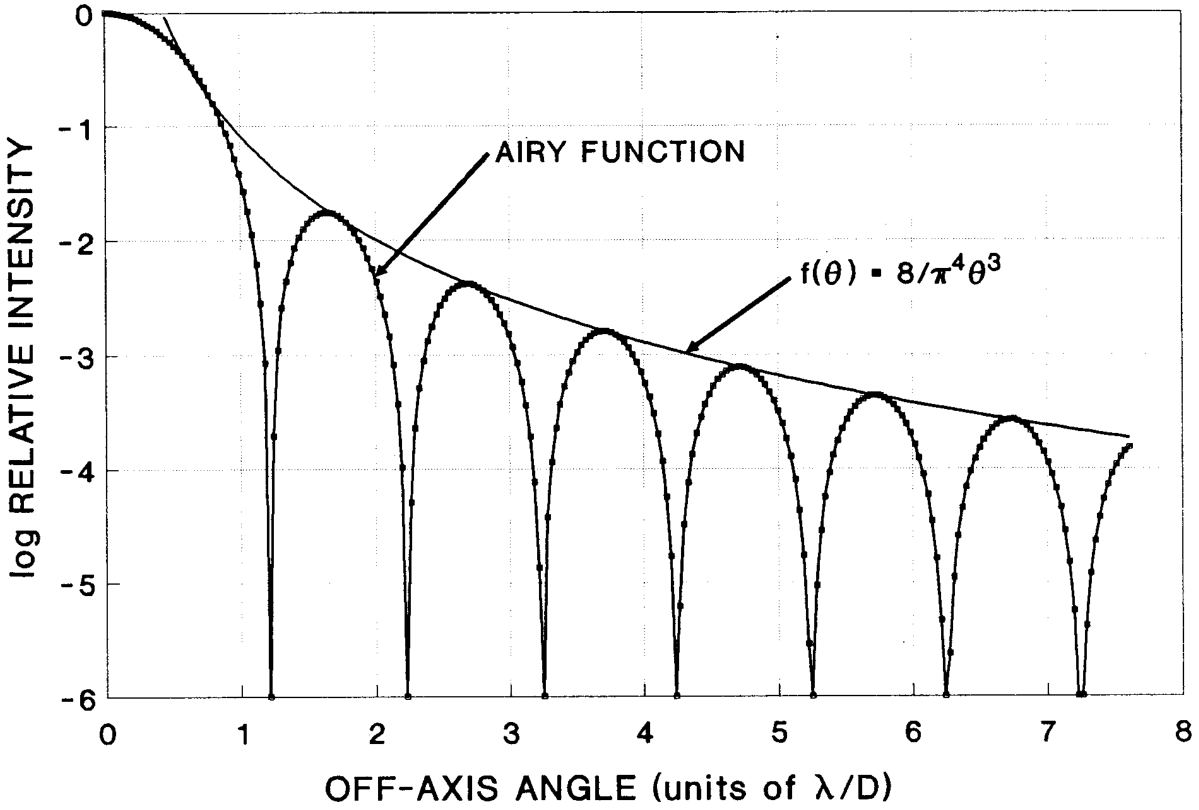
$$I(R,\theta') = \frac{P\pi a^2}{\lambda^2 R^2}\left[\frac{2\,\mathrm{J}_1\left(\pi(D/\lambda)\sin\theta'\right)}{\left(\pi(D/\lambda)\sin\theta'\right)}\right]^2 \tag{2.59}$$

where $\mathrm{J}_1(x)$ is the Bessel function of the first kind of order 1, $D=2a$ is the diameter of the aperture, λ is the wavelength of the illuminating radiation, and P is the total power incident on the aperture. The squared quantity in the rectangular brackets is normalized to unity at zero angle and also normalized to unity total area (the integral over all angles is unity). If we assume that the angle θ' is small then we may assume that $\sin\theta' \sim \theta'$. If we further define the angle θ is the angle θ' measured in units of the characteristic diffraction angle λ/D, then Eq. (2.59) can be rewritten as

$$I(R,\theta) = \frac{P\pi a^2}{\lambda^2 R^2}\left[\frac{2\,\mathrm{J}_1(\pi\theta)}{(\pi\theta)}\right]^2 \tag{2.60}$$

This function is shown graphically in Figure 2-12.

Figure 2-12. The Airy function – diffraction from a circular aperture.



As indicated in Figure-12 and shown photographically in Figure 2-13, the diffraction pattern is a bright central spot (or lobe) surrounded by concentric rings of rapidly decreasing intensity. The position of the minimum between the central lobe and the first ring is given by

$$\theta_{\min} = 1.220\lambda / D . \tag{2.61}$$

For reasons that will be discussed in Chapter 11, this angle is often associated with the limiting angular resolution of sensors with aperture *D* and operating at wavelength *λ*.

For values of *θ* > *1*, the Bessel function in Eq. (2.60) can be approximated by

$$J_1(x) = \sqrt{2 / \pi x}\ \cos(x - (3\pi / 4)) \tag{2.62}$$

and the envelope of the peaks can be determined by setting *cos x = 1*. This result shown here, occasionally finds utility in applications such as determining the angular distance over which a bright source like the sun might produce veiling glare in a sensor, or how fast the radar cross section of a circular disc falls off with off-normal reflectance angle. The envelope function *f(θ)* is explicitly illustrated in Figure 12-2 and is given by

$$f(\theta) = \left[\frac{2\,\mathrm{J}_1(\pi\theta)}{(\pi\theta)}\right]^2 = \left(\frac{2\sqrt{2/\pi\,\pi\theta}\,\cos(\pi\theta-(3\pi/4))}{\pi\theta}\right)^2$$

$$\xrightarrow{\text{Envelope}} \left(\frac{2\sqrt{2/\pi^2\theta}}{\pi\theta}\right)^2 = \frac{8}{\pi^4\theta^3} \qquad (2.63)$$

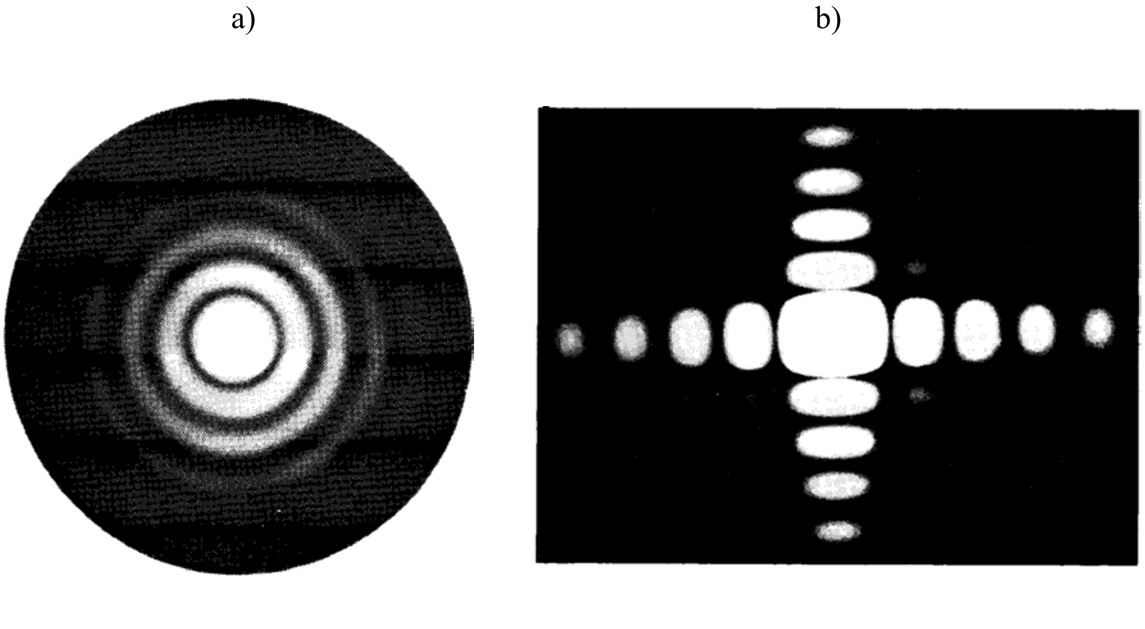
The intensity of the first bright ring is seen to be 17 dB (see Appendix A on units for a description of logarithmic units) below the intensity of the central lobe. One must move an angular distance at least *9.36 λ/D* before the sidelobe intensity falls below -40 dB.

The diffraction pattern from a rectangular aperture of horizontal side *2a* and vertical side *2b* is given by the function

$$I(R,\theta,\phi) = \frac{P4ab}{\lambda^2 R^2}\left(\frac{\sin(\pi(2a/\lambda)\sin\theta)}{\pi(2a/\lambda)\sin\theta}\right)^2\left(\frac{\sin(\pi(2b/\lambda)\sin\phi)}{\pi(2b/\lambda)\sin\phi}\right)^2 \qquad (2.64)$$

where θ is the angle with respect to the horizontal and ϕ is the angle with respect to the vertical. We will see the function {*(sin x)/x = sinc x*} time and again. It will be especially prominent when we discuss radar cross sections. A photograph of the diffraction pattern from a rectangular aperture is compared with the diffraction pattern from a circular aperture in Figure 2-13.

Figure 2-13. Photographs of the diffraction patterns from a a) circular aperture and b) rectangular aperture.[2]

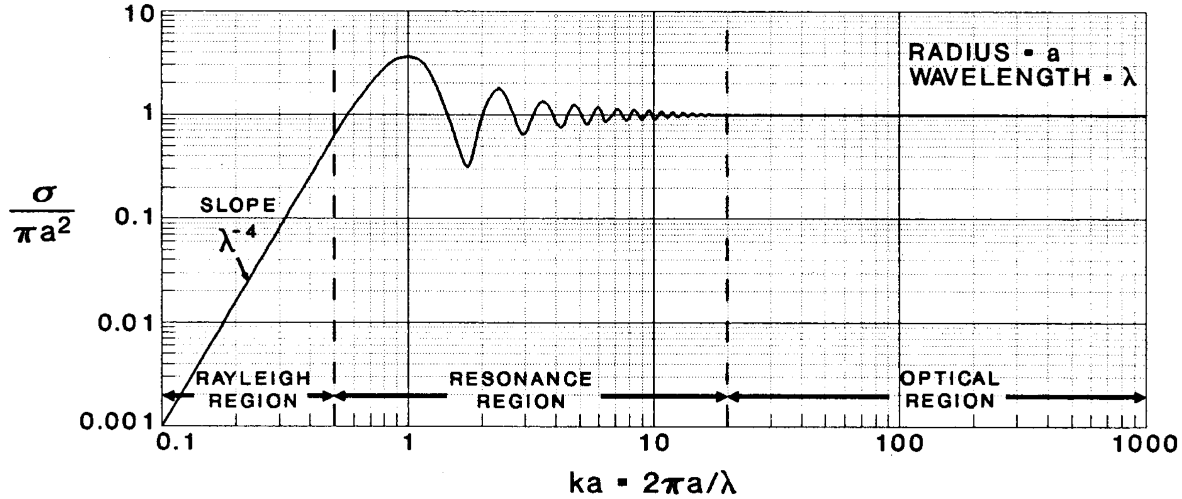


Mie Scattering

Mie scattering is the scattering of electromagnetic radiation by spherical particles having a dielectric constant and/or a conductivity which differs from the medium surrounding the particle [4]. Calculation of this phenomena is fairly complex and will not be described here. However, the results of such calculations are extremely interesting and can be qualitatively applied to understanding many phenomena encountered in the propagation of electromagnetic radiation. First, we must understand the concept of cross section. Assume that a plane wave of radiation is incident on a unit area. If we place a scatterer (an object, or in the Mie scattering problem, a small spherical particle), then part of the radiation will be blocked by the particle. This blocked radiation is either absorbed or it is scattered into all directions. Due to diffraction, a particle may actually block more or less area than the physical cross sectional area of the particle. The scattering cross section of an object is the apparent (as opposed to physical) size of the object based on the fraction of the incident radiation that is affected. Let the unit area be 1 cm^2. If the physical cross sectional area of the sphere is $\pi a^2 = 0.0001$ cm^2 but scatters 1% of the incident radiation, then the total scattering cross section is 0.01 x 1 cm^2 = 0.01 cm^2 (100 times the real physical cross section). If the fractional scattered radiation is measured as a function of angle, it is possible to determine a differential cross section (usually given units of cm^2/sr).

The most commonly presented Mie scattering result is a graph of the total scattering cross section as a function of wavelength of the scattered radiation. Figure 2-14 shows the Mie scattering cross section of a perfectly conducting metallic sphere of radius *a*. The calculated curve is normalized to the physical spherical cross section πa^2. The abscissa of the plot is $ka = 2\pi a/\lambda$. There are three obviously different physical regions. When $ka > 20$, that is, when the wavelength is much smaller than the particle radius, the scattering cross section is exactly equal to the physical cross section. This is usually referred to as the optical region. Diffraction is a minor contributor in the optical region.

Figure 2-14. Normalized Mie scattering cross section of a perfectly conducting metallic sphere.



When $ka < 0.5$, that is when the wavelength is much larger than the particle radius, the scattering cross section falls off as the inverse fourth power of the wavelength ($\sigma \propto \lambda^{-4}$). This region is known as the Rayleigh region. In between the Rayleigh and optical regions, the cross section oscillates up and down about the optical value πa^2. This region is called the resonance region. The oscillations are due to interferences between waves which are scattered from the front surface of the sphere and waves which are diffracted completely around the sphere.

This qualitative shape of the Mie curve (Rayleigh region, resonance region, and optical region) will be valid for virtually any spherical particle. However, the exact shapes for weakly conducting, non-conducting (dielectric), and/or absorbing particles will vary depending on the real and imaginary parts of the refractive index and the conductivity. The strength and number of oscillations are strongly affected by the material characteristics. The presence of a strong absorption in a region of the spectrum will cause significant dips in the cross section in that region. Furthermore, clouds of particles will have particle of different radii. In this case the appropriate Mie curve must be averaged over the distribution in sizes (radius a). The resulting averaged curve will usually not exhibit significant resonance oscillations (larger particles will have their scattering peaks at the minima of smaller particles). However, unless the range of particle sizes extends over many orders of magnitude, the average curve will still have pronounced Rayleigh and optical regions with a smooth shoulder replacing the resonance region. Even non-spherical particles will have scattering curves that possess Rayleigh regions, optical regions, and resonance regions. However, the non-spherical particle curves will exhibit significant polarization dependences and have different angular distributions of scattered radiation. Even complex objects such as ships or airplanes will have a flat region in a σ versus $1/\lambda$ plot that corresponds to an optical region (wavelength small compared to any characteristic scale length of the target) and to λ^{-4} region that corresponds to the Rayleigh region (wavelength large compared to the maximum dimension of the target).

The virtual indifference of the qualitative Mie curve to external conditions allows its use in many different situations. We shall see that it can be used to describe the attenuation of electromagnetic radiation by rainfall and by clouds and fogs from the microwave region through the visible region. It can be used to explain why the sky is blue. Air molecules act as dielectric particles that are small compared to visible light. Thus scattering occurs in the Rayleigh region. Blue light (small wavelengths) from the sun is thus more strongly scattered than red light (large wavelengths). Mie scattering can also explain why most military smokes are not very effective in the infrared. We will go into more detail on these and other phenomena later in this volume.

References

[1] King James Version, The Holy Bible (World Publishing Co., Cleveland OH).

[2] Born, Max and Wolf, Emil, Principles of Optics 6th Ed. (Cambridge University Press, Cambridge UK, 1980) pp.36-47 (Fresnel reflection), 370-401 (diffraction).

[3] Hale, G. M. and Querry, M. R., “Optical Constants of Water in the 200 nm to 200 um wavelength region,” *Applied Optics*, 12, 555-563 (1973).

[4] Van de Hulst, H. C., Light Scattering by Small Particles (Dover Publications, New York NY, 1981).

Problems

2-1. From what does the existence of electromagnetic radiation follow as an inevitable outcome?

2-2. What functional form do solutions of the equation

$$\frac{d^2\xi}{dx^2} - \frac{1}{a^2}\frac{d^2\xi}{dt^2} = 0 \tag{2.65}$$

take? The quantity ξ represents any of a number of real physical quantities which will not be delineated here. Equation (2.65) is one of the classic equations of physics. What name is given to equations of the form of Eq. (2.65)?

2-3. You need to find a window material for a dual mode seeker operating at 10 µm and at 95 GHz. Gallium Arsenide has a complex refractive index of 3.277 + i6.1x10^{-7} at 9.6 µm and 3.606 + i0.0012 at 95 GHz. What is the absorption coefficient of gallium arsenide at 9.6 µm? At 95 GHz? What is the loss tangent at 95 GHz. Windows are nominally 1 cm thick. From an absorption perspective only is gallium arsenide likely to be a good window material.

2-4. The refractive indices before and after an interface between two media are n_1 and n_2, respectively. If the angle of incidence is 45°, derive an expression for the reflection coefficient (E_R^2/E_0^2).

2-5. What theoretical construct is used to derive diffraction and interference relations?

2-6. Two point sources of electromagnetic radiation are separated by a distance a. If the intensity of one source (#1) is twice that of the second source (#2), derive an expression for the interference pattern at an arbitrary distance. Use the geometry of Figure 2-9.

2-7. Using Huygens' principle and the empirical diffraction patterns for the circular and rectangular apertures as guides, estimate (sketch) the diffraction pattern expected at some distance from a knife edge. A knife edge can be assumed to be an infinite half-plane that blocks radiation propagation next to an empty infinite half-plane.

2-8. A radar reflecting target is made from a spherical, metallic-coated balloon. The diameter of the balloon is 20 cm. What is a good estimate of the scattering cross section σ at a frequency of 300 MHz? At a frequency of 3 GHz? At a frequency of 30 GHz? Which of these estimates is most likely to be in error?

CHAPTER 3

PROPAGATION OF ELECTROMAGNETIC RADIATION. II. – WEATHER EFFECTS

Introduction

The characteristics of the propagation medium have the most profound influence on sensor performance. We begin our studies with the important case of sensors employing propagation of electromagnetic radiation through the atmosphere. To fully understand the performance of such sensors we must understand the constituents of the atmosphere, the structure of the atmosphere, and how both the constituents and structure varies with time and location. This requires at least a minimal understanding of both meteorology (weather) and climatology.

The most profound impact of the atmosphere on general sensor performance comes through its ability to attenuate electromagnetic signals. The **attenuation** (often called **extinction**) consists of two parts: absorption and scattering. **Absorption** is the conversion of electromagnetic energy into internal excitations of the atmospheric constituents or into heat. **Scattering** is the alteration of the direction of propagation of a unit of electromagnetic energy without any significant reduction in the magnitude of that energy. Extinction (both absorption and scattering) tends to reduce the strength of signals that sensors are trying to detect. Scattering (especially of other sources of electromagnetic radiation) tends to affect the noise that tends to make detection of the signal more difficult. Therefore both scattering and absorption are important to electromagnetic sensor systems. It is important to remember that

$$EXTINCTION \ = \ SCATTERING \ + \ ABSORPTION$$

Knowledge of two of the three components allows calculation of the third.

Before proceeding to address the details of extinction, it is useful to introduce a concept that has great practical utility and is a baseline input to many extinction calculation. **Visibility** is the greatest range at which an unaided human observer can detect objects of interest. It has two working definitions. In daytime, visibility is the greatest range at which a very dark object can be detected against the sky at the horizon. At night, visibility is the greatest range at which an unfocused, moderately intense light source can be detected. In practice, visibility is estimated by looking at appropriate objects with known ranges. The visibility is somewhere between the range of the farthest object you can see and the closest object you can't see.

To place visibility into a more scientific perspective, the meteorological community defined a related term, the **meteorological range** (or sometimes the **meteorological visibility range**), V. It is defined by the Koschmieder formula [1]

$$V = \frac{1}{\beta} \ln \frac{1}{\varepsilon} = \frac{3.912}{\beta} \qquad (3.1)$$

where β is the scattering coefficient at 550 nm wavelength and ε is a threshold contrast for detection. For human visual detection it has been found that ε is approximately 0.02. For most aerosols in the visible, absorption is negligible and β is usually approximated by the extinction coefficient.

Visibility is a subjective measure; meteorological range is quantitative. Visibility is easily measured without expensive instruments and is reported in routine weather measurements. Meteorological range requires expensive instruments to measure. It is commonly used as an input into atmospheric propagation computer codes. The two are often (and incorrectly) used interchangeably. If only an observed visibility, V_{obs}, is available, the meteorological range can be estimated by

$$V = \left(1.3 \pm 0.3\right) \times V_{obs} \ . \qquad (3.2)$$

Meteorological range has been grouped into 10 international visibility codes shown in Table 3-1. [2]

Table 3-1. International Visibility Codes

CODE NUMBER	WEATHER CONDITION	METEOROLOGICAL RANGE (km)	SCATTERING COEFFICIENT (km^{-1})
0	Dense Fog	< 0.05	> 78.2
1	Thick Fog	0.05 - 0.20	78.2 - 19.6
2	Moderate Fog	0.2 - 0.5	19.6 - 7.82
3	Light Fog	0.5 - 1.0	7.82 - 3.91
4	Thin Fog	1.0 - 2.0	3.91 - 1.96
5	Haze	2.0 - 4.0	1.96 - 0.954
6	Light Haze	4.0 - 10.0	0.954 - 0.391
7	Clear	10.0 - 20.0	0.391 - 0.196
8	Very Clear	20.0 - 50.0	0.196 - 0.078
9	Exceptionally Clear	> 50.0	< 0.078
–	Pure Air	277	0.0141

Many atmospheric constituents contribute to atmospheric scattering and/or absorption. These are summarized in Table 3-2. Molecules are extremely small "particles". It is not unreasonable to expect that since the characteristic dimensions "*a*" are very small (atmospheric molecules are of the order of one nanometer in dimension) compared to the wavelength (hundreds of nanometers), that molecules would scatter radiation similar to the Rayleigh region of the Mie scattering curve. This is indeed observed and is called **Rayleigh scattering**. This is a very weak effect. The extinction coefficient for Rayleigh scattering at 550 nm in a sea-level atmosphere is only 0.0141 km^{-1}. The corresponding visibility is 277 km. All molecules contribute to Rayleigh scattering. The sky is blue because of Rayleigh scattering. Due to the $1/\lambda^4$ dependence of the scattering strength, blue light at 400 nm is scattered 13 times more strongly than deep red light at 760 nm. Since the scattering during the middle of the day (when the sun is well above the horizon) is not strong enough to completely attenuate the blue component, everywhere we look there is an equally strong blue excess over red. Sunsets tend to be red, because at the low grazing angles at sunset allow a propagation path long enough for all of the blue light to be attenuated. Only the red light remains.

Some molecules contribute to absorption. It is useful to distinguish between different kinds of atmospheric molecules. Seven gases (O_2, N_2, CO_2, Ar, CH_4, N_2O, and CO) constitute the **uniformly mixed gases**. Uniformly mixed means that the relative proportions of each gas are essentially independent of altitude or geographic location. Of the uniformly mixed gases, two have very significant absorptions. Oxygen (O_2) has two strong absorption lines in the microwave region and one strong absorption line at the red edge of the visible region. Carbon dioxide (CO_2) absorbs at a large number of wavelengths throughout the infrared region.

Two abundant gases are not uniformly mixed. As we shall see, water vapor (H_2O) is concentrated near the earth's surface, while ozone has a maximum concentration in the stratosphere. Water vapor absorbs at many places throughout the infrared and microwave regions. Ozone has limited absorption in the infrared, but is an intense absorber of ultraviolet radiation in the region from 200-300 nm. The significance of the absorption of both of these molecules will be discussed in detail later.

There are many **trace gases** (twenty of which are of sufficient significance to warrant inclusion in atmospheric extinction models and are explicitly named in the table). Many of these trace gases are the result of natural processes: photochemical degradation of organic chemical vapors emitted by plants, decay and decomposition of dead organic matter, emission from volcanic (or other geothermal) events, and production by lightning or solar wind radiation interaction with more abundant atmospheric species. A few are characteristically human industrial pollutants. None of the trace gases have absorption strong enough to influence any sensor that isn't designed to explicitly detect those trace gases. The same is true of artificial contaminant vapors (such as chemical warfare agents or HAZMAT).

In the atmosphere there is a special type of molecular absorption that is not commonly encountered elsewhere, the so-called **continuum absorption**. Continuum absorption results when collisions between molecular species alter the energy levels of those molecules in such a way that

Table 3-2. Potential contributors to atmospheric extinction.

MOLECULAR SCATTERING (RAYLEIGH SCATTERING)				
MOLECULAR ABSORPTION				
- UNIFORMLY-MIXED GASES	– O_2	N_2	CO_2	Ar
	CH_4	N_2O	CO	
- WATER VAPOR (H_2O)				
- OZONE (O_3)				
- TRACE GASES	– NO	SO_2	NO_2	NH_3
	HNO_3	OH	HF	HCl
	HBr	HI	ClO	OCS
	H_2CO	HOCl	HCN	H_2O_2
	CH_3Cl	C_2H_2	C_2H_6	PH_3
- CONTAMINANTS	– VOLATILE ORGANIC COMPOUNDS			
	– CHEMICAL WARFARE AGENTS			
	– POLLUTANTS			
CONTINUUM ABSORPTION (COLLISION-INDUCED ABSORPTIONS)				
- WATER VAPOR				
- NITROGEN				
PARTICULATE SCATTERING AND ABSORPTION				
- BOUNDARY LAYER PARTICULATES	– HAZE AEROSOLS (RURAL, URBAN, MARITIME, DESERT)			
	– FOG			
	– ICE FOG			
	– DUST			
	– SMOKE			
	– ARTIFICIAL OBSCURANTS			
- UPPER TROPOSHERIC "AEROSOLS"				
- STRATOSPHERIC "AEROSOLS"	– METEORITIC DUST			
	– VOLCANIC DUST			
- CLOUDS				
PRECIPITATION (HYDROMETEORS)				
- DRIZZLE				
- RAIN				
- SNOW				
- GRAUPEL				
- HAIL				
- SLEET				

absorptions now occur where in the absence of collision no absorptions would normally occur. As two molecules collide and subsequently move apart, the energy levels morph continuously from their isolated values to some altered collisional values and back again. Thus, the collisionally altered absorptions are very much broader than the absorptions of the isolated molecules, hence the use of the adjective "continuum". It should be noted that collisions can induce absorptions where no absorption is possible in the isolated molecule.

Two atmospheric molecules have significant continuum absorption: water vapor and nitrogen. Water molecules interact strongly with other water molecules, but interact only weakly with nitrogen and oxygen. Thus, water-water collisions have a strong collisional absorption. Since the rate at which water molecules collide with other water molecules is proportional to the square of the partial pressure (molecular density) of water vapor, one would expect that the importance of continuum absorption to be much higher when the partial pressure of water vapor is high (humid, sea level) than when it is low (dry, moderate altitudes and above). This is indeed observed. The strong collisional dependence of absorption of water dramatically manifests itself in the absorption of liquid water. We have already seen that liquid water is virtually opaque everywhere except for a narrow window in the blue-green region. Nitrogen is a molecule that at low pressure has no visible, infrared, or microwave absorption. However, there is a collision-induced absorption between 4 μm and 5 μm that must be incorporated in predictions of infrared attenuation.

Isolated molecules are not the only constituents of the atmosphere. Particulates can be found at all altitudes and geographic locations. Particulates small enough that they are at least partially subject to Brownian motion and remain airborne for extended periods of time are often referred to as aerosols. Aerosol has a technical definition meaning an airborne suspension of a liquid or solid. Atmospheric aerosols may contribute to absorption in some spectral regions (depending on their chemical constitution) and will contribute to scattering in all regions with wavelengths comparable to or smaller than the particle size. The scattering behavior can usually be described as "Mie-like".

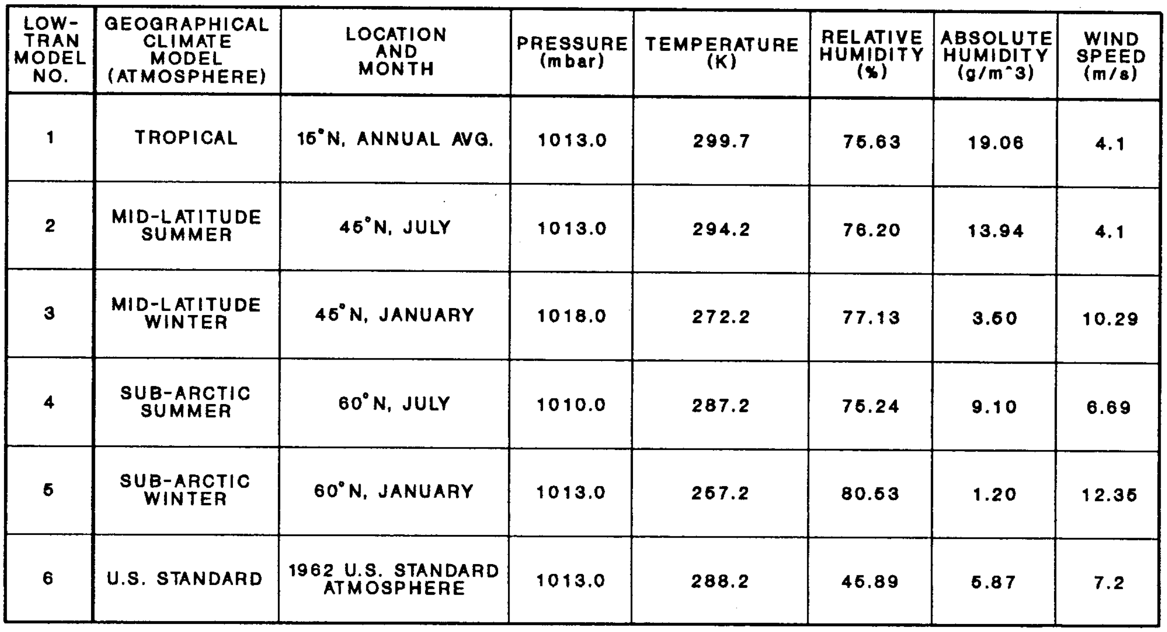
In the lower few kilometers of the atmosphere (the boundary layer) we encounter haze aerosols, fogs, smoke, dust, and artificial obscurants (military smokes). At higher altitudes there are clouds and volcanic dust. At very high altitudes there is a steady influx of meteoritic dust. Hazes generally produce visibilities of a few kilometers. Fogs can produce visibilities of 100 m or less, as can clouds. Smoke and dust from distant sources can be so thin as to produce only mild reductions in visibility or so dense that visibilities are measured in single-digit meters. Even volcanic dust has high variability. Those who remember the eruption of Mt. St. Helens in Washington or Mt. Pinatubo in the Philippines both of which dumped inches of falling ash over thousands of square miles of territory, can attest to the worst cases.

A final contributor to atmospheric extinction is the class of hydrometeors. Hydrometeor means "water in the air". Technically, it includes fogs and clouds. However, from a practical perspective the term is often associated with "falling water" or precipitation. There is drizzle (liquid water with drop sizes less than 0.5 mm), rain (liquid water with drop sized greater than 0.5 mm), snow (crystalline solid water), hail (large balls of frozen water formed by rain being cycled up and down between high liquid content regions and freezing regions of a cloud), graupel (small pellets of snow formed by partial melting and refreezing of snow), sleet (mixed rain and snow), among many other forms. The primary characteristic of hydrometeors is their large size (they tend to be large enough to fall with a steady terminal velocity). As a result their effects extend much farther into the microwave region than aerosol effects. The magnitudes of hydrometeor effects depend strongly on the quantity of water (converted to an equivalent liquid water precipitation rate) that falls and on the type of hydrometeor (rain does not produce as large an effect as does snow for equivalent liquid water precipitation rate).

Structure of the Atmosphere

The Earth's atmosphere is not uniform in either space or time. There are very distinct variations in atmospheric characteristics as one moves from one pole towards the equator and on to the other pole. There are variations as one moves from coastal regions to interior regions of the continents and as one moves from lowlands to mountainous areas. There are major variations as the seasons change from summer to fall to winter to spring and back to summer again. The foremost atmospheric attenuation propagation code, LOWTRAN, has defined six atmospheric models that cover several of regions of interest at their maximum seasonal differences. [3]-[11] These are summarized in Table 3-3. These six models will show up in virtually all of the charts that follow in this section.

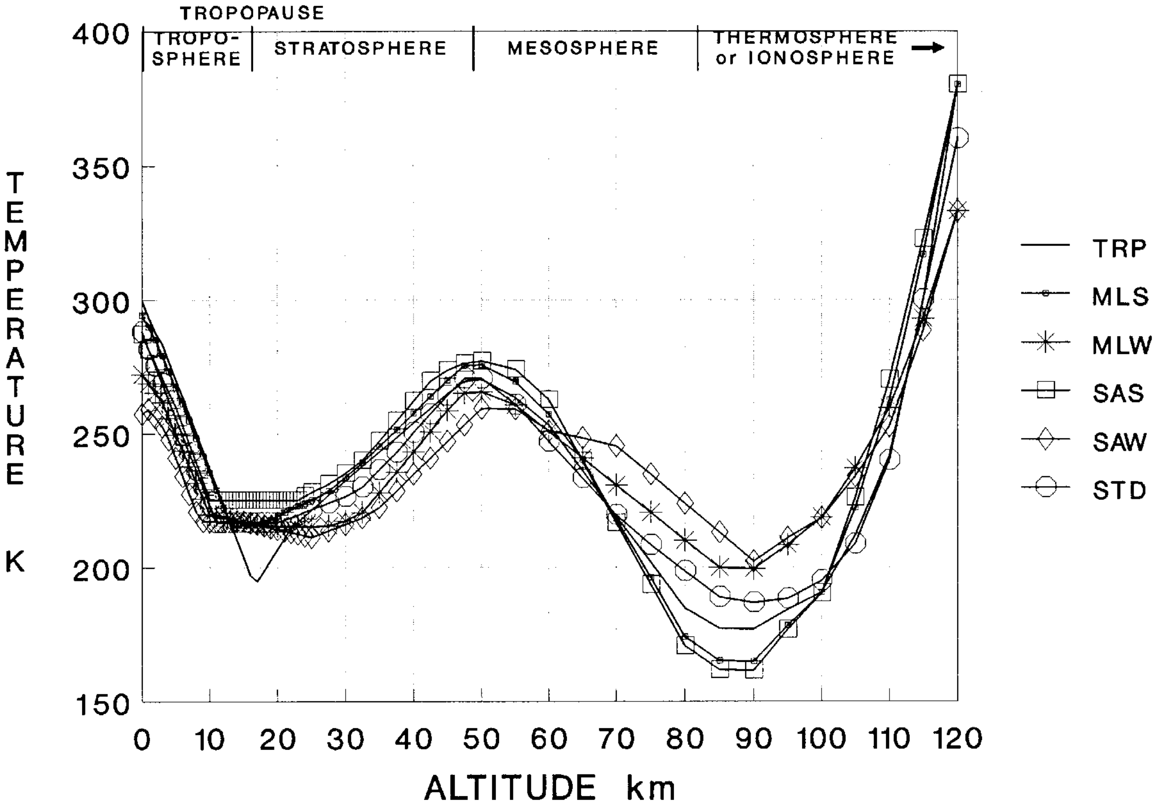
Table 3-3. Ground-level characteristics of LOWTRAN climate models.[11]



LOW-TRAN MODEL NO.	GEOGRAPHICAL CLIMATE MODEL (ATMOSPHERE)	LOCATION AND MONTH	PRESSURE (mbar)	TEMPERATURE (K)	RELATIVE HUMIDITY (%)	ABSOLUTE HUMIDITY (g/m^3)	WIND SPEED (m/s)
1	TROPICAL	15°N, ANNUAL AVG.	1013.0	299.7	75.63	19.06	4.1
2	MID-LATITUDE SUMMER	45°N, JULY	1013.0	294.2	76.20	13.94	4.1
3	MID-LATITUDE WINTER	45°N, JANUARY	1018.0	272.2	77.13	3.50	10.29
4	SUB-ARCTIC SUMMER	60°N, JULY	1010.0	287.2	75.24	9.10	6.69
5	SUB-ARCTIC WINTER	60°N, JANUARY	1013.0	257.2	80.53	1.20	12.35
6	U.S. STANDARD	1962 U.S. STANDARD ATMOSPHERE	1013.0	288.2	45.89	5.87	7.2

Figure 3-1 show the variation of temperature with altitude over the range 0 - 120 km for the six LOWTRAN climate models. All of the models exhibit the same qualitative behavior. From sea level the temperature falls steadily with increasing altitude until roughly 10 km (possibly as high as 20 km) altitude. There may or may not be a thin layer in which the temperature is virtually constant with increasing altitude. Next the nominal downward trend reverses and the temperature increases with increasing altitude until roughly 50 km. The trend reverses a second time and the temperature falls again with increasing altitude until roughly 85 km at which point the trend reverses for a third time. The temperature then increases rapidly with altitude reaching extremely high values at altitudes above 120 km.

Figure 3-1. Atmospheric temperature as a function of altitude. The divisions of the atmosphere derive from the characteristic temperature variations.[11]



It should be noted that the term "temperature" used here is based on the average kinetic energies of the particles. At each altitude the particles are assumed to be in local thermodynamic equilibrium. The particle speeds are thus given by the Maxwell speed distribution law

$$p(v) = 4\pi \left(\frac{M}{2\pi kT} \right)^{3/2} v^2 e^{-Mv^2/2kT} \tag{3.3}$$

where M is the mass of the particle and v is its speed. If v_{av} is the average particle velocity, then the temperature can be defined as

$$T = \pi\, Mv_{av}^2 / 8k = \pi\, E_{av} / 2k \tag{3.4}$$

where $E_{av} = Mv_{av}^2/2$ is the average particle kinetic energy. The human sensation of temperature depends not only on the average particle energy, but also on the rate of energy transfer to/from our bodies. This depends on the density of particles (pressure) as well. Even a very hot "temperature" would feel cold to our skin at the low pressures of the upper altitudes.

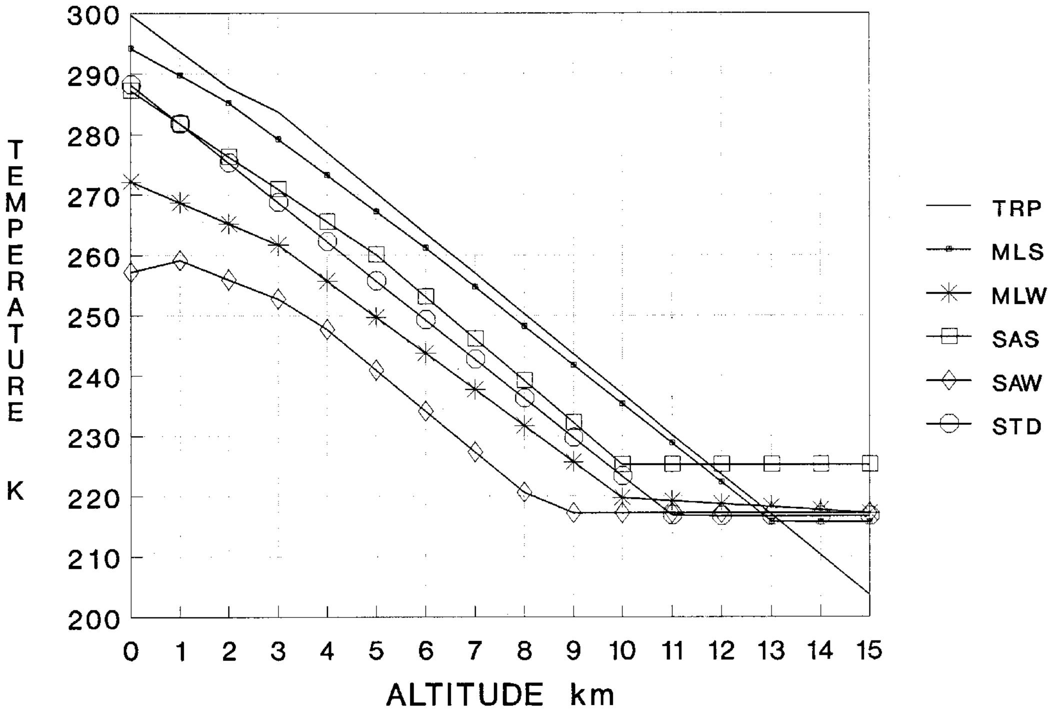
The temperature dependence described above has significance because it form the basis for dividing the atmosphere into discrete layers (as indicated very approximately in Figure 3-1). The lowest layer is the troposphere. The temperature variations that commonly occur in the troposphere are shown in Figure 3-2. The average tropospheric decrease in temperature is 6.5 K per km. This is true over most of the earth. However, as indicated by the curve for sub-arctic winter, there is usually a thin (1-2 km) inversion layer over snow-covered ground. It should be noted that Figures 3-1 through 3-9 in this section were plotted using data extracted directly from the LOWTRAN7 source code.[11]

The troposphere extends from sea level to the tropopause at an altitude typically between 10 and 20 km. The tropopause is defined as the altitude at which the decrease in temperature ceases and is replaced by either an inversion (increase in temperature with altitude) or by a neutral layer (temperature constant with altitude). The tropopause marks the lower edge of the stratosphere, the second layer extended from the tropopause to roughly 50 km altitude. The top of the stratosphere is called the stratopause and is the point at which temperature inversion ceases and the decrease in temperature with altitude increase resumes. Above the stratosphere is the layer referred to as the mesosphere that extends from the stratopause to roughly 80-90 km. The top of the mesosphere is the mesopause which is the altitude at which the temperature inversion resumes. Above the mesopause is the outer layer of the atmosphere called the thermosphere or the ionosphere.

Distinctly different phenomena are associated with each layer. Most of the phenomena that are associated with weather occur in the troposphere. Virtually all fogs, clouds, and precipitation, as well as all forms of wind storms (cyclones, anti-cyclones, and tornados) occur in the troposphere. Winds in the troposphere tend to increase as altitude increases. The jet streams are narrow bands of extremely strong winds that exist at high altitudes below the tropopause. Except in the vicinity of the jet streams and large cyclonic storms, the temperature inversion tends to prevent the troposphere and the stratosphere from having much interaction or exchange of materials. Most phenomena that affect sensor performance exist in the troposphere. However, a few important ones do not. As we shall see shortly, most of the "ozone layer" lies within the stratosphere. The same is true of an occasionally significant dust layer produced by volcanic eruptions. The ionosphere, by virtue of having a significant ionization, will affect the propagation of radio waves. The increase and decrease of this ionization as the sun rises and sets will also produces diurnal changes in the earth's magnetic fields.

The temperature decrease with increasing altitude in the troposphere is called the **lapse rate** [12]. Any increase in temperature with increasing altitude has an inverted lapse rate and is called a **temperature inversion**. The lapse rate is whatever temperature profile actually exists. The global average lapse rate is roughly 6.4 ^{o}C per km altitude (2 ^{o}C per 1000 feet). It has no direct connection to the **adiabatic lapse rate**. The adiabatic lapse rate is the change in temperature that a mass of air would experience as it increased in altitude without being able to exchange heat with its surroundings. Heating or cooling depends only on the expansion or contraction of the gas as the altitude changes.

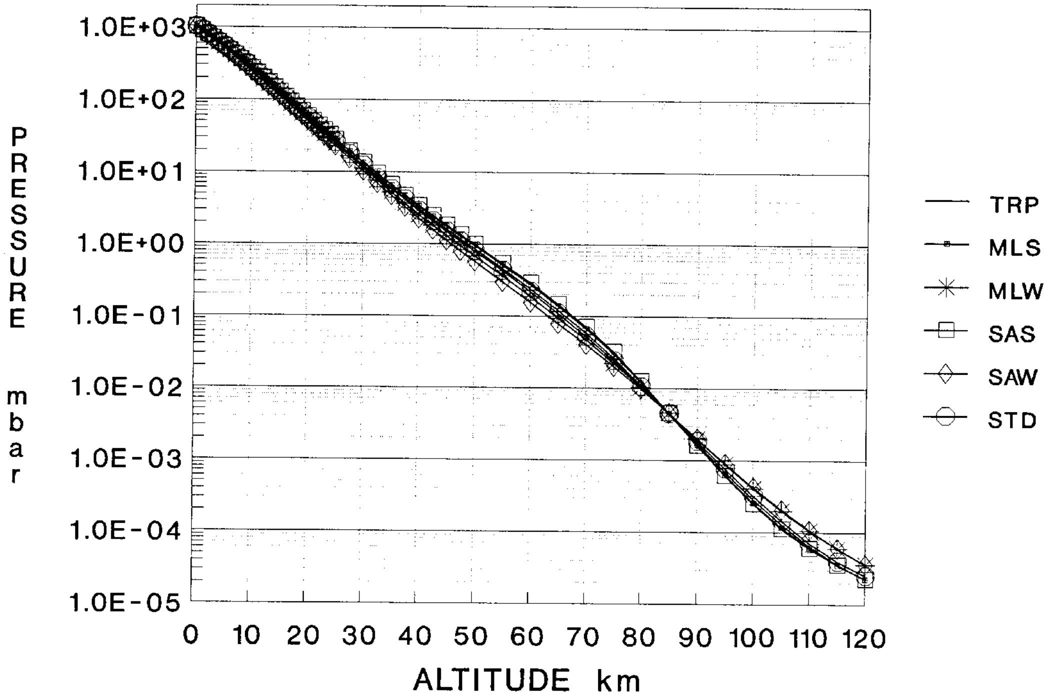
Figure 3-2. Temperature dependence of the troposphere.[11]



The adiabatic lapse rate is roughly 9 °C per kilometer (3 °C per 1000 feet) for dry air. The atmospheric lapse rate is usually less than the adiabatic rate because the atmosphere is constantly in motion. Turbulence, precipitation, and convection all lead to mixing of energy content from one altitude to another. Such mixing processes violate the adiabatic assumption. The adiabatic lapse rate is the temperature change in a volume of gas brought about by vertical motion. If the adiabatic lapse rate is greater than the actual lapse rate (for example, $T_{air} = T_{gas}$ = 30 °C at sea level, T_{air} = 24 °C and T_{gas} = 21 °C at 1000m), a rising volume of hot air will cool faster than the surrounding air. Its buoyancy will therefore be less than the surrounding air and the air volume will fall. This is a stable atmospheric condition. If the adiabatic lapse rate is less than the lapse rate of the surrounding air (for example, $T_{air} = T_{gas}$ = 30 °C at sea level, T_{air} = 18 °C and T_{gas} = 21 °C at 1000m), a rising volume of air will cool slower than the surrounding air. Its buoyancy will be higher than the surrounding air, and it will continue to rise. This is an unstable condition. The atmosphere is stable on the average. However, solar heating of the earth's surface (and the air in contact with it) or the forced flow of cold air masses over warmer air masses can both produce unstable atmospheric condition. Unstable conditions affects the local weather and, as we shall see later, they affect the employment of military obscurants (smoke) and chemical and biological agents.

The pressure variation with altitude is shown in Figure 3-3. Pressure does not exhibit much variation according to locale and tends to fall off exponentially with altitude. The exponential scale

Figure 3-3. Atmospheric pressure as a function of altitude.[11]



height is roughly 6.5 km. That is, the pressure falls off as

$$p(H \text{ in km}) = p(0)e^{-H/6.5} \tag{3.5}$$

H is the altitude. The pressure at 15 km is roughly 10% of the pressure at sea level. It can be safely assumed that the uniformly-mixed gases (O_2, N_2, CO_2, etc.) maintain their relative abundances at altitudes up to 80 km (the bottom of the ionosphere). Their absolute abundances will scale exponentially with altitude exactly as the pressure.

Several other important species do not vary in such well-behaved fashion. Figures 3-4 and 3-5 show the altitude dependence of water vapor concentration. In the troposphere water vapor decreases by four orders of magnitude as the altitude varies from sea level to only 12 km. That is, 99.99% of all atmospheric water vapor lies below 12 km altitude; 90% lies below 3 km altitude. The implication of this is that if a sensor has its performance limited by water vapor for sea level operation, then the performance of that sensor at high altitudes may not suffer the same limitations – there is no water vapor at high altitudes. The same sensor operated at sea level may be useless against surface targets, but have acceptable performance against high altitude targets because of the differences in integrated concentration along the slant path versus the horizontal path.

Another molecular species whose concentration does not exponentially decrease with increasing altitude is ozone. The altitude dependence of ozone concentration is shown in Figures

3-6 and 3-7. There is a significant concentration of ozone at sea level. Some but not all of this is man-made. A considerable amount is produced by atmospheric electrical activity. The majority, however, is produced in a thick layer surrounding the tropopause. This layer is commonly referred to as the ozone layer. The ozone layer is roughly 10 km thick and centered around 20 km altitude. In the ozone layer, solar ultraviolet radiation breaks atmospheric molecular oxygen into oxygen atoms. When one of these atoms collides with an oxygen molecule, an ozone molecule is formed. Ozone has the useful property of strongly absorbing ultraviolet radiation between 200 nm and 300 nm. Such radiation is extremely harmful to life causing cell death and genetic mutations. It is also a potent skin carcinogen. Without this layer, life on the earth's surface may not be possible. The absorption of solar ultraviolet radiation in the stratosphere is a major contributor to the temperature inversion that is characteristic of this layer of the atmosphere. A side effect of the strong ultraviolet absorption by the ozone layer, is the creation of a "solar blind" spectral region from 200-300 nm at sea level. The solar absorption is so strong that the world is pitch black in this spectral region. Only fires and extremely hot objects emit any detectable amount of radiation in this region.

Another class of atmospheric species that is non-uniform in its altitude dependence is the class of atmospheric aerosols. The altitude dependence of the major background aerosols is shown for the fall and winter seasons in Figure 3-8 and the spring and summer seasons in Figure 3-9. The boundary layer aerosols are confined almost entirely within the first 2000 m of altitude. It takes a truly massive storm for dust, haze, and other aerosols to be lifted to higher altitudes. Sensors that are severely affected by limited visibility due to high aerosol concentrations at the surface may work very well in almost every airborne application. The explicit dependence on visibility indicates that the extinction coefficient is a linear function of the aerosol concentration. Above 2000 m there is a background aerosol whose concentrations fall off more or less smoothly. In the stratosphere, this aerosol is primarily residual volcanic dust. In the mesosphere and above, the volcanic component is gradually replaced by meteoritic dust. Right in the vicinity of the tropopause, there is frequently a relatively thick layer of volcanic dust caused by recent (within months to a few years at most) eruptions. If an eruption is violent enough to cause much of its ash to reach the tropopause, the jet stream will blow it into a band that circles the world and gradually spreads out. The strength of this layer can vary by two or even three orders of magnitude above the background levels. Over time, the volcano dust will gradually settle out.

With the exception of rare nacreous clouds and noctilucent clouds, clouds are a strictly tropospheric phenomenon. Most clouds are restricted to thin layers of the atmosphere (typically a few hundred meters to a few kilometers thick). Fogs are clouds that form at the surface and extend upwards a few meters to a few hundred meters. Clouds that do not contact the ground tend to form in three regions: low (0.5 - 2 km cloud base), medium (2 -7 km cloud base), and high (5-13 km cloud base). All three types of clouds may be present at the same time. Only cumulus, stratocumulus, and cumulonimbus clouds tend to be more than a few kilometers thick. The tops of cumulonimbus clouds may extend slightly into the stratosphere. Nacreous clouds occur high in the stratosphere. Having an iridescent "mother of pearl" appearance, nacreous clouds appear to be associated with air masses forced violently upward by high mountain ranges. Noctilucent clouds are tenuous,

Figure 3-4. Atmospheric water vapor concentration as a function of altitude.[11]

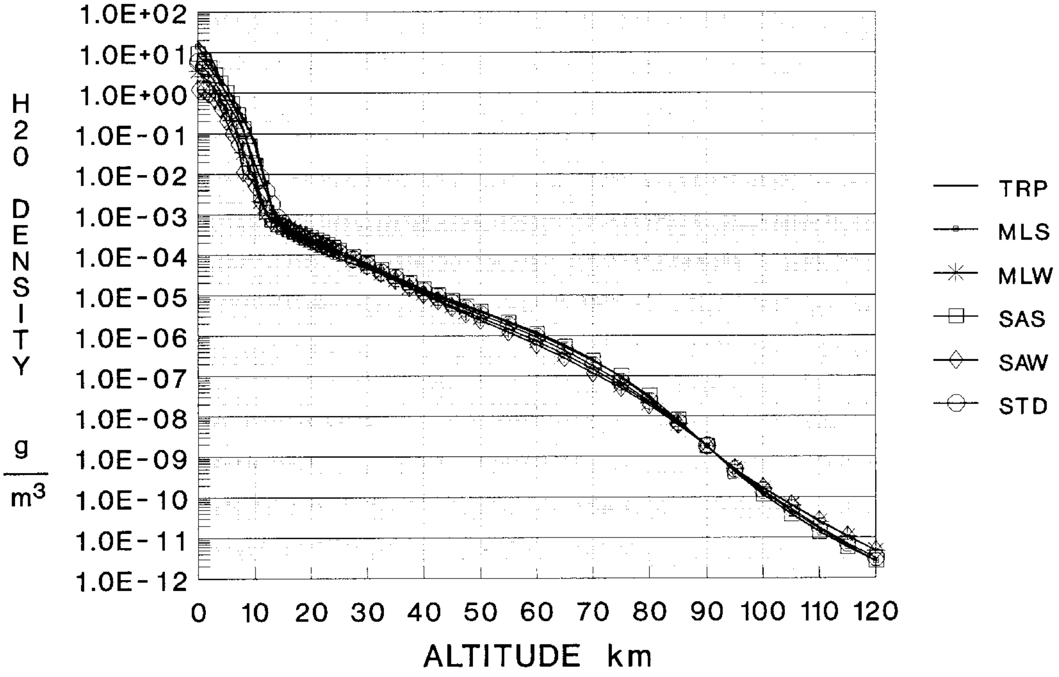


Figure 3-5. Altitude dependence of tropospheric water vapor concentration.[11]

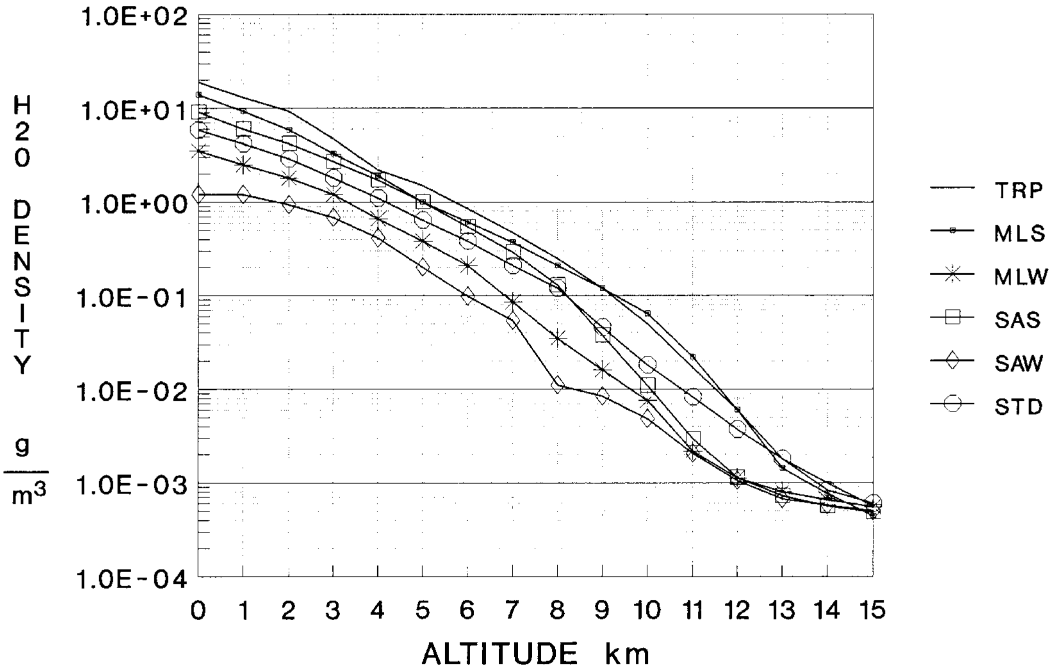


Figure 3-6. Ozone concentration as a function of altitude.[11]

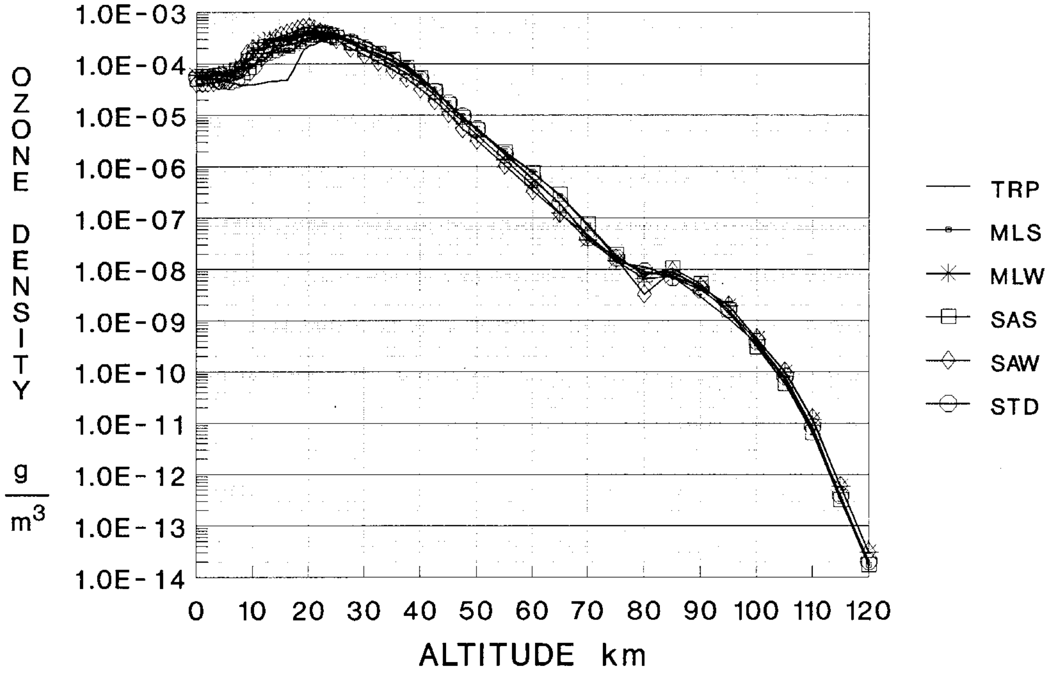


Figure 3-7. Altitude dependence of tropospheric and stratospheric ozone concentration.[11]

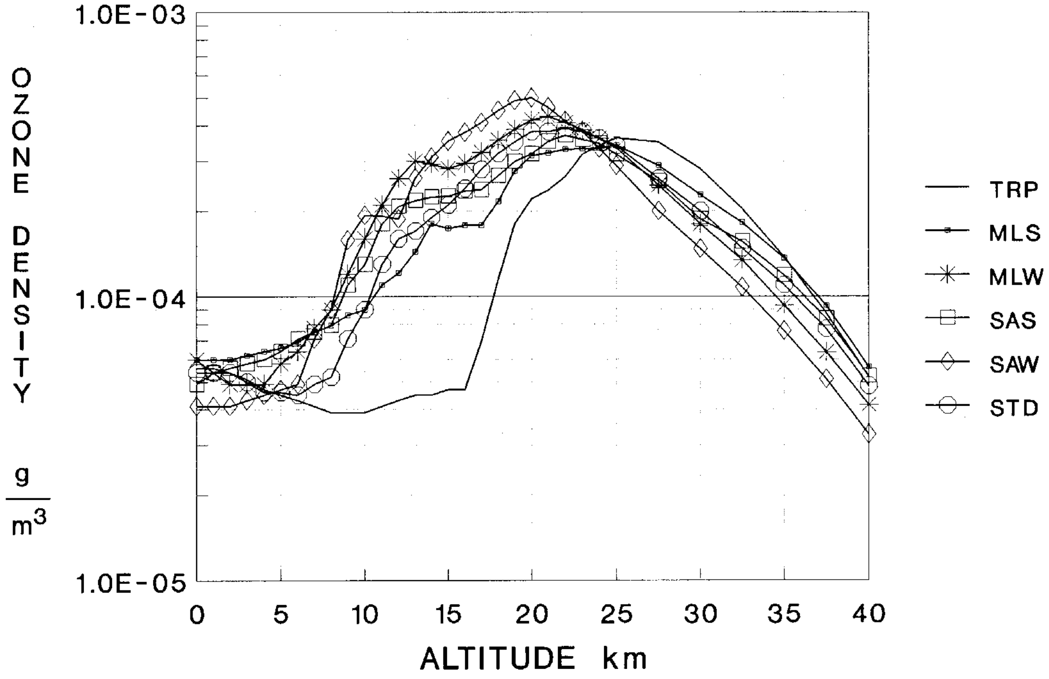


Figure 3-8. Aerosol concentrations as a function of altitude during the fall and winter.[11]

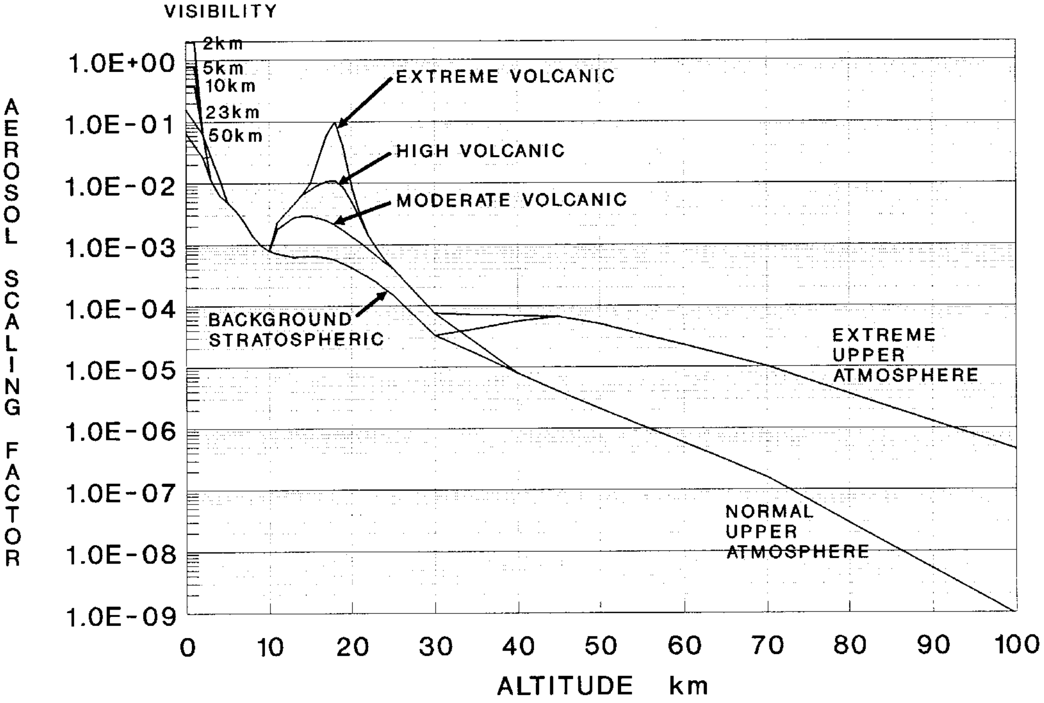
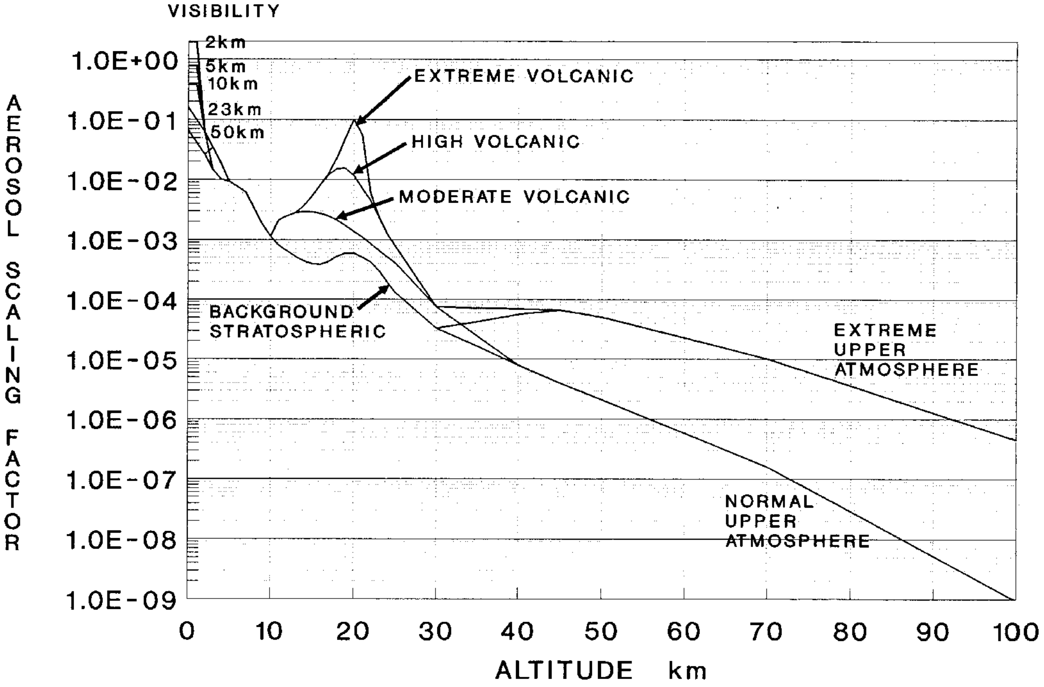


Figure 3-9. Aerosol concentrations as a function of altitude during the spring and summer.[11]

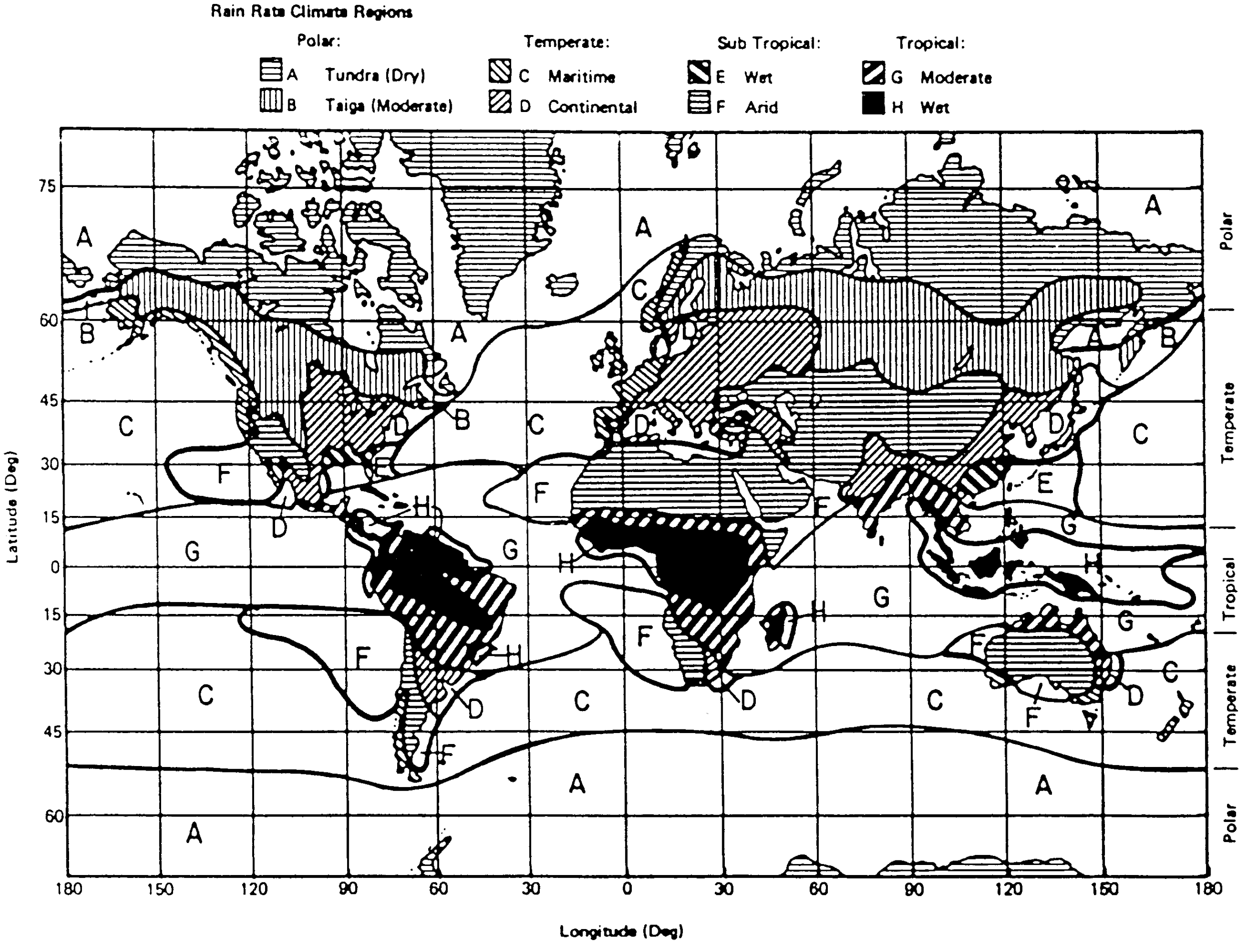


cirrostratus-like clouds occurring at the mesopause (80-90 km). Their origin is uncertain. They can be seen only when directly illuminated by sunlight against a twilight sky in the absence of intervening lower clouds.

Rain tends to be associated with low or medium clouds. Rain also tend to come in two forms: steady rains of light to moderate intensity and intense rains of short duration. The former tend to occur from most low and medium cloud types. The rain density is zero at the cloud top (2-10 km), increases toward the cloud bottom, and then remains steady or slightly decreasing (due to potential droplet evaporation) until the ground. The latter rain type tends to be associated with large cumulus or cumulonimbus (thunderstorm) clouds. Here the rain (or frozen rain) extends from near the tropopause all the way down to the ground). Snow has much the same altitude characteristics as rain. Hail is only associated with cumulonimbus clouds and will again be present from the ground to near the tropopause.

The geographic (latitude and longitude) structure of the atmosphere is strongly tied to whether an area is over the ocean or the land, whether it is equatorial, polar or in between, whether it is coastal or inland, and where it is located with respect to mountain ranges and prevailing winds and ocean currents. There have been a number of classifications of global climate into regions. One substantial classification is with respect to total rainfall, rainfall rate, and latitude. As shown in Figure 3-10, this classification scheme creates eight basic climate types based on latitude (and thus range of temperatures) and total rainfall. Subsequent studies of several of the indicated regions has broken the eight basic climate types into a number of sub-types. Some of this data is included in the section on hydrometeors later in this chapter.

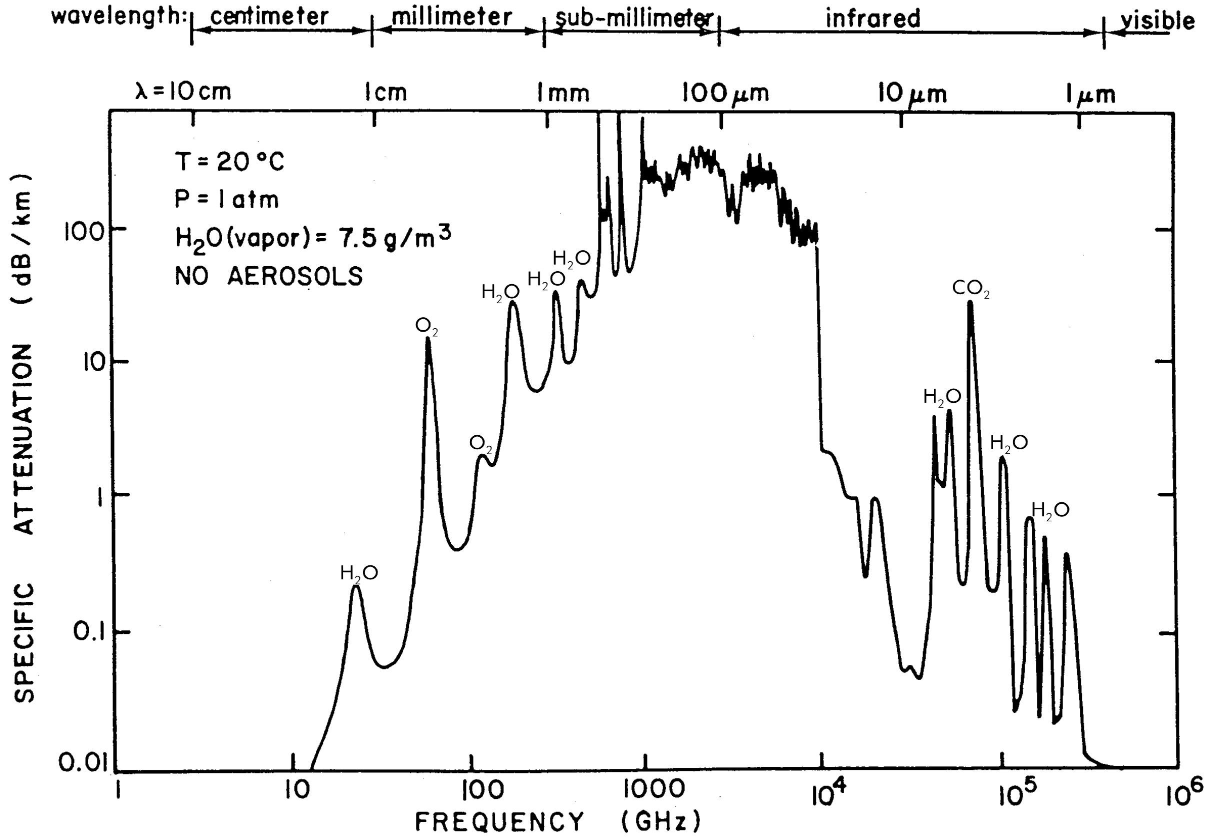
Figure 3-10. Global climate regions.[13]



Molecular Absorption

Figure 3-10 shows a calculation of the atmospheric extinction for a hypothetical atmosphere ($T = 293.15$ K, $p = 1013$ mbar, and $h = 7.5$ g/m^3 H_2O concentration) with no particulates of any kind. [14] Molecular scattering has been ignored, although its inclusion would only produce a tiny change (< 0.1 dB/km) in the region to the right of the 1 μm tick mark. Thus Figure 3-10 shows only the molecular absorption of the atmosphere. One thing is immediately apparent. There are two broad regions where the absorption is non-existent. One is in the radio frequency and microwave region and the second is in the visible. These regions of low absorption are called "atmospheric windows". The second observation is a region from roughly 20 μm to 500 μm where the attenuation routinely exceeds 100 dB/km. The attenuation is so high that virtually no atmospheric sensors are ever built to operate in this regime. A third observation is that there are a number of strong absorption peaks. In the microwave region, the absorptions have been identified as being due to oxygen and water vapor. In the infrared region, the absorptions are due mostly to carbon dioxide and water vapor. The deepest valleys between some of the absorption peaks are also called atmospheric windows. Prominent examples are near 35 GHz, 95 GHz, 8-14 μm, 4.5-5.5 μm, and 3 -4 μm. Atmospheric windows play an important role in sensor design; sensors are designed to operate in atmospheric

Figure 3-10. Molecular absorption of the atmosphere.[14]



windows. It is unusual that any sensor designed to operate outside of an atmosphere window would function properly anywhere in the atmosphere. In fact, even nature takes advantage of atmospheric windows. Most animals have eyes that see in the visible specifically because there is an atmospheric window in the visible.

To have a thorough understanding of the effects of atmospheric propagation on sensor performance we need to study the absorption lines in the microwave region and the infrared region in more detail. Let us begin in the microwave. One of the best sources of attenuation information is a model developed by Crane [15]. Microwave attenuation data was collected by analysis of satellite communication signal-to-noise ratios. Crane analyzed the data to develop a semi-empirical model for predicting microwave attenuation. Crane's coefficients were determined by regression of the measured propagation data. He assumed that the major contributors to variation in attenuation would be frequency, humidity, and temperature (the widths of the absorption peaks will be temperature dependent because they are pressure-broadened and the collision frequency depends on the average molecular velocity). The attenuation was assumed to scale linearly with the atmospheric pressure. Since sea level pressure does not vary more than a few percent, except in hurricanes, it was not included in the regression. It is easy to scale any of the following results to pressures other than 1013 mbar if desired.

In the following we assume that T = ambient temperature (in K), h = absolute humidity (in g/m^3), ν= microwave frequency (GHz), and the ambient pressure is 1013 mbar. The attenuation per km at sea level (in dB/km) is assumed to have the form

$$\gamma(\nu) = a(\nu) + b(\nu)\cdot h - c(\nu)\cdot(T - 273.15) \qquad (3.6)$$

where $a(\nu)$, $b(\nu)$, and $c(\nu)$ are frequency dependent regression coefficients. The attenuation over earth-to-space path at the zenith (in dB) is assumed to have the form

$$A(\nu, 90°) = \alpha(\nu) + \beta(\nu)\cdot h - \varepsilon(\nu)\cdot(T - 273.15) \qquad (3.7)$$

where $\alpha(\nu)$, $\beta(\nu)$, and $\varepsilon(\nu)$ are the regression coefficient for this path type. An atmospheric scale height (in km) may be calculated from

$$H(\nu) = A(\nu, 90°)/\gamma(\nu). \qquad (3.8)$$

Using this scale height, the attenuation over a slant path to space (in dB) may be found from one of the following two relations

$$A(\nu,\theta) = \frac{A(\nu,90°)}{\sin\theta} = A(\nu,90°)\csc\theta \qquad \theta > 10^{\circ} \qquad (3.9a)$$

or

$$A(\nu,\theta) = \frac{2A(\nu,90°)}{\sin\theta + \left[\sin^2\theta + \left(2H(\nu)/8500\right)\right]^{1/2}} \qquad \theta \leq 10^{\mathrm{o}} \qquad (3.9b)$$

It should be noted that because this is a regression based on real data, residual attenuations (i.e., those for which *h=0* or *T=273.15 K*) will have considerably larger errors than attenuations at normal conditions. This will be obvious in the charts that follow.

Table 3-4 shows Crane's regression coefficients at a number of frequencies. Should the reader have the need to obtain accurate estimates at frequencies other than those listed, the following interpolation procedure should be used. *Y* will denote a coefficient value, and ν will denote the frequency. First, a slope and intercept are calculated using values from the table.

$$m = \log[Y1 / Y2] / \log[\nu 1 / \nu 2] \qquad \text{Slope} \qquad (3.10)$$

and

$$b = \log(Y2) - m \cdot \log(\nu 2) \qquad \text{Intercept} \qquad (3.11)$$

Then, the interpolated value is obtained by direct substitution into the relation

$$\log[Y(\nu)] = m \cdot \log(\nu) + b\,. \qquad (3.12)$$

Cranes's interpolation technique is more accurate than simple linear interpolation because the attenuation coefficients are logarithmic quantities. Thus a logarithmic interpolation is more accurate.

Figure 3-11 shows the calculated attenuation coefficients versus frequency for four different humidity conditions. Note that the h = 0 curve has small bumps at frequencies we have already associated with as being due entirely to water vapor. At zero humidity there should be no absorption peaks at these frequencies. The small bumps that show up are artifacts of the regression. Figure 3-12 shows the analogous plots for earth-to-space zenith paths. Eqs. (3.6) through (3.12) are routinely incorporated into computer codes that are used to predict the performance of microwave sensors and microwave communication links. We will refer back to these results when we begin to discuss sensors operating in the microwave region.

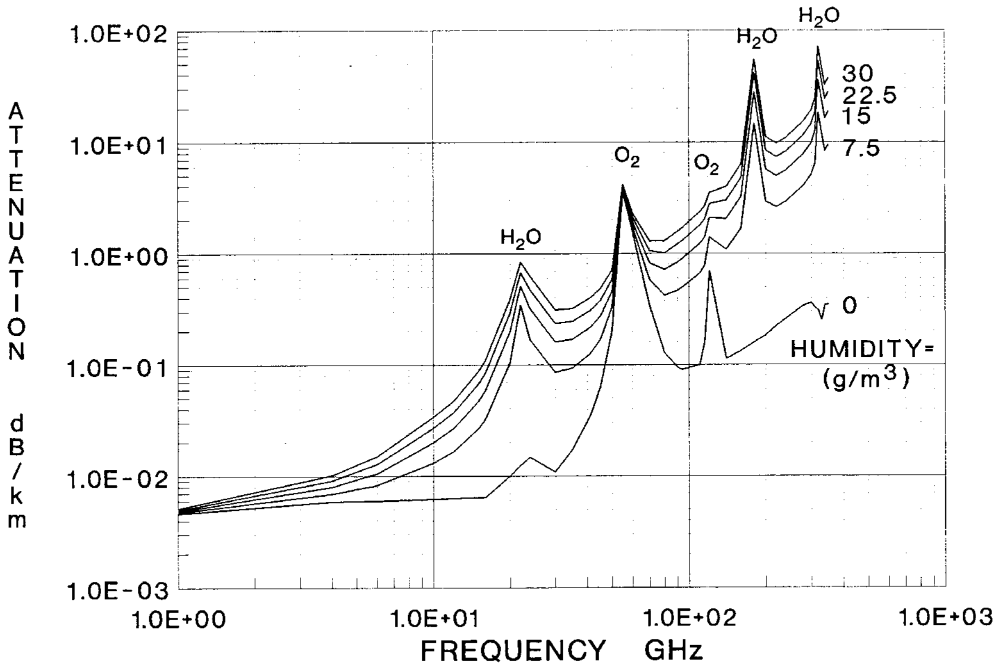
In the preceding we have used the concept of absolute humidity. Absolute humidity is the amount of water vapor (in g/m^3) in the air. At any temperature, there is a maximum absolute humidity which corresponds to complete saturation of the air by water vapor. This absolute humidity corresponding to air saturated with water vapor can be computed from the expression [16]

Table 3-4. Crane's regression coefficients vs. frequency for microwave attenuation over horizontal paths (a, b, c) and earth-to-space zenith paths (α, β, ε). [15]

ν (GHz)	a(ν)	b(ν)	c(ν)	α(ν)	β(ν)	ε(ν)
1	0.00588	0.0000178	0.0000517	0.0334	0.0000028	0.000112
4	0.00802	0.000141	0.000085	0.0397	0.000276	0.000176
6	0.00824	0.0003	0.0000895	0.0404	0.000651	0.000196
10*	0.00878	0.000919	0.000103	0.0427	0.00210	0.000278
12	0.00898	0.00137	0.000108	0.0436	0.00318	0.000315
15	0.00953	0.00269	0.000125	0.0461	0.00634	0.000455
16	0.00976	0.00345	0.000133	0.0472	0.00821	0.000536
20	0.0125	0.0125	0.000101	0.056	0.0346	0.00155
22	0.181	0.0221	0.000129	0.076	0.0783	0.0031
24	0.0162	0.0203	0.0000563	0.0691	0.0591	0.0025
30	0.0179	0.01	0.00028	0.085	0.0237	0.00133
35	0.0264	0.0101	0.000369	0.123	0.0237	0.00149
41	0.0499	0.0121	0.00062	0.237	0.0284	0.00211
45	0.0892	0.014	0.00102	0.426	0.0328	0.00299
50	0.267	0.0171	0.00251	1.27	0.0392	0.00572
55	3.93	0.022	0.0158	24.5	0.049	-0.00121
60*	1.797	0.0252	0.00999	10.17	0.0566	0.0027
70	0.449	0.0319	0.00443	2.14	0.0732	0.0104
80	0.16	0.0391	0.0013	0.705	0.0959	0.00586
90	0.113	0.0495	0.000744	0.458	0.122	0.00574
94	0.106	0.054	0.000641	0.417	0.133	0.00594
100*	0.110	0.0614	0.000642	0.422	0.151	0.00663
110	0.116	0.0749	0.000644	0.431	0.185	0.00785
115	0.206	0.0826	0.00185	0.893	0.203	0.0113
120	0.985	0.0931	0.0115	5.35	0.221	0.0363
140	0.123	0.129	0.000372	0.368	0.319	0.0119
160	0.153	0.206	0.000784	0.414	0.506	0.0191
180	1.13	1.79	-0.00237	0.281	5.04	0.192
200	0.226	0.366	0.00167	0.562	0.897	0.0339
220	0.227	0.316	0.000174	0.543	0.777	0.0276
240	0.258	0.356	-0.000119	0.601	0.879	0.0307
260*	0.296	0.423	-0.00009	0.679	1.042	0.0365
280	0.336	0.497	-0.000066	0.76	1.22	0.0428
300	0.379	0.629	0.000808	0.853	1.54	0.0551
310	0.397	0.812	0.00286	0.905	1.97	0.0735
320	0.732	2.36	0.00467	1.66	6.13	0.238
330	0.488	1.61	0.00945	1.13	3.94	0.155
340	0.475	1.06	0.00519	1.07	2.56	0.0969
350	0.528	1.23	0.00722	1.2	2.96	0.114

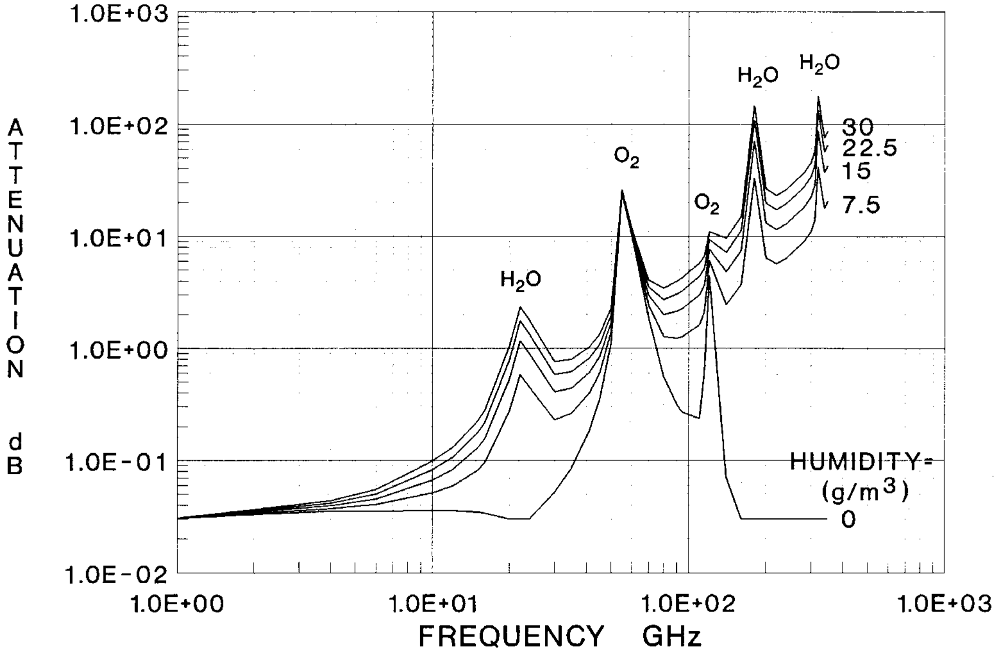
* Interpolated Values

Figure 3-11. Microwave attenuation coefficient versus frequency for horizontal sea level paths.



NOTE: ZERO-HUMIDITY CURVE IS APPROXIMATE

Figure 3-12. Microwave attenuation coefficient versus frequency for surface-to-space zenith path.



NOTE: ZERO-HUMIDITY CURVE IS APPROXIMATE

$$\begin{aligned}\rho_{SAT} &= 17.39\left(\frac{300}{T}\right)^{6} 10^{\left(10-9.834\frac{300}{T}\right)} \\ &= 1.2677\times 10^{26}\,\frac{10^{-2950.2/T}}{T^{6}}\end{aligned} \qquad (3.13)$$

and is shown in Figure 3-13. If the air has a known amount of absolute humidity, this same curve may be used to find the dewpoint. The dewpoint is the temperature at which cooling air becomes saturated with a fixed absolute amount of water vapor. If the temperature drops below the dewpoint, some of the water must condense out as fog or dew. The highest recorded dewpoint is roughly 34 °C. This corresponds to 37 g/m^3 of absolute humidity. If a problem specification requires outdoors operation in absolute humidities greater than 37 g/m^3, then the problem specification is suspect.

Relative humidity is a term commonly used in news hour weather reports. It is defined as

$$\textit{Relative Humidity} = \textit{Absolute Humidity/Saturation Humidity} \times \textit{100\%} \qquad (3.14)$$

For example, if the temperature is 30 °C, then the saturation humidity is 30 g/m^3. If at the same time the absolute humidity is 20 g/m^3, then the relative humidity is 20/30 x 100% = 67%. If the dewpoint is 10 °C, then the absolute humidity is 9.5 g/m^3. If the actual temperature is 20 °C, then the saturation humidity is approximately 17 g/m^3. The relative humidity is thus 9.5/17 x 100% = 56%.

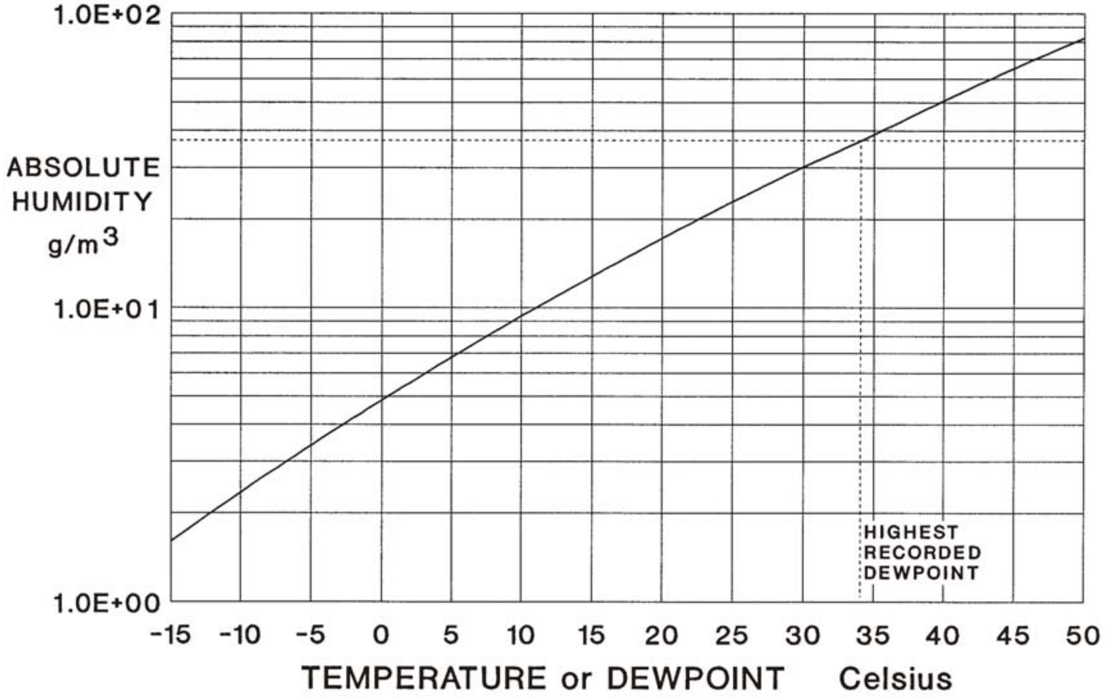
The attenuation coefficient in the infrared is seldom evaluated analytically or even semi-empirically. Its calculation is usually performed using one of a few computer programs that have been developed over the course of the last several decades. LOWTRAN is the most widely known. [3]-[11] It was developed by the then Air Force Geophysics Laboratory for low resolution (>20 cm^{-1}) atmospheric transmission and atmospheric radiance analyses. It uses band model calculations for frequencies (wavelengths) from

$$\begin{aligned} 0 \;\le\; &\Delta\bar{\nu} \;\le\; 50{,}000\ \mathrm{cm}^{-1} \\ 0.2\ \mu\mathrm{m} \;\le\; &\lambda \;\le\; \infty \\ 0 \;\le\; &\nu \;\le\; 1.499\times 10^{15}\ \mathrm{Hz} \end{aligned}$$

LOWTRAN will provide outputs for frequencies in the microwave region. However, the poor resolution (20 cm^{-1}) makes all data below 600 GHz or above 500 μm to have questionable validity.

LOWTRAN includes:
- 6 user-selectable atmospheric climate models with 33-layer atmospheres (see Table 3-3)
- Numerous user-selectable aerosol models
- Rain, fog, and cloud models
- Multiple scattering solar irradiance contributions
- AFGL (0-120 km) atmospheric profiles

Figure 3-13. Absolute humidity of saturated air as a function of temperature.[16]



- User-specified path geometries (incl. horizontal, slant, surface-to-space, etc.)
- Atmospheric refraction
- User-definition of atmospheric or aerosol models is permitted.

LOWTRAN is written in FORTRAN. The latest version is LOWTRAN7 which was released in Spring 1989. A PC-based executable version is available from Ontar Corporation in Brookline, MA as is the original FORTRAN source code.[11] The Air Force selected Ontar to be the long-term support contractor for this code. The source code contains a wealth of hidden information (incorporated as block data files) on the atmosphere and the various models used in the calculations. This data extracted from the LOWTRAN7 source code was used to create many of the charts in this Chapter.

LOWTRAN has now been replaced by MODTRAN.[17] MODTRAN is a moderate resolution version of LOWTRAN. It uses almost all of the same LOWTRAN models but permits calculations to 1 cm^{-1} resolution. MODTRAN outputs for frequencies below 30 GHz and wavelengths longer than 1 cm are still suspect. Several versions of MODTRAN are also available from Ontar.[18]

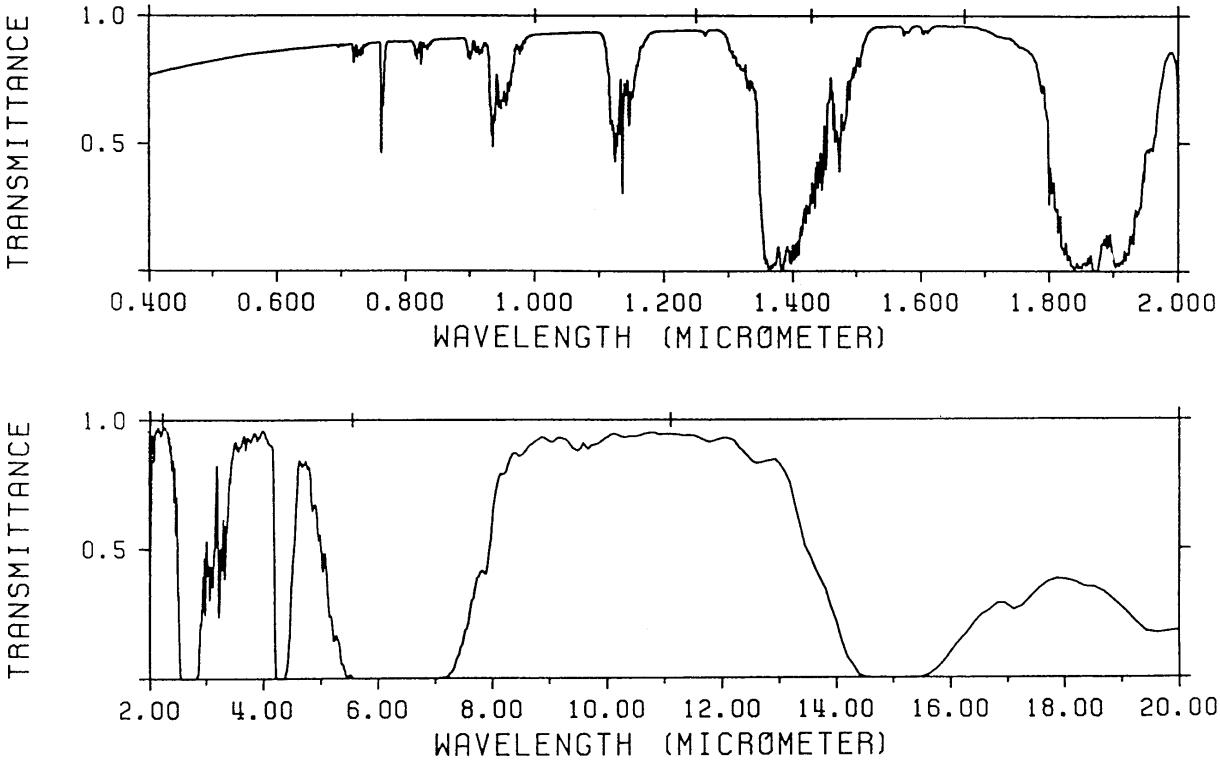
HITRAN [19] was the high resolution companion to LOWTRAN. It used a completely different scheme (rigorous line-by-line addition of spectral contributions as opposed to a simpler absorption band model) for computing molecular absorption. This permits calculations to any resolution desired by the user. Otherwise it used most of the same models as LOWTRAN (such as the path models and models for aerosols, fogs, clouds, rain, and refraction). HITRAN is extremely useful for calculating transmission of laser outputs. HITRAN was replaced a number of years ago by FASCODE [20] which uses the same scheme as HITRAN but implements a much faster computational algorithm. FASCODE and HITRAN both are available from Ontar.[21],[22]

There are other atmospheric transmission codes, such as the Army's EOSAEL family.[23] However, the author has never used them and cannot report their characteristics with any certainty.

Let us now examine the visible and infrared portions of the spectrum. Figure 3-14 shows the attenuation in the spectral regions from 0.4 to 2.0 μm and from 2.0 to 20.0 μm. The calculations were performed using LOWTRAN7 for the specific inputs shown in the figure caption. Note that the wavelength scale is different for the two plots. Transmittance is the fraction of the incident radiation that is not absorbed or scattered. It is directly related to the extinction coefficient α through the relation

$$T = e^{-\alpha R} \tag{3.15}$$

Figure 3-14. LOWTRAN7 calculation of the transmittance of the atmosphere in the infrared. The calculation assumes the mid-latitude winter climate model, with a rural aerosol model with 23 km visibility, and a 1 km horizontal path at sea level.[11]



where R (range) is the distance traveled. There is obviously considerable structure. In fact, in several wavelength regions there are two, three, or even four different species making significant contributions to the extinction.

A better understanding of the qualitative character can be obtained by using the same LOWTRAN 7 output to identify the contributions from each species. This is shown in figures 3-15 and 3-16. Figure 3-15 covers the visible and near infrared spectrum. There are a few isolated peaks associated with the uniformly-mixed gases. The one sharp line at roughly 0.78 µm is due to oxygen; the remaining peaks are due to carbon dioxide. Water vapor is seen to be the major contributor to the extinction structure with a number of regularly spaced absorption bands. Each band contains a myriad of narrower absorption lines. The water vapor continuum absorption is seen to be a significant contributor although not as important as the absorption from isolated water molecules. Molecular (Rayleigh) scattering is seen to be weak contributor, showing up only as having a significant effect below 0.5 µm. Aerosol scattering is much stronger, contributing a smooth background extinction throughout the visible and near infrared.

The mid infrared (Figure 3-16) presents an even more complicated picture. There are strong absorptions from the uniformly-mixed gases. In this instance the absorber is almost entirely CO_2. Water vapor and the water vapor continuum provided strong absorption at a number of bands in this region. There is a limited aerosol extinction virtually everywhere. Please note that the visibility in Figure 3-16 was chosen as 5 km; in Figure 3-15 it was 23 km. The choice of lower visibility makes the aerosol contribution more pronounced without affecting anything else. Ozone and nitrogen continuum absorption contribute small amounts of absorption in two places. A brief summary of the character of infrared molecular absorption could read: no significant molecular absorption in the visible; broad but well-spaced water vapor absorption bands dominate the near infrared; carbon dioxide and water vapor vie for strongest absorber throughout most of the mid infrared with only a few atmospheric windows being relatively free of absorption.

Figure 3-15. Contributors to atmospheric extinction in the visible and near infrared.[11]

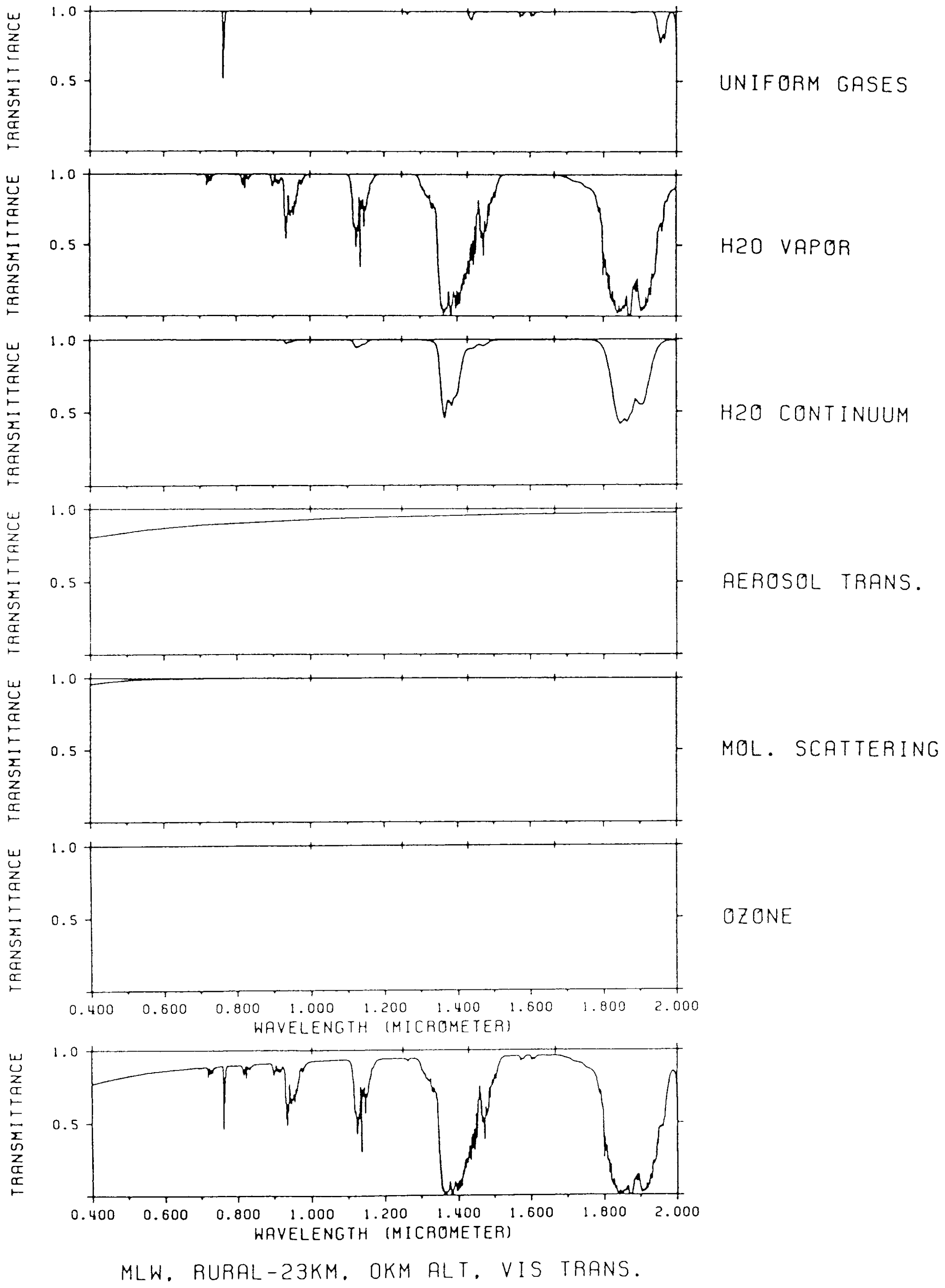
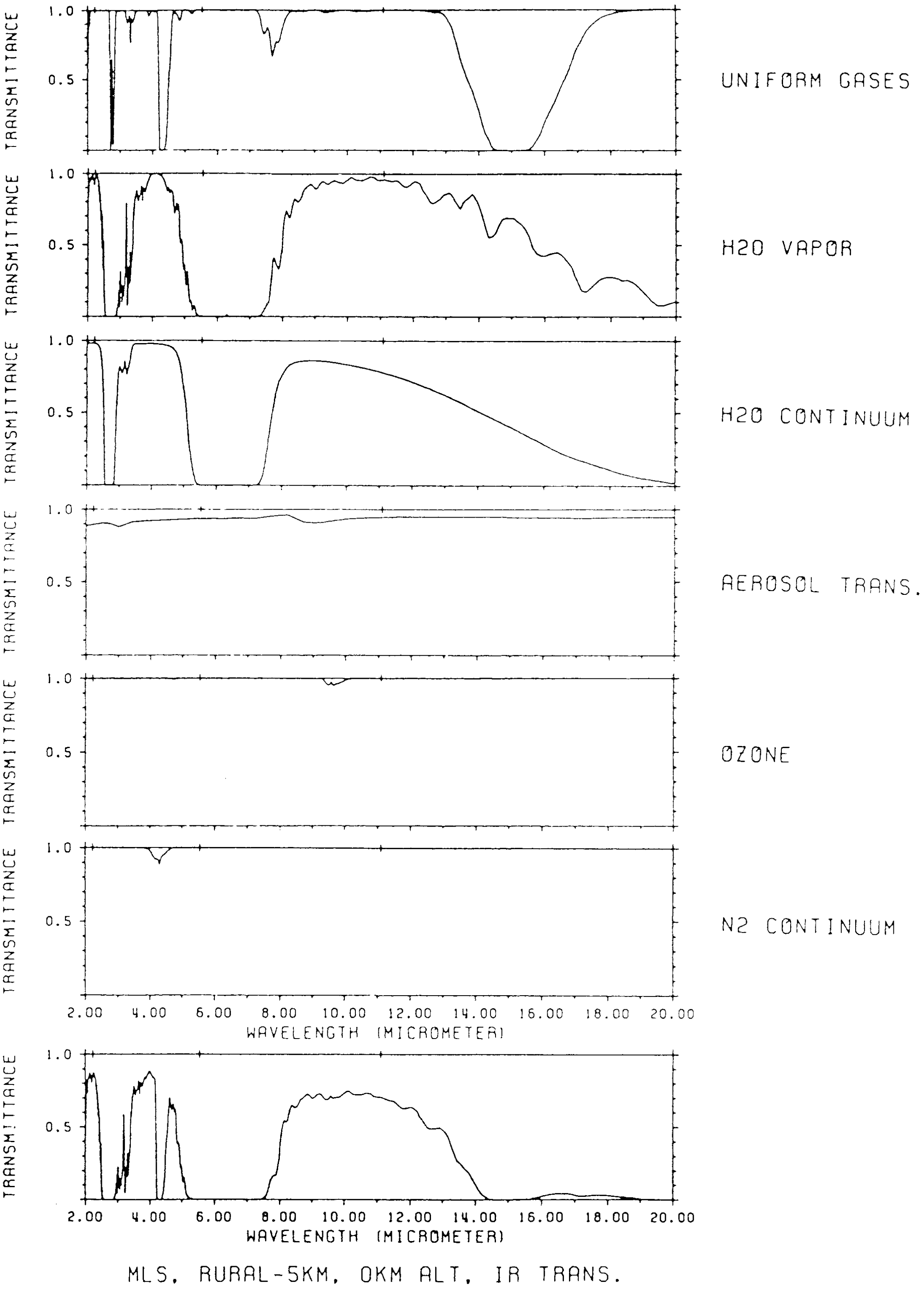


Figure 3-16. Contributors to atmospheric extinction in the mid infrared.[11]



Aerosol Scattering

Figures 3-15 and 3-16 each have one graph that shows the effects of aerosol scattering on the visible and infrared transmission through the atmosphere. The aerosols addressed by these graphs are the so-called “haze” aerosols. These aerosols are formed naturally from condensation of atmospheric water vapor around tiny particles of soot, salt, organic condensates, or dust (of soil, meteoritic, or volcanic origin). Depending on the origin of an aerosol, it can have vastly different attenuation characteristics. Figure 3-17 shows the relative extinction coefficient as a function of wavelength for the four major tropospheric aerosol models used by LOWTRAN. Before examining the differences between the aerosols it is important to understand the meaning of the term “relative” in our description. Note that all four aerosols have a relative extinction of “1.0" at 0.55 µm. This wavelength is the wavelength at which meteorological range (visibility) is defined. The extinction at any other wavelength is relative to the extinction at 0.55 µm. Given a declaration of visibility from some source, the extinction coefficient at 0.55 µm can be determined from Eq. (3.1). This value then defines the absolute extinction coefficient at 0.55 µm. Multiplying any relative extinction coefficient at wavelength λ (from the curves) by the absolute extinction coefficient at 0.55 µm gives the absolute extinction coefficient at wavelength λ. This necessity to know the visibility in order to predict the aerosol attenuation at other wavelengths is one of the peculiarities of LOWTRAN. It is also one region that visibility remains a critical meteorological parameter in estimating sensor performance.

Figure 3-17. Relative extinction coefficient of tropospheric aerosols.[11]

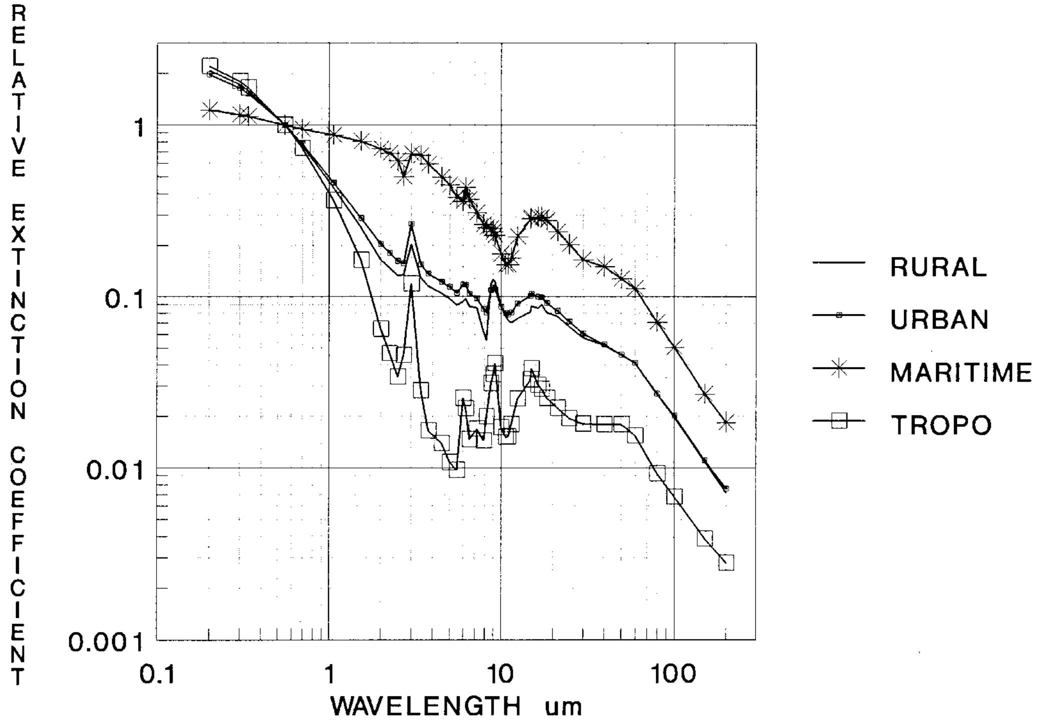


Figure 3-18. Correlation of visibility and atmospheric extinction at 10.6 μm.[24]

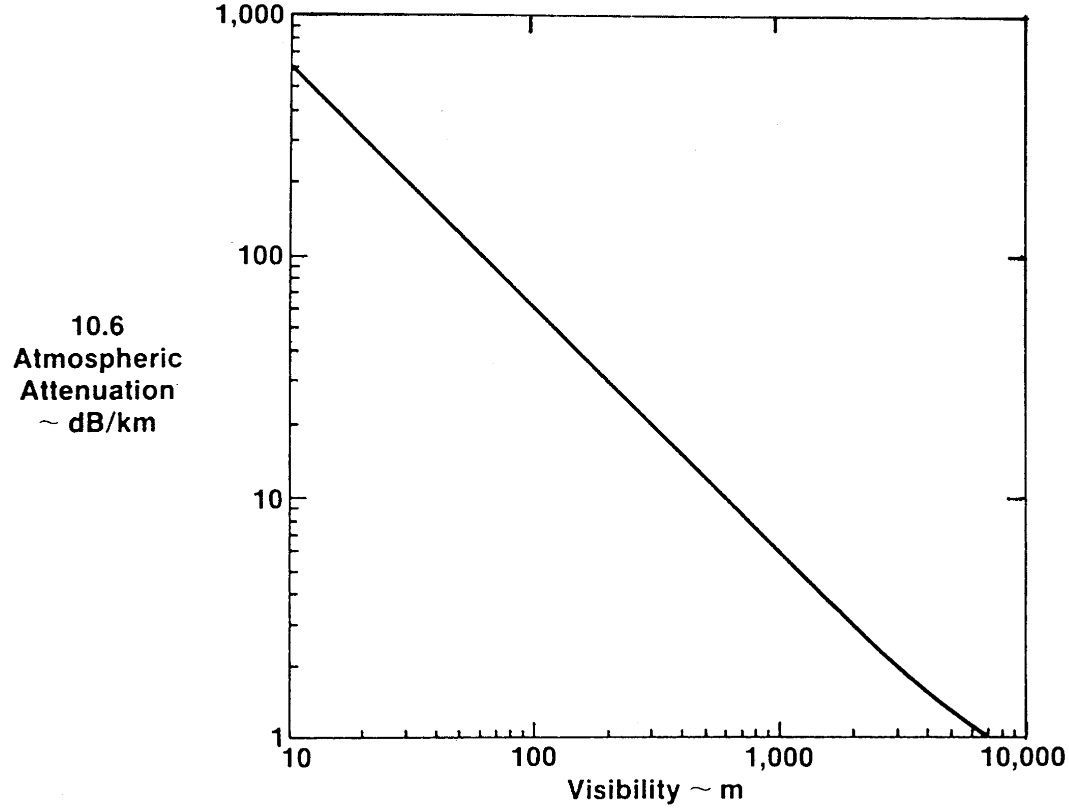


Figure 3-18 shows a correlation of visibility with measured atmospheric extinction at 10.6 μm. For visibilities below 3 km, the curve is a virtually perfect fit to

$$\textit{Extinction} = \textit{Constant} \, / \, \textit{Visibility}. \tag{3.16}$$

At visibilities above 3 km, the attenuation becomes more and more dominated by molecular extinction. This explains the slight hook at one end of the plot. Correlations like this have reinforced time and again the ability to use visibility to predict extinction at other wavelengths.

Examination of the curves in Figure 3-17 indicates that depending on the source of the aerosol, there can be more than a factor of ten difference in extinction in the 3-5 μm region of the infrared, although a factor of five is more characteristic in the 8-12 μm region. Figure 3-19 shows the relative absorption coefficient of the tropospheric aerosols. Relative in the sense of Figure 3-19 refers to the fraction of the total extinction produced by absorption. The extinction coefficient is the sum of the absorption and scattering coefficients

$$\alpha_{EXT}(\lambda) = \alpha_{ABS}(\lambda) + \alpha_{SCAT}(\lambda) \tag{3.17}$$

If $A(\lambda)$ is the relative extinction coefficient at wavelength λ, and $B(\lambda)$ is the relative absorption coefficient, then

$$\alpha_{EXT}(\lambda) = A(\lambda)\,\alpha_{EXT}(0.55\mu\text{m}) \tag{3.18}$$

and

$$\alpha_{ABS}(\lambda) = B(\lambda)\,\alpha_{EXT}(\lambda)\,. \tag{3.19}$$

Examination of Figure 3-19 shows that the relative absorption is small for all of the tropospheric aerosols. That is, the tropospheric aerosols have considerably more scattering than absorption. This is especially true in the 3-5 and 8-12 μm bands.

Since most tropospheric aerosols have a significant water content, it might be expected that relative humidity would strongly affect the extinction. Figure 3-20 shows the relative extinction coefficient of maritime aerosols (nominally salt water aerosols) as a function of relative humidity, while Figure 3-21 shows the relative absorption for those same aerosols under the same conditions. As expected, relative humidity does have a significant impact on both the extinction coefficient and absorption of the maritime aerosol. In general, higher humidity leads to higher extinctions and higher absorption in the infrared. Similar humidity effects are observed with the other aerosols.

Figure 3-19. Relative absorption coefficient of tropospheric aerosols.[11]

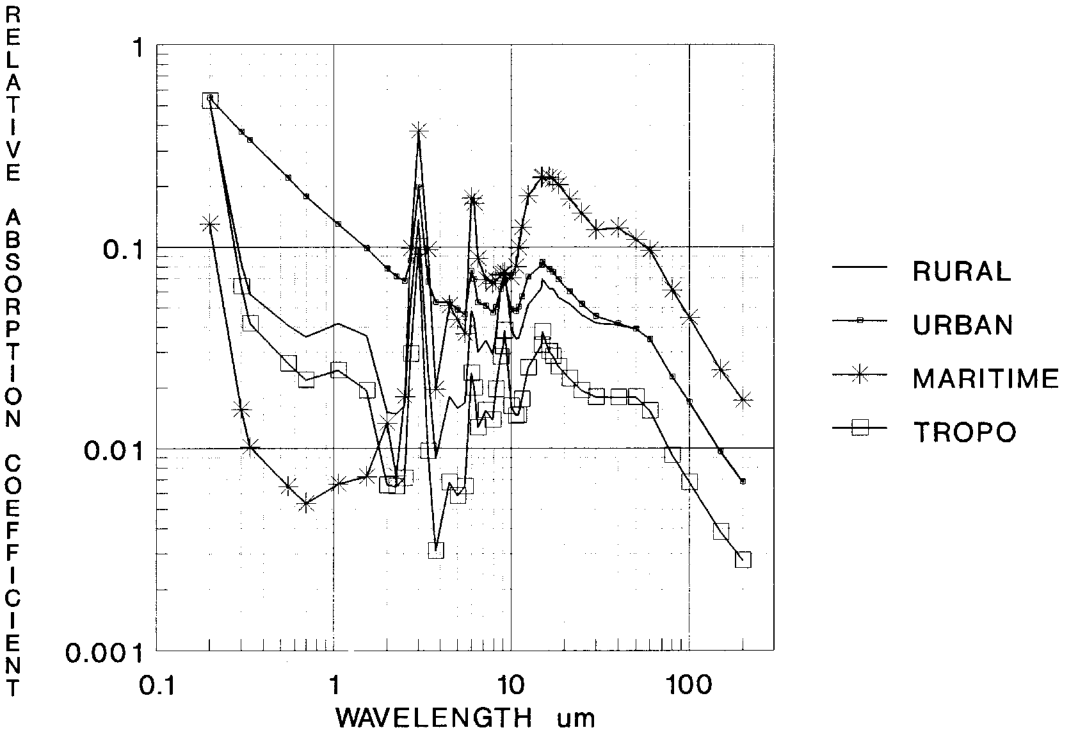


Figure 3-20. Relative extinction of maritime aerosols as a function of relative humidity.[11]

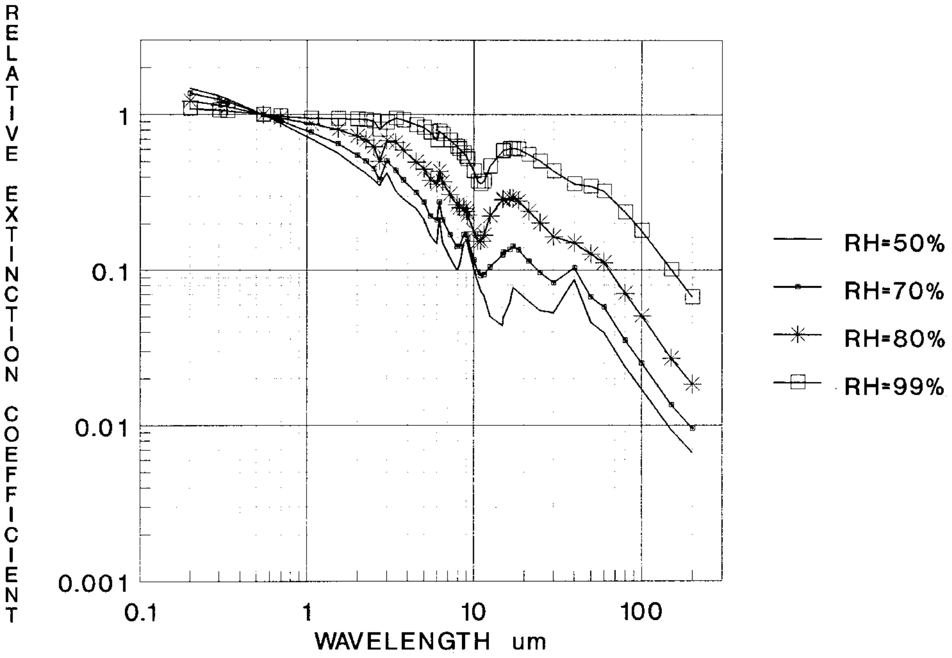
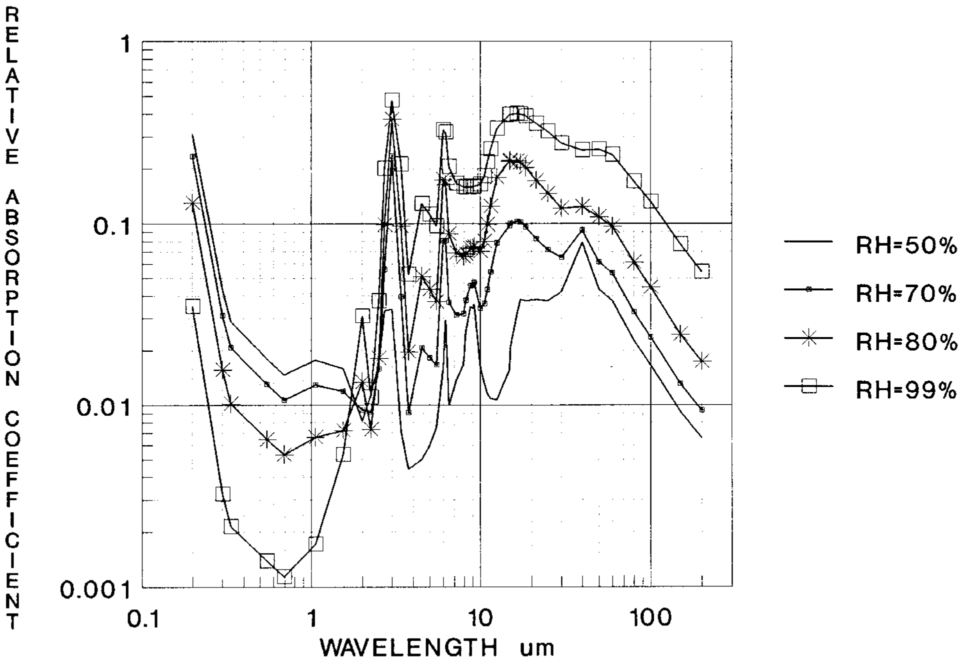


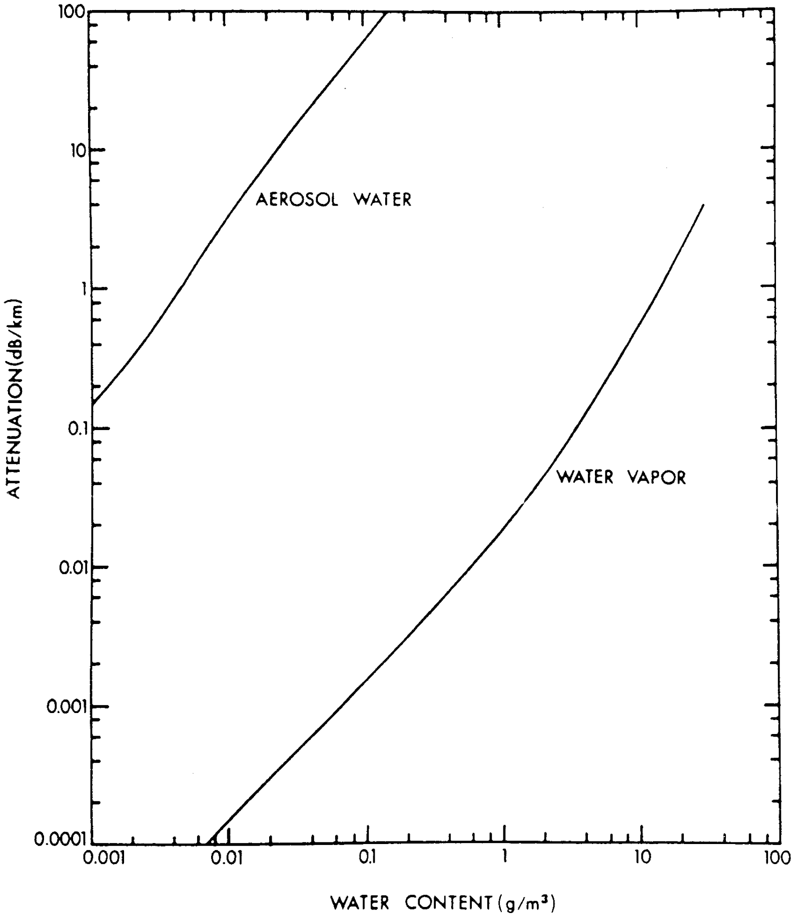
Figure 3-21. Relative absorption of maritime aerosols as a function of relative humidity.[11]



Hydrometeors - Fogs and Clouds

As defined earlier, hydrometeors are "water in the air", that is, rain, fog, clouds, snow, etc. In an earlier section we saw how important the absolute humidity was in defining the molecular absorption contribution to the total extinction. Just as important, however, is the fraction of that humidity that is in the form of liquid water. Figure 3-23 compares the attenuation of liquid water with that of water vapor as a function of total water content (humidity). As is evident, the attenuation of the liquid is four orders of magnitude greater than the vapor. This should not come as a complete surprise. Remember how important the water vapor continuum was to absorption in the infrared. That extinction, which was comparable in magnitude to the direct vapor absorption was caused by a relatively small number of water molecules in collision with each other. In liquid water, every molecule is in continual collision with its neighbors. The continuum absorption dominates all other effects. It does so to such a great deal that liquid water is a strong absorber from the yellow

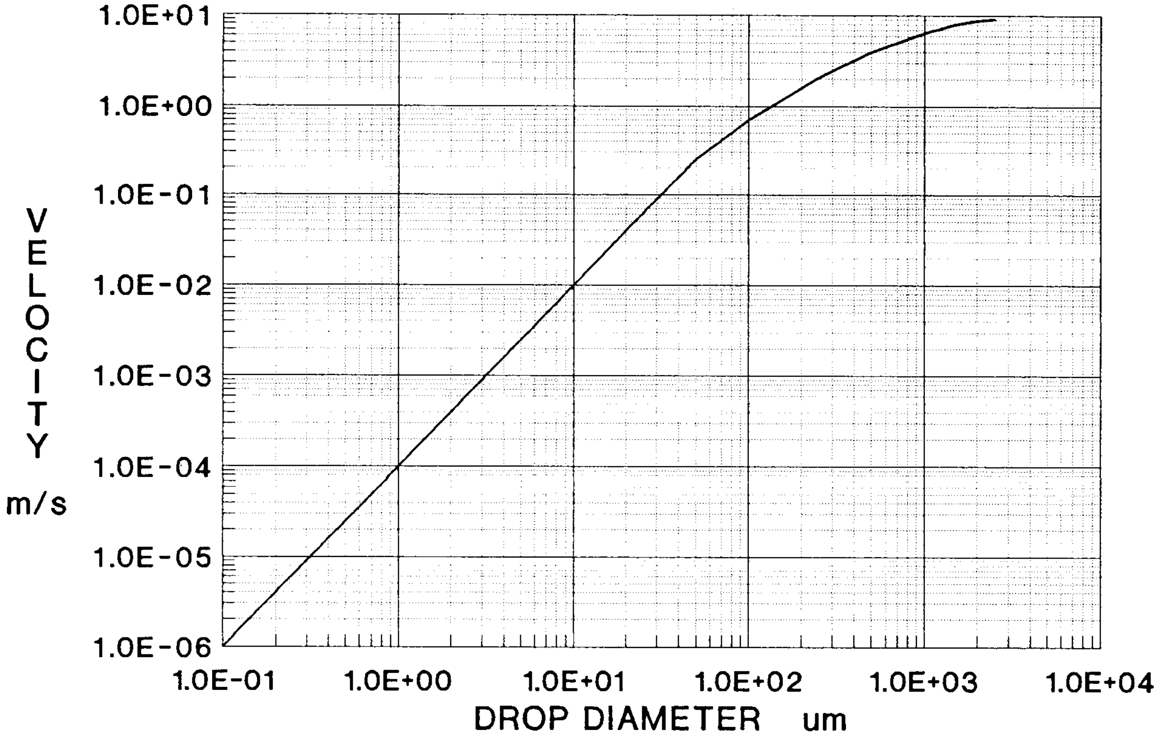
Figure 3-22. Comparison of the attenuation produced by different concentrations of liquid water and water vapor.



region of the visible to the VLF region of the radio spectrum. The upshot of this is that if there is only 0.01 g/m^3 of liquid water in the air, its extinction will be comparable to that of 30 g/m^3 of water vapor. Furthermore, clouds and fogs can have liquid water content exceeding 1 g/m^3. Thus, hydrometeors can be the dominant source of weather-related attenuation in some portions of the electromagnetic spectrum.

There are two basic classes of hydrometeors: the small particles that remain suspended in the air for minutes to hours and the large particles that fall to the surface within seconds to minutes. Figure 3-23 shows the terminal velocity of water droplets as a function of their size.[25] Note that droplets with radii greater than 100 μm have terminal velocities greater than 1 m/s. Unless driven upwards by strong winds, these large droplets will fall the few kilometers to the ground in a few tens of minutes or less. They make up rain and drizzle. Particles less than 10 μm have terminal velocities less than 1 cm/s. These small particles will take minutes to settle a few meters. Their motion will be dominated more by wind motion (even mild breezes) than to their downward settling. These particles make up the majority of fogs and clouds. The particles in the intermediate zone, those tens of μm in diameter, will settle out of low-lying hydrometeor formations in a few minutes. They form the "mist" component of fogs and are the particles which rapidly coalesce into larger raindrops that will fall rapidly out of the clouds. Snow particles do follow the same curves. Almost all falling snow flakes will have velocities of 1-2 m/s. Graupel velocities can approach 2-3 m/s.

Figure 3-23. Terminal velocity versus droplet diameter for water droplets. [25]



It should be noted that the theoretical terminal velocity of a spherical particle scales as the square root of the material density. Thus, the curve of Figure 3-23 can be used to describe the "fallout" of "spherical" particles other than water by multiplying the terminal velocity by the square root of the density. For example, dust particles have densities between 1.5 and 2.6 g/m^3. Such particles should fall between 1.2X and 1.6X faster than raindrops of the same size.

The generic extinction behavior of the large water droplets and small water droplets is shown in Figure 3-24. Both exhibit behavior characteristic of Mie scattering. The small particles have maximum extinction in the visible and infrared. The extinction then decreases rapidly becoming negligible in the microwave region (below 10 GHz). Just like the aerosol extinction described in the preceding section, the small particle extinction will scale inversely with the visibility. The large particles have more or less constant extinction from the visible to the millimeter-wave region. The extinction then falls off rapidly, becoming negligible below in the UHF and VHF regions. The total large particle extinction scales with the total large particle concentration (essentially with precipitation rate).

We address the small suspended aerosols in the current section. Large hydrometeors or precipitation are covered in the following section. In the visible and infrared regions, hydrometeor aerosols are treated like any other aerosol. Figures 3-25 and 3-26 show the relative extinction and relative absorption of two different fog aerosols. **Radiation fogs** are formed when the emission of

Figure 3-24. Summary of hydrometeor effects on electromagnetic radiation.[14]

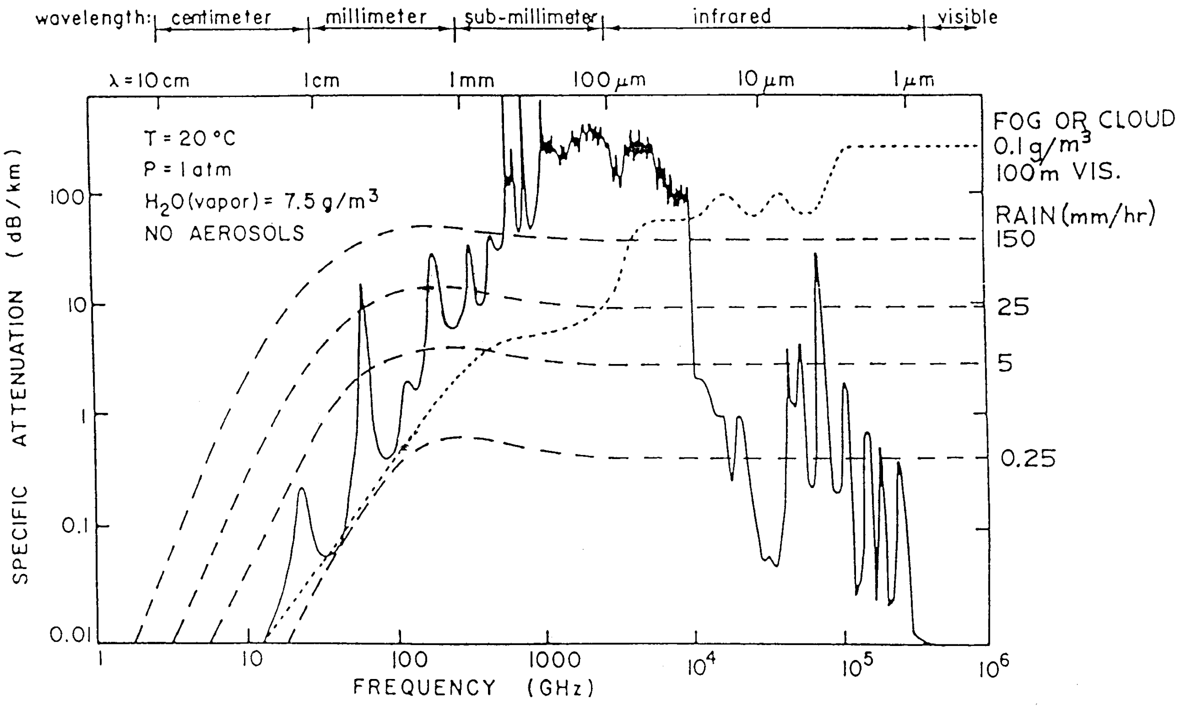


Figure 3-25. Relative extinction coefficient of two different generic fog aerosols.[11]

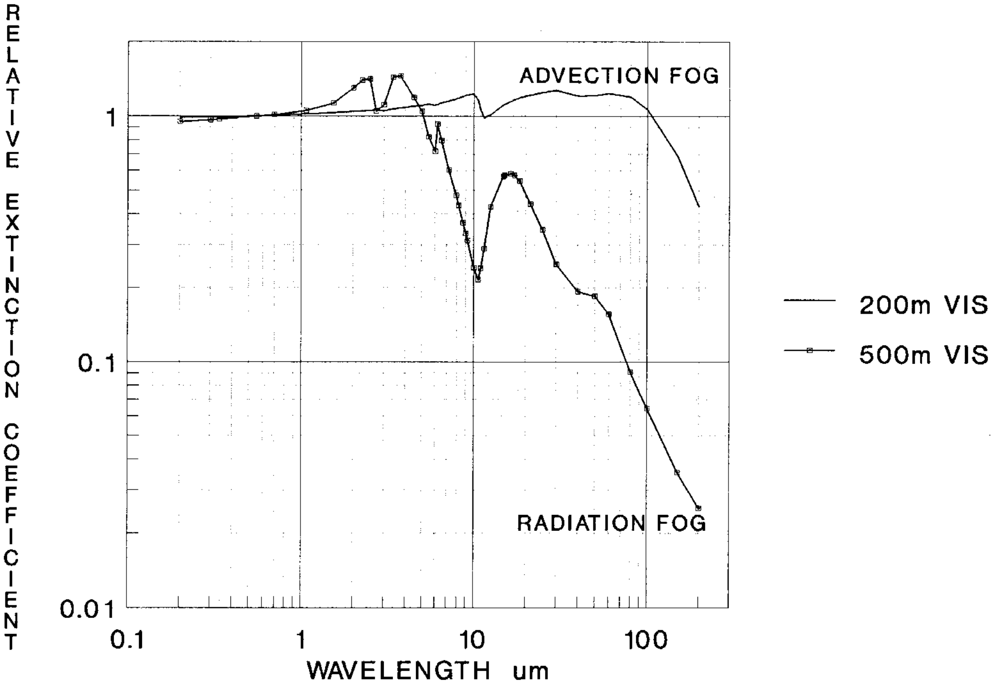
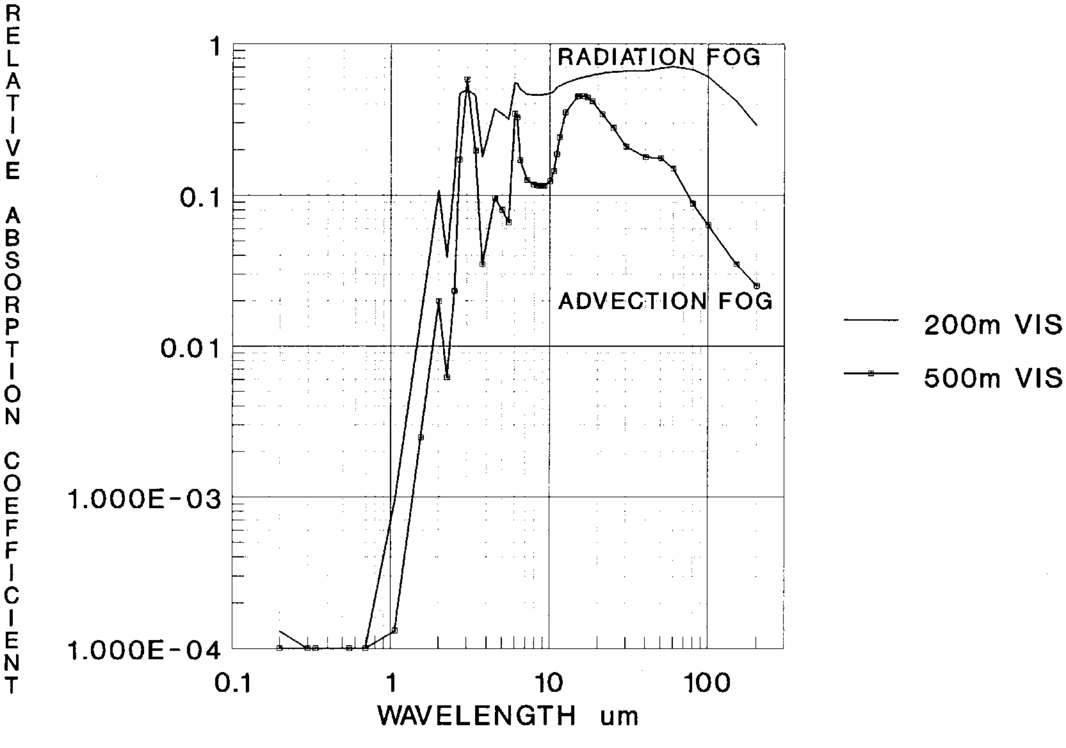


Figure 3-26. Relative absorption of two generic fog aerosols.[11]



thermal radiation from an airmass causes it to cool below its dewpoint. When the air has cooled sufficiently, some trigger (the presence of nucleation particles, turbulence, cosmic rays – this is what causes tracks to form in a Wilson cloud chamber, etc.) causes the water in the supersaturated air to condense into thousands of tiny droplets (fog). Radiation fogs evolve with time. The initial concentration of water is fixed by the amount that was present in the supersaturated air. There is little or no horizontal transport when conditions are optimal for radiation fog formation. As time goes by, collision between the tiny droplets cause them to coalesce into larger and larger droplets, until they grow large enough to settle out of the air. Because there is no replacement of the small droplets, the average particle size in an evolving radiation fog grows with time while the average density of particles decreases with time. The "Tule fogs" that regularly menace traffic in the Central Valley of California are classic examples of radiation fogs.

Advection fogs form when warmer, moister air flows over a colder surface. Advection fogs tend to evolve spatially rather than over time. Small particles are continually being formed as new masses of warm, moist air contact the cold surface. As the fog drifts, the particles coalesce as in a radiation fog. Greater degrees of evolution correspond to greater distances from the "generation" zone. The **sea fogs** that regularly form off the California coast and frequently roll in over the hills around San Francisco Bay are classic examples of advection fogs.

Other types of fog include steam fogs, warm-front fogs, and up-slope fogs. **Steam fogs** result when cold air moves over warm water. The high evaporation rate from the warm water supersaturates the cold air and wisps of fog will form. If the body of water is large enough, the wisps may coalesce into a dense fog. Sea smoke is a form of steam fog. **Warm-front fog** occurs at such fronts when warm rain falls through colder air beneath it. Under certain conditions, the evaporation of this rain can produce a supersaturated mass of air can be formed at ground level. Such fogs can be extremely dense. **Up-slope fogs** occur when warm, moist air is forced by wind up the slopes of mountains. As the air adiabatically cools, it can become supersaturated and a fog will form. Such fogs can be extremely dense.

Both radiation fogs and advection fogs tend to be relative thin in terms of height. Some fogs may be only a few meters thick. It is not unknown for "steam fogs" on lakes to be so dense that a person in a rowboat can barely see the front of his boat, but so thin that if that person stands up, his head is above the fog layer. Steam fogs seldom exceed 10 m thickness.[25] Valley fogs (a form of radiation fog) are often thin enough that only the valley bottoms are in the fog; hills fifty meters higher in altitude may be above the fog. Fifteen to one hundred meters is a reasonable estimate for the thickness (height) of any fog layer, although thicknesses as high as 200-300 meters are not unknown.

The World Meteorological Organization has recognized ten basic cloud types or genera. The descriptions of these ten genera are drawn from their International Cloud Atlas [26], [25]. The cloud types and their shorthand notations (two letters in parentheses) are:

(Ci) **Cirrus** – Detached clouds in the form of white, delicate filaments or white or mostly white patches or narrow bands. These clouds have a fibrous (hair-like) appearance, or a silky sheen, or both.

(Cc) **Cirrocumulus** – Thin, white patch, sheet, or layer of cloud without shading, composed of very small elements in the form of grains, ripples, etc., merged or separate, and more or less regularly arranged; most of the elements have an apparent width of less than one degree.

(Cs) **Cirrostratus** – Transparent, whitish cloud veil of fibrous (hair-like) or smooth appearance, totally or partly covering the sky, and generally producing halo phenomena.

(Ac) **Altocumulus** – White or grey, or both white and grey, patch, sheet or layer of cloud, generally with shading, composed of laminae, rounded masses, rolls, etc., which are sometimes partly fibrous or diffuse, and which may or may not be merged; most of the regularly arranged small elements usually have an apparent width of between one and five degrees.

(As) **Altostratus** – Greyish or bluish cloud sheet or layer of striated, fibrous or uniform appearance, totally or partly covering the sky, and having parts thin enough to reveal the sun at least vaguely, as through ground glass. Altostratus does not show halo phenomena.

(Ns) **Nimbostratus** – Grey cloud layer, often dark, the appearance of which is rendered diffuse by more or less continually falling rain or snow which in most cases reaches the ground. It is thick enough throughout to blot out the sun. Low, ragged clouds frequently occur below the layer with which they may or may not merge.

(Sc) **Stratocumulus** – Grey or whitish, or both grey and whitish, patch, sheet or layer of cloud which almost always has dark parts, composed of tessellations, rounded masses, rolls, etc., which are non-fibrous (except for virga – precipitation trails) and may or may not be merged; most of the regularly arranged small elements have an apparent width of more than five degrees.

(St) **Stratus** – Generally grey cloud layer with a fairly uniform base, which may give drizzle, ice prisms or snow grains. When the sun is visible through the cloud its outline is clearly discernible. Stratus does not produce halo phenomena (except possibly at very low temperatures). Sometimes stratus appears in the form of ragged patches.

(Cu) **Cumulus** – Detached clouds, generally dense and with sharp outlines, developing vertically in the form of rising mounds, domes, or towers, of which the bulging upper part often resembles a cauliflower. The sunlit parts of these clouds are mostly brilliant white; their bases are relatively dark and nearly horizontal. Sometimes cumulus is ragged.

(Cb) **Cumulonimbus** – Heavy and dense cloud, with a considerable vertical extent, in the form of a mountain or huge towers. At least part of its upper portion is usually smooth, or fibrous, or striated, and nearly always flattened; this part often spreads out in the shape of an anvil or vast plume. Under the base of this cloud, which is often very dark, there are frequently low ragged clouds either merged with it or not, and precipitation, sometimes in the form of virga.

Figure 3-27 provides a rough illustration of each of these cloud types along with the altitudes at which they are commonly encountered.

The characteristics of the ten cloud types are given in Table 3-5. Summarized here are typical base heights (the height of the base of the cloud above the ground), cloud thicknesses, temperatures at the base of the cloud, constituents (the physical form of water comprising the cloud), and associated precipitation (if any). The ten cloud types are commonly aggregated into three or four classes based on their location in the atmosphere. The three basic classes are low (cloud base below 2 km altitude), medium (cloud base between 2 and 7 km), and high (cloud base between 5 and 13 km). Low clouds are sometimes further distinguished between those with and without extensive vertical development (vertical structure and extended thickness). A quick examination of the list also reveals that clouds can also be classified by general shape: stratiform (layered), cumuliform (heaped), or cirriform (fibrous).

To permit prediction of attenuation in the visible and the infrared, several cloud models have been developed for use with LOWTRAN and its related codes. These include cumulus, altostratus, stratus, stratocumulus, and nimbostratus cloud types. Models have been developed for two kinds of cirrus (small ice crystals averaging 4 μm in length and large ice crystals averaging 64 μm in length) as well. For the five low to medium cloud types, the liquid water content (the weight of water in liquid form per unit volume of cloud) is shown graphically in Figure 3-28 as a function of altitude and in tabular form in Table 3-6. Table 3-7 summarizes the characteristics of the particles that make up all of the LOWTRAN hydrometeor models. Each "cloud" is assumed to have a number density of particles versus particle size that behaves as

$$n(r) = ar^{\alpha}e^{-br^{\gamma}} \tag{3.20}$$

with the total number density of particles being obtained by integration of the expression above, and

Figure 3-27. The ten basic cloud types.

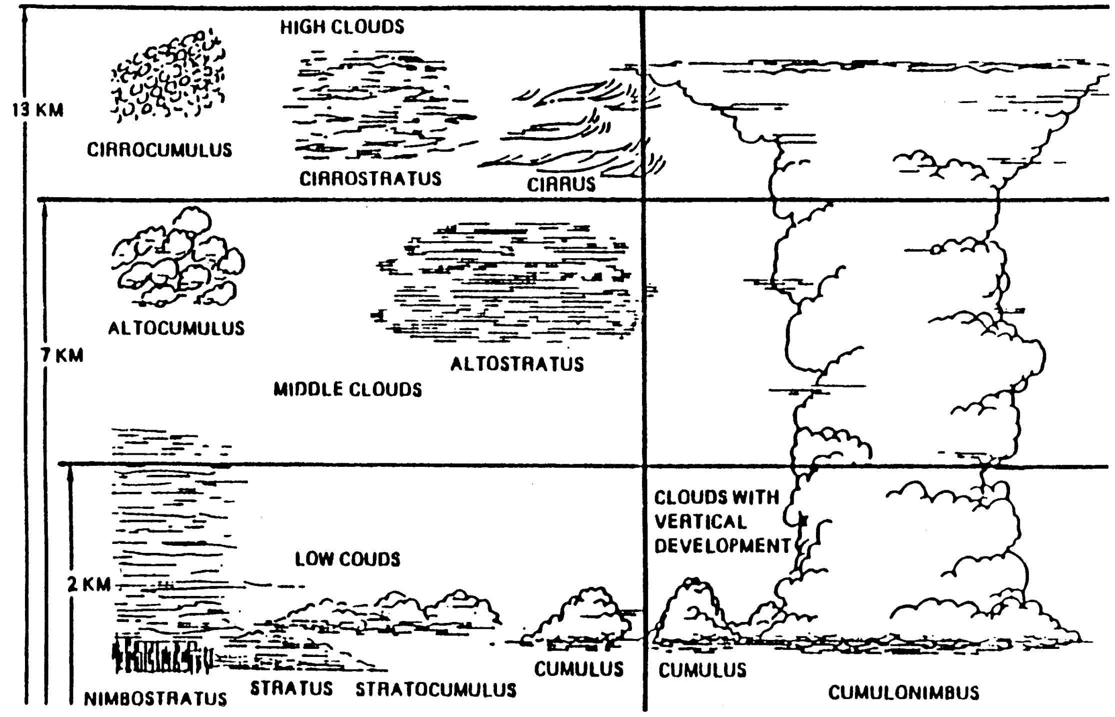


Table 3-5. Characteristics of the ten cloud genera.

CLOUD GENUS (Type)	CLOUD CONSTITUENTS	BASE HEIGHT (km)	CLOUD THICKNESS (km)	BASE TEMP. (°C)	TYPE OF PRECIPITATION
Ci (High)	Ice Crystals	7 - 10	0.1-0.2 thin to 1 thick	-20 to -60	None
Cc (High)	Ice Crystals	6 - 8	0.2 - 0.4	-20 to -60	None
Cs (High)	Ice Crystals	6 - 8	0.1 to	-20 to -60	None Several
Ac (Middle)	Water Droplets, Ice Crystals, or Snowflakes	2 - 6	0.1 - 1.0	+10 to -30	Rare, Light Snow or Rain
As (Middle)	Ice Crystals (thin cloud) Water Droplets (thick cloud)	3 - 5	0.1 thin, 6-8 thick	+10 to -30	Light Rain, Light to Medium Snow
Sc (Low)	Water Droplets (7 - 8 Microns)	0.5 - 1.5	0.2 - 0.8	+15 to -5	Usually none; Occasional Lt. Rain or Snow
St (Low)	Water Droplets (4 Microns)	0.1 - 0.7	0.2 - 0.8	+20 to -5	Drizzle, Light Rain or Snow
Ns (Vert. Devel.)	Ice Crystals, Water Droplets (10 Microns)	0.1 - 1.0	1 to Several	+10 to -5	Continuous Rain or Snow, Light to Heavy
Cu (Vert. Devel.)	Water Droplets (6 Microns)	0.8 - 1.5	0.15 - 1.5	+15 to -5	Occasional Light Rain or Snow
Cb (Vert. Devel.)	Water Drops, Ice, Ice Crystals, Hail	0.4 - 1.0	3 - 15	+15 to -5	Moderate to Heavy Rain, Snow, or Hail

Figure 3-28. Liquid water content vs. altitude of the LOWTRAN cloud models.[11]

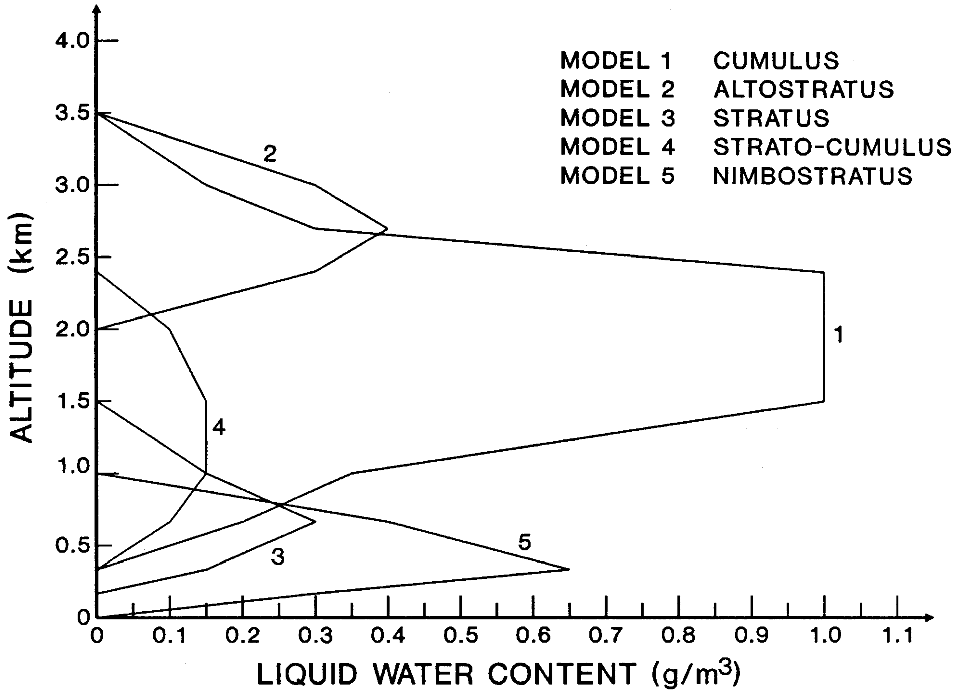


Table 3-6. Characteristics of the LOWTRAN cloud and rain models.[11]

LAYER HEIGHT (km)	CLOUD MODELS					CLOUD AND RAIN MODELS				
	LIQUID WATER CONTENT (g/m^3)					RAIN RATE (mm/hr)				
	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5	MODEL 6	MODEL 7	MODEL 8	MODEL 9	MODEL 10
	CUMULUS	ALTO-STRATUS	STRATUS	STRATO-CUMULUS	NIMBO-STRATUS	DRIZZLE FROM STRATUS	LIGHT RAIN FROM NIMBO-STRATUS	MODERATE RAIN FROM NIMBO-STRATUS	HEAVY RAIN FROM CUMULUS	EXTREME RAIN FROM CUMULUS
0.00	0.00	0.00	0.00	0.00	0.00	2.00	5.0	12.5	25.0	75.0
0.16	0.00	0.00	0.00	0.00	0.30	1.78	4.0	10.5	21.5	70.0
0.33	0.00	0.00	0.15	0.00	0.65	1.43	3.4	8.0	17.5	65.0
0.66	0.20	0.00	0.30	0.10	0.40	1.22	2.6	6.0	12.0	60.0
1.00	0.35	0.00	0.15	0.15	0.00	0.86	0.8	2.5	7.5	45.0
1.50	1.00	0.00	0.00	0.15	0.00	0.22	0.2	0.8	4.2	20.0
2.00	1.00	0.00	0.00	0.10	0.00	0.00	0.0	0.2	2.5	12.5
2.40	1.00	0.30	0.00	0.00	0.00	0.00	0.0	0.0	1.0	7.0
2.70	0.30	0.40	0.00	0.00	0.00	0.00	0.0	0.0	0.7	3.5
3.00	0.15	0.30	0.00	0.00	0.00	0.00	0.0	0.0	0.2	1.0
3.50	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.2
4.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.0

Table 3-7. Characteristics of the LOWTRAN hydrometeor particle distributions.[11],[27]

				NUMBER	LIQUID H_2O	MODAL RADII		$0.55\mu m$
HYDRO-	SIZE DISTRIBUTION PARAMETERS			DENSITY	CONTENT	(No.)	(Mass)	EXTINCT.
METEOR	(γ = 1 for all distributions)			N	W	R_N	R_M	k
TYPE	α	a	b	(cm^{-3})	($g\text{-}m^{-3}$)	(μm)	(μm)	(km^{-1})
HEAVY FOG, ADVECTION	3	0.27	0.3	20	0.37	10.0	20.0	28.74
MED. FOG, RADIATION	6	607.5	3.0	20	0.02	2.0	3.0	8.672
CUMULUS	3	2.604	0.5	250	1.00	6.0	12.0	130.8
STRATUS	2	27.0	0.6	250	0.29	3.33	8.33	55.18
STRATO-CUMULUS	2	52.734	0.75	250	0.15	2.67	6.67	35.65
ALTO-CUMULUS	5	6.268	1.111	400	0.41	4.5	7.2	91.04
NIMBO-STRATUS	2	7.678	0.425	200	0.65	4.7	11.76	87.08
CIRRUS	6	$2.21\text{x}10^{-12}$	0.09375	0.025	0.06405	64.0	96.0	1.011
THIN CIRRUS	6	0.011865	1.5	0.5	$3.128\text{x}10^{-4}$	4.0	6.0	0.0831

is therefore,

$$N = a\gamma^{-1} b^{-(\alpha+1)/\gamma} \Gamma\left(\frac{\alpha+1}{\gamma}\right). \tag{3.21}$$

In the visible and infrared regions of the spectrum, the extinction of clouds is handled very similar to that of fogs. Figures 3-28 and 3-29 show the relative extinction coefficient and relative absorption of the five low to medium cloud types. The extinction coefficient at 0.55 μm of the cloud at maximum liquid water content (LWC) is given in the last column of Figure 3-7. The 0.55 μm extinction coefficient is scaled by the LWC as a function of position along the propagation path (Figure 3-27). It is further scaled to accommodate other wavelengths using Figure 3-28. Figures 3-30 and 3-31 show the relative extinction coefficient and relative absorption of the two cirrus models. Attenuation due to cirrus uses the 0.55 μm extinction coefficient multiplied by the cirrus thickness (user selectable) and further modified by the relative extinction data of Figure 3-30 to accommodate other wavelengths.

Figure 3-29. Relative extinction coefficient for five cloud aerosols.[11]

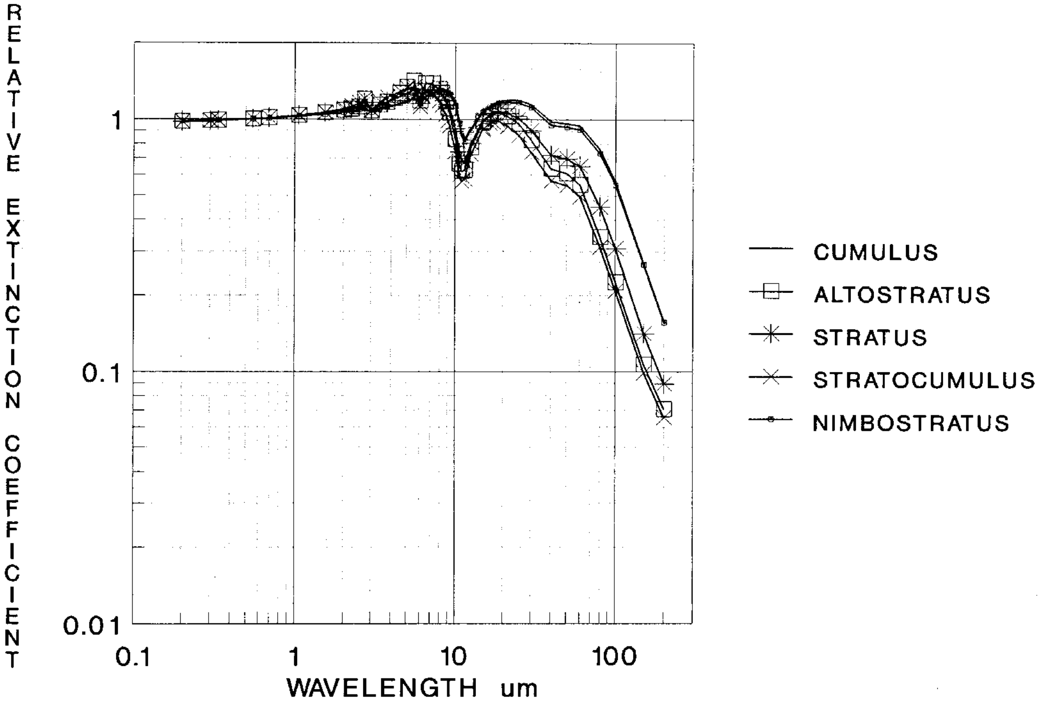


Figure 3-30. Relative absorption for five cloud aerosols.[11]

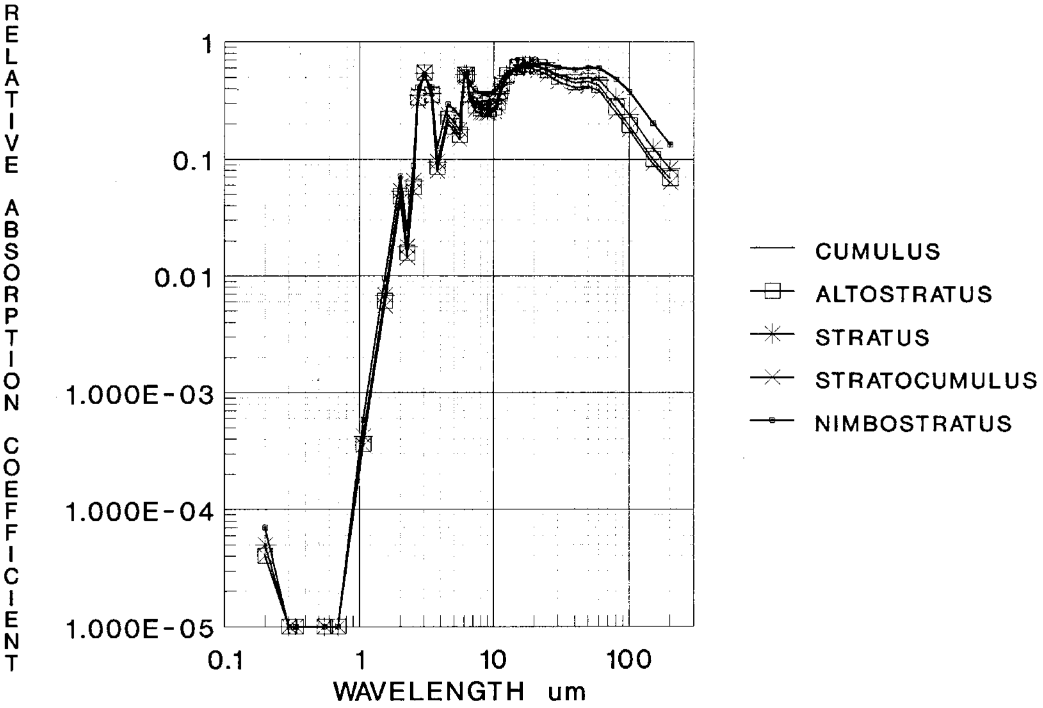


Figure 3-31. Relative extinction coefficient for two cirrus aerosols.[11]

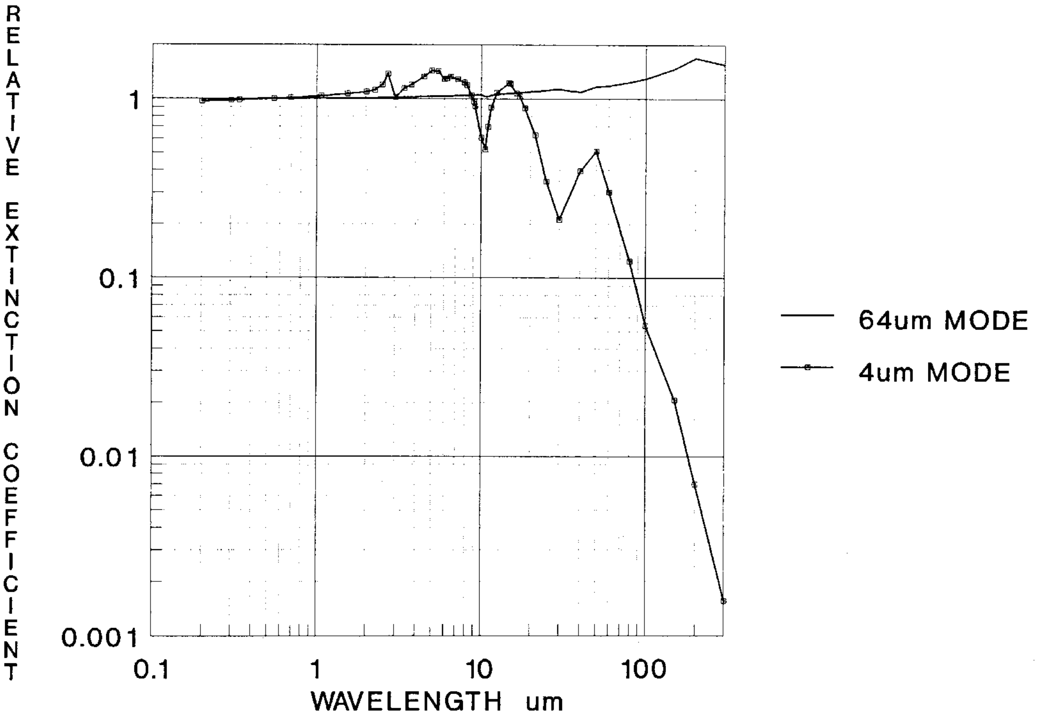
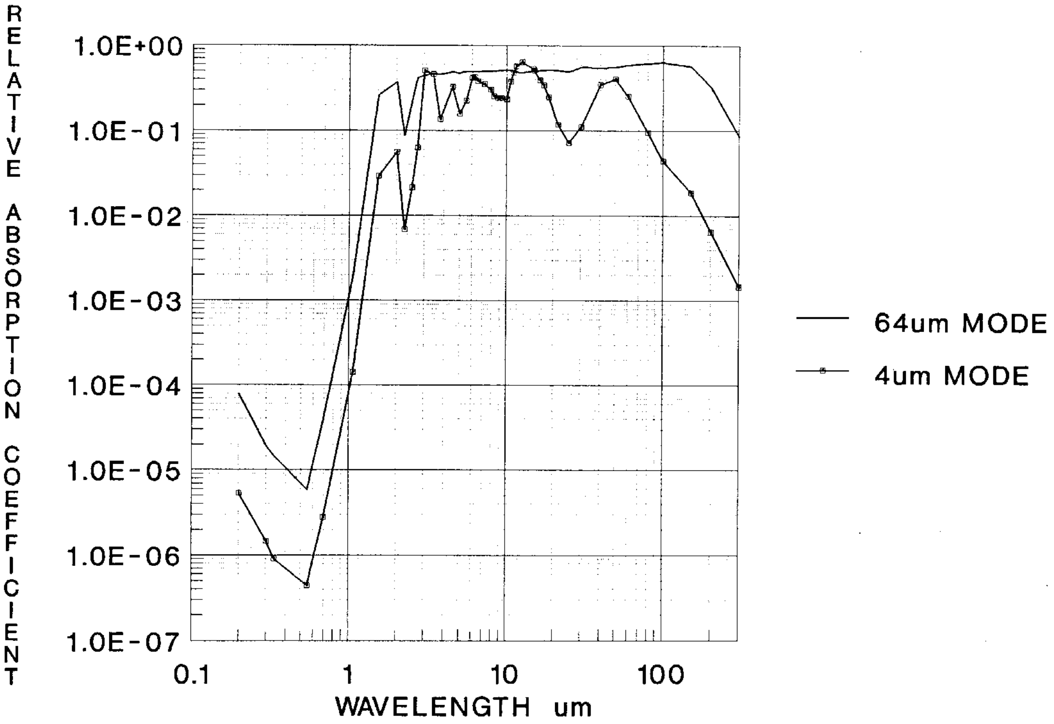


Figure 3-32. Relative absorption for two cirrus aerosols.[11]



In the microwave region, the attenuation from the small hydrometeors can be estimated from the complex refractive index. The refractivity *N* of air is related to the refractive index *n* by

$$N = (n-1)\times 10^{6} \qquad \text{ppm} \tag{3.22}$$

The refractive index of air is so close to one that large relative changes do not show up explicitly. The refractivity is a large number in which even small changes in refractive index can be observed directly. Many atmospheric constituents contribute directly to the refractivity. The contribution from liquid water, when the water droplets are small enough to be treated in the Rayleigh approximation, is given by [28]

$$N_W = \frac{\alpha_W}{0.182\nu} = 4.50\,w\left[\frac{\varepsilon''}{(\varepsilon'+2)^2+\varepsilon''^2}\right] \qquad \text{ppm} \tag{3.23}$$

where ν is the microwave frequency, α_W is the absorption coefficient (in dB/km), ε' and ε'' are the real and imaginary components of the dielectric constant. Using Debye theory, the dielectric constant component of water were determined to be [29]

$$\varepsilon'(\nu) = \frac{(72.12+103.3(\theta-1))}{\left(1+(\nu/\nu_P)^2\right)} + \frac{(1.97)}{\left(1+(\nu/\nu_S)^2\right)} + 3.51 \tag{3.24}$$

and

$$\varepsilon''(\nu) = \frac{(72.12+103.3(\theta-1))(\nu/\nu_P)}{\left(1+(\nu/\nu_P)^2\right)} + \frac{(1.97)(\nu/\nu_S)}{\left(1+(\nu/\nu_S)^2\right)} \tag{3.25}$$

where the primary and secondary relaxation frequencies are

$$\nu_P = 20.09 - 142(\theta-1) + 294(\theta-1)^2 \quad \text{GHz} \tag{3.26}$$

and

$$\nu_S = 590 - 1500(\theta-1) \quad \text{GHz} \tag{3.27}$$

and θ is a reduced temperature given by

$$\theta = 300/T \tag{3.28}$$

The variants of these equations used in [28] have been used to determine the attenuation coefficient of hydrosols (fog or cloud) for a liquid water content (LWC) $w = 1$ g/m^3. These are shown in Table

Table 3-8. Hydrosol (haze, fog, cloud) attenuation for a mass concentration of $w = 1$ g/m^3.

TEMPERATURE		ATTENUATION COEFFICIENT (dB/km)					
		FREQUENCY (GHz)					
T	θ	1	10	30	100	200	300
0	1.0983	0.001	0.097	0.82	5.4	9.3	10.7
25	1.0062	0.001	0.051	0.45	4.2	10.8	15.3

3-8. The attenuation at other water contents scales linearly. If A is a coefficient from the table, then the LWC-dependent coefficient is

$$A(w) = A \cdot w\,. \tag{3.29}$$

In Table 3-9, the LWC data from Table 3-7 has been combined with the data in Table 3-8, to estimate the maximum attenuation coefficient provided by each of the LOWTRAN fog and cloud models. From the data it appears that fog and clouds are negligible contributors to attenuation at 1 GHz, marginal (<1 dB maximum path attenuation for reasonable paths) contributors at 10 GHz, but serious attenuation sources at 30 GHz and above.

Table 3-9. Maximum attenuation coefficients for the LOWTRAN hydrometeor models.

HYDRO-	TEMP.	MAX.	MAX. ATTENUATION COEFFICIENT (dB/km)					
METEOR	T	LWC	FREQUENCY (GHz)					
TYPE	(°C)	$(g-m^3)$	1	10	30	100	200	300
HEAVY FOG,	0	0.37	0.00037	0.0359	0.303	2.00	3.44	3.96
ADVECTION	25		0.00037	0.0189	0.167	1.55	4.00	5.66
MED. FOG,	0	0.02	2×10^{-5}	0.00194	0.0162	0.108	0.186	0.214
RADIATION	25		2×10^{-5}	0.00102	0.009	0.084	0.216	0.326
CUMULUS	0	1.00	0.001	0.097	0.82	5.4	9.3	10.7
	25		0.001	0.051	0.45	4.2	10.8	15.3
STRATUS	0	0.29	0.00029	0.0281	0.238	1.57	2.70	3.10
	25		0.00029	0.0148	0.131	1.22	3.13	4.44
STRATO-	0	0.15	0.00015	0.0146	0.123	0.810	1.40	1.61
CUMULUS	25		0.00015	0.0076	0.068	0.630	1.62	2.30
ALTO-	0	0.41	0.00041	0.0398	0.336	2.21	3.81	4.39
CUMULUS	25		0.00041	0.0209	0.185	1.72	4.43	6.27
NIMBO-	0	0.65	0.00065	0.0631	0.533	3.51	4.65	6.96
STRATUS	25		0.00065	0.0332	0.293	2.73	7.02	9.95

Hydrometeors - Precipitation

Rain attenuation can be determined from Mie scattering theory if the raindrop size distribution is known. Most measurements of the size distribution have attempted to fit the results to a curve of the form

$$\frac{dN}{dD} \equiv n(D) = N_0 e^{-\Lambda(R)\,D} \tag{3.30}$$

where N_0 is the number density scale factor, $\Lambda(R)$ is an exponent dependent on the rain rate R (in mm/hr), and D is the drop diameter (in mm). From Mie theory the extinction coefficient α can be written as [9]

$$\alpha = \int_0^\infty dD\, \frac{\pi D^2}{4}\, Q_{EXT}\left[\pi D/\lambda, m(\lambda)\right] \frac{dN}{dD} \tag{3.31}$$

where Q_{EXT} is the Mie Extinction Efficiency and $m(\lambda)$ is the complex refractive index of water. For raindrops in the visible and infrared spectral regions, $D >> \lambda$. Therefore, $Q_{EXT} \sim 2$, independent of wavelength or particle diameter. Thus, we have

$$\alpha_{RAIN} = \frac{\pi}{2} N_0 \int_0^\infty dD\, D^2\, e^{-\Lambda D} = \pi N_0 \Lambda^{-3} \tag{3.32}$$

LOWTRAN uses the Marshall-Palmer raindrop distribution

$$n(D) = 8000\, e^{-4.1 R^{-0.21} D} \tag{3.33}$$

which leads to the following expression for the extinction due to rain

$$\alpha_{RAIN} = 0.365 R^{0.63} \quad \text{km}^{-1} \tag{3.34}$$

or

$$A_{RAIN} = 1.585 R^{0.63} \quad \text{dB / km}\,. \tag{3.35}$$

In the microwave spectral region, this simplified analysis does not suffice. However, more sophisticated calculations have yielded tabular coefficients (Table 3-10) for calculating the attenuation of rain in the microwave as a function of frequency ν and polarization (subscript p = v for vertical, h for horizontal, and c for circular) using an expression of the form

$$A_{RAIN}(\nu) = a_p(\nu) R^{b_p(\nu)} \quad \text{dB / km}\,. \tag{3.36}$$

Table 3-10. Coefficients for rain attenuation calculations for horizontal, vertical, and circular-polarization.[30],[31]

ν (GHz)	a_h	a_v	a_c	b_h	b_v	b_c
1	0.000039	0.000035	0.000037	0.912	0.880	0.896758
2	0.000154	0.000138	0.000146	0.963	0.923	0.944096
4	0.00065	0.000591	0.000621	1.121	1.075	1.099093
6	0.00175	0.00155	0.00165	1.308	1.265	1.287803
7	0.00301	0.00265	0.00283	1.332	1.312	1.322636
8	0.00454	0.00395	0.004245	1.327	1.310	1.319091
10	0.0101	0.00887	0.009485	1.276	1.264	1.270389
12	0.0188	0.0168	0.0178	1.217	1.200	1.208978
15	0.0367	0.0335	0.0351	1.154	1.128	1.141593
20	0.0751	0.0691	0.0721	1.099	1.065	1.082707
25	0.124	0.113	0.1185	1.061	1.030	1.046219
30	0.187	0.167	0.177	1.021	1.000	1.011093
35	0.263	0.233	0.248	0.979	0.963	0.971484
40	0.350	0.310	0.330	0.939	0.929	0.934303
45	0.442	0.393	0.4175	0.903	0.897	0.900176
50	0.536	0.479	0.5075	0.873	0.868	0.870640
60	0.707	0.642	0.6745	0.826	0.824	0.825048
70	0.851	0.784	0.8175	0.793	0.793	0.793000
80	0.975	0.906	0.9405	0.769	0.769	0.769000
90	1.060	0.999	1.0295	0.753	0.754	0.753485
100	1.120	1.060	1.090	0.743	0.744	0.743486
120	1.180	1.130	1.155	0.731	0.732	0.731489
150	1.310	1.270	1.290	0.710	0.711	0.710492
200	1.450	1.420	1.435	0.689	0.690	0.689495
300	1.360	1.350	1.355	0.688	0.689	0.688498
400	1.320	1.310	1.315	0.683	0.684	0.683498

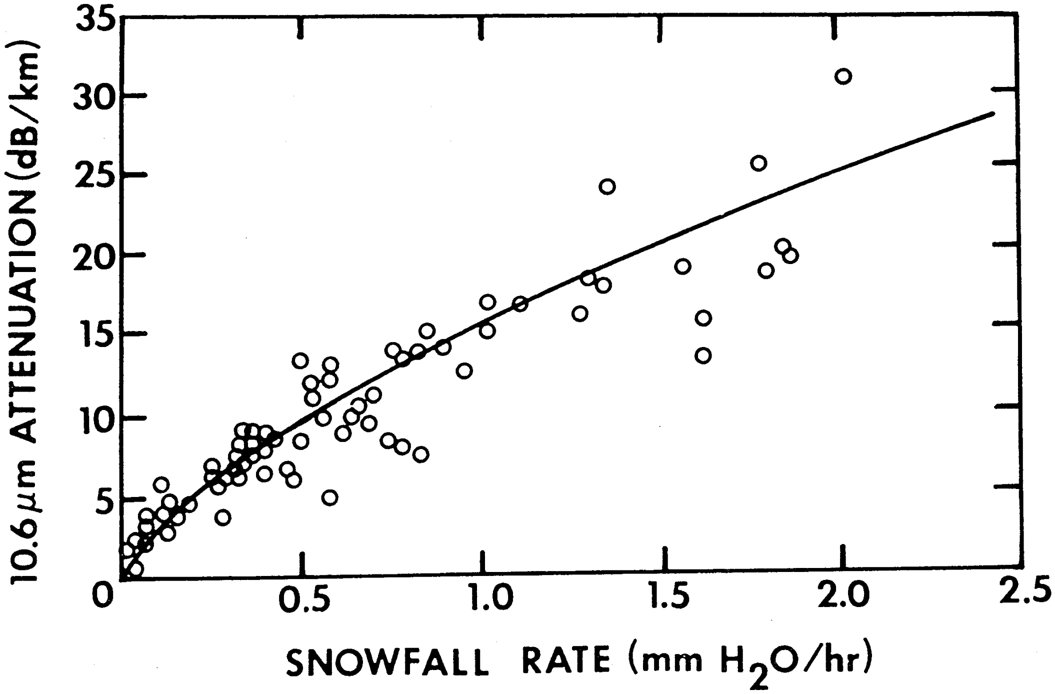
The preceding expressions are applicable to determining the attenuation due to rain. They are not applicable to snow. There are many fewer measurements of snow attenuation, and due to the non-spherical character of the particles, still fewer theoretical analyses. Let us consider some of the available results. If the visibility *V* through the snow is known, then the extinction (scattering – the extinction of snow will be mostly scattering) coefficient can be determined from the relation [32]

$$\alpha = 3.912 / V \quad km^{-1} . \tag{3.37}$$

This relation can be used with some reservations in the visible and infrared spectral regions (because the Mie scattering efficiency should be in the constant "optical" region and should be independent of wavelength). This relation can also be used for rain attenuation.

If the visibility is not known, we must rely on a few measurements. Figure 3-33 shows the data obtained at 10.6 μm during one snowfall in the Soviet Union.[33]

Figure 3-33. Attenuation at 10.6 μm wavelength as a function of snowfall intensity.[33]



The Soviet authors also measured attenuation at 0.63 μm. They obtained the following fits to their data

$$A_{10.6} = 15.1R^{0.71} \quad \mathrm{dB/km} \tag{3.38}$$

and

$$A_{0.63} = 10.8R^{0.53} \quad \mathrm{dB/km} \tag{3.39}$$

Despite the statements made in the preceding paragraph there appears to be some wavelength dependence to the attenuation. Given the complexity of the shapes of snowflakes, there is no reason that the attenuation in the visible and infrared must be independent of wavelength. Nevertheless, the differences between attenuations at different wavelengths are small compared to the attenuation differences between snow and rain. The author fit an equation of the Marshall-Palmer $R^{0.63}$ form to the 10.6 μm Soviet data and obtained

$$\alpha_{SNOW} = 3.7R^{0.63} \quad \mathrm{km^{-1}} \tag{3.40}$$

and

$$A_{SNOW} = 16R^{0.63} \quad \mathrm{dB/km}\,. \tag{3.41}$$

The author's fit gives a result for snow that is almost exactly a factor of ten higher extinction coefficient than that for rain.

A Defense Department handbook [32] on atmospheric attenuation give several data points for snow attenuation in the millimeter-wave region of the spectrum. Table 3-11 presents coefficients of the equation

$$\alpha_{SNOW} = cR^d \quad \text{km}^{-1} \tag{3.42}$$

for three kinds of snow: dry snow (powder snow), moist snow (small flakes that form firm snowballs when compacted), and wet snow (large "gloppy" particles that readily stick together).

Table 3-11. Millimeter-wave snow extinction parameters.[32]

SNOW	35 GHz		94 GHz	
TYPE	c	d	c	d
DRY	0.0125	1.60	0.08	1.26
MOIST	0.160	0.95	0.31	0.75
WET	0.235	0.95	0.61	0.75

This limited data gives snowfall rate dependences which are considerably different than those in the visible and infrared, and although not considerably different from rain. Additional snow data can be found in Reference [34]. They do little to clarify the overall picture. About the only things that can be said about attenuation due to snow, is (1) it will be significant from the millimeter-wave region to the visible region of the spectrum, and (2) snow attenuation will be higher than rain attenuation for the same precipitated equivalent liquid water content.

Statistics of Weather Effects

Weather is a random variable. We are unable to accurately predict the weather in most locations more than a few days into the future. However, using statistics of past weather conditions, we can estimate with reasonable accuracy the fraction of the time that given conditions will occur. This statistical data is of even more importance to system designers than point data. For example, if we know that rain greater than 5 mm/hr will degrade the performance of a sensor design, it is useful to know what fraction of the time in the locales of interest that the rain rate will exceed 5 mm/hr. If the number is very large we will have a problem. If it is very small, the degradation may be deemed insignificant.

A number of groups at various times have for their own reasons collected and analyzed weather statistics. Occasionally, this data is made available to the community as a whole. One such set of statistical data was published by Dynetics, Inc. for the Army Materiel Command, Smart Weapons Management Office, in the form of cumulative probability curves on a wall chart.[35] These curves are reproduced here as Figures (3-34) through (3-37). Cumulative probability curves for visibility, humidity, rain rate, and cloud base height are presented for three nominal geographic environments: European Lowlands, Mideast Deserts, and Central America. A cumulative probability curve is a plot of the "*Probability that Variable X has a value less than or equal to x*" as a function of "*x*". When X is at its minimum value, $P(X_{min}) = 0$. When X is at its maximum value, $P(X_{max}) = 1$. When X is at its median value, $P(X_{median}) = 0.5$.

Cumulative probability curves are often used in the following manner. A probability value that is acceptable from a system perspective is selected. For example, it may be acceptable for a system to suffer degradation from excessive humidity a maximum of 10% of the time in Central America. That is, the humidity must be below some as yet unknown maximum acceptable value for 90% of the time. Picking an 0.9 cumulative probability and finding the curve for humidity in Central America (Figure 3-35, we find a humidity value of slightly less than 21 g/m^3. This value now become a derived system specification. That is, the system must work without degradation at a value of 21 g/m^3. If it can do that then it will work at least 90% of the time in Central America (barring other variables that also limit performance).

In another example, it was desired that a sensor operate under at least 95% of all visibility conditions. This requires that the cumulative probability of having the an unacceptable visibility can only be 5%. Looking at Figure 3-34 we find that a 7 km visibility is acceptable in Central America and the Mideast, but in the European Lowlands, the 5% cumulative probability value corresponds to only 1.5 km. The systems engineer is left with a decision as to whether a rather benign 7 km visibility requirement (i.e., the sensor must perform adequately when the visibility is as low as 7 km) or a rather severe 1.5 km visibility requirement should be imposed. The resolution of this problem would depend on the exact nature of the mission need.

The four variables in the Dynetics data set are generally very useful, but they are by no means complete. Data is desirable for other geographic locations (such as Southeast Asia, the

Figure 3-34. Cumulative probability for visibility in three geographic locations.[35]

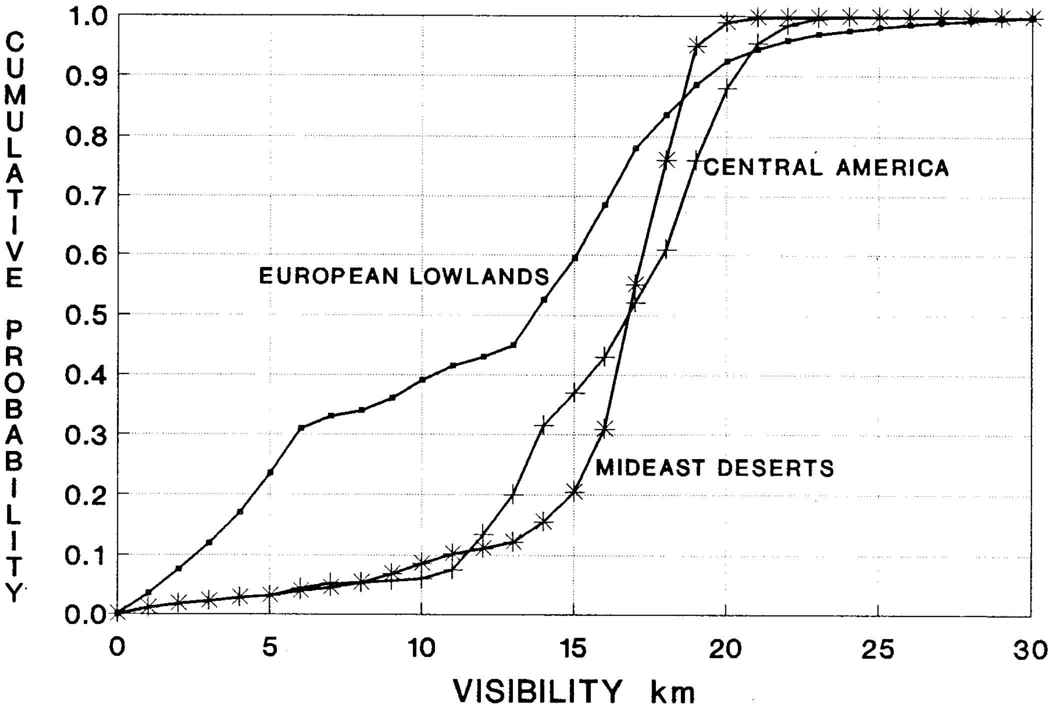


Figure 3-35. Cumulative probability for absolute humidity in three geographic locations.[35]

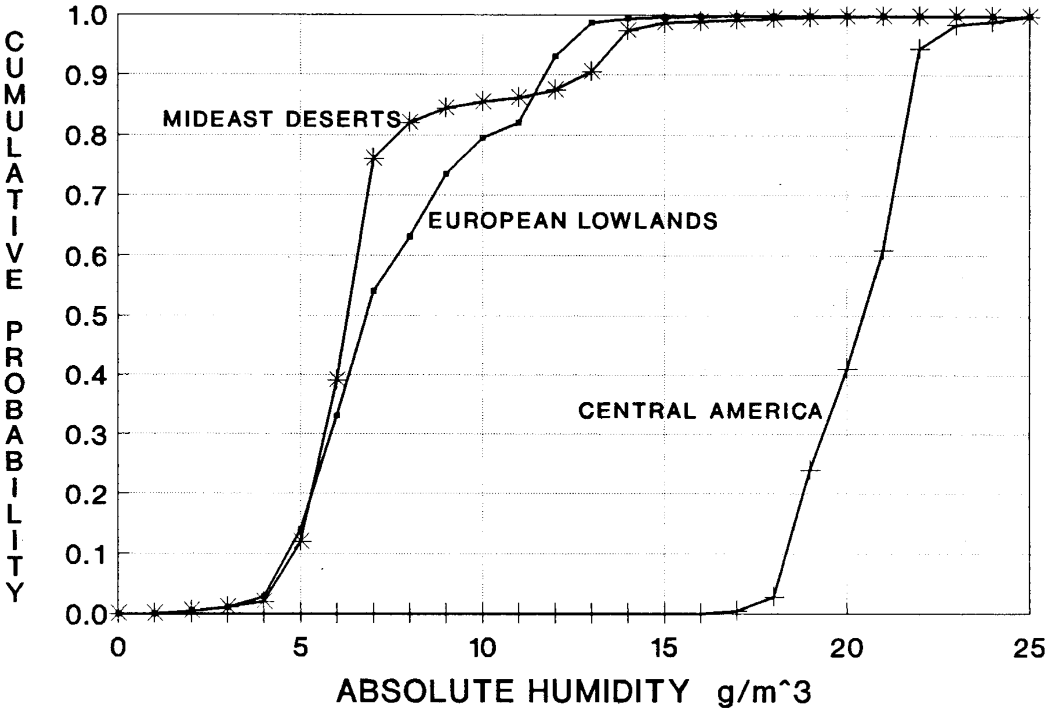


Figure 3-36. Cumulative probability for rain rate in three geographic locations.[35]

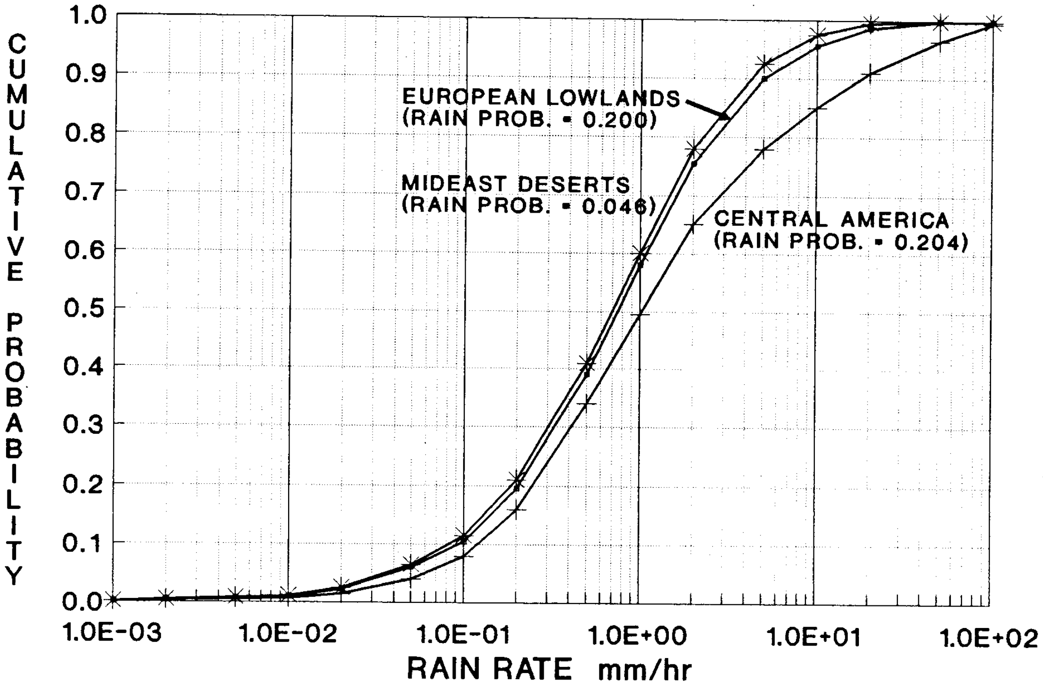
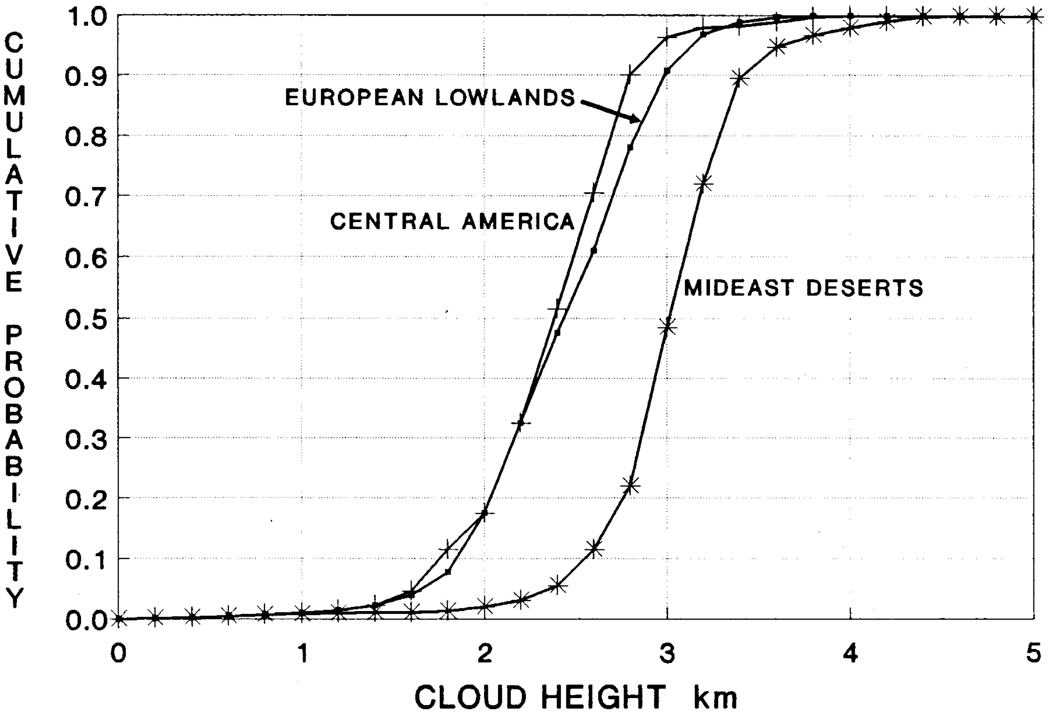


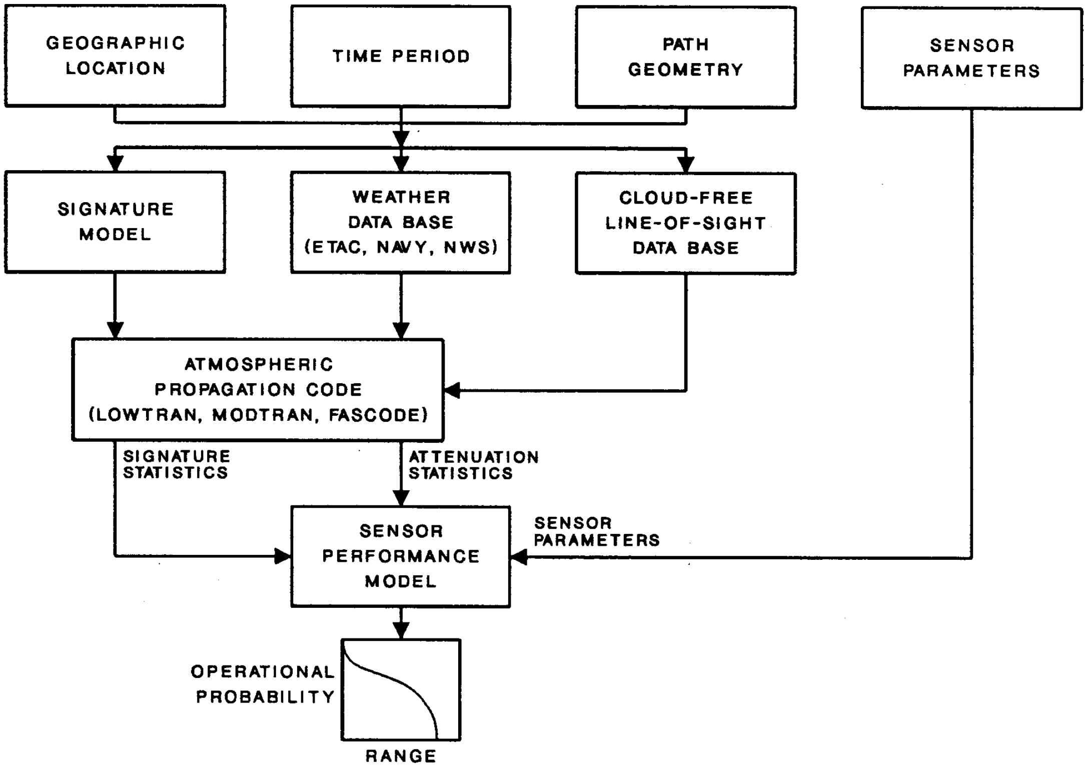
Figure 3-37. Cumulative probability for cloud base height in three geographic locations.[35]



Korean Peninsula, or the Antarctic) and for other meteorological variables (including wind speed, ambient pressure, air temperature, percent cloud cover, etc).

A relatively complete database was developed by Dr. Ken Hepfer of the Dahlgren Division of the Naval Surface Warfare Center (NSWC) in June 1993 for the Electro-Optical Cost and Operational Effectiveness Analysis (COEA).[36] The observations were "randomly" selected from thousands of weather ship measurements using a filter that guaranteed that the database included an equal number of observations (96 each) from the Baltic Sea, the Yellow Sea, the Gulf of Oman, and the Caribbean Sea (25% North latitudes, 50% mid-latitudes, and 25% tropics) equally divided between the months of the year (8 observations per month per location). Each observation contains all of the weather data collected by each site. This includes air temperature, sea temperature, pressure, relative humidity, absolute humidity, visibility, wind speed, wind direction, rain rate, cloud cover, and turbulence strength. This "R384" database has two main uses in sensor system analysis. In one application, a maritime sensor-plus-atmospheric model can be exercised using each of the 384 weather observations. The 384 performance predictions provide the basis for a "world-wide" operational utility analysis (see Figure 3-38), that is, an analysis which precisely defines what fraction of the time during a year, a given season, or given time of day, a stated or desired level of performance can be obtained. Such analyses can serve as a means to fairly compare sensors with

Figure 3-38. Elements of an operational utility analysis.



possibly significantly different design parameters, design philosophies, or principle of operation.

In the second application, the 384 observations can be input to a standard model such as LOWTRAN to determine values for atmospheric transmission in a few explicit wavebands. NSWC performed such calculations for several infrared wavebands and several millimeter-wave wavelengths. The results were then statistically analyzed to develop cumulative probability of attenuation curves. For the infrared bands, the attenuation was modeled as

$$T = e^{-\alpha R^{\beta}} \tag{3.43}$$

where T is the transmission over a range R (in km), α is an "attenuation factor" and β is a correction factor that account for non-uniform attenuation across a waveband. For a very narrow waveband, β would reduce to 1, and α would become the extinction coefficient. Table 3-12 shows the two attenuation coefficients α and β for the 3-5 μm and 8-12 μm bands at a number of selected percentile points of the cumulative probability distributions. For example, if one wanted to design a 3-5μm sensor to work in at least 80% of world-wide maritime weather conditions, then from the 80-percentile numbers, one should design the sensor to work with α = 0.75743 and β = 0.57004. Alternatively, if an 8-12 μm sensor was capable of operating in weather with α = 0.72255 and β = 0.92409, then that sensor should be capable of operating at least 95% of the time anywhere in the world's maritime environment.

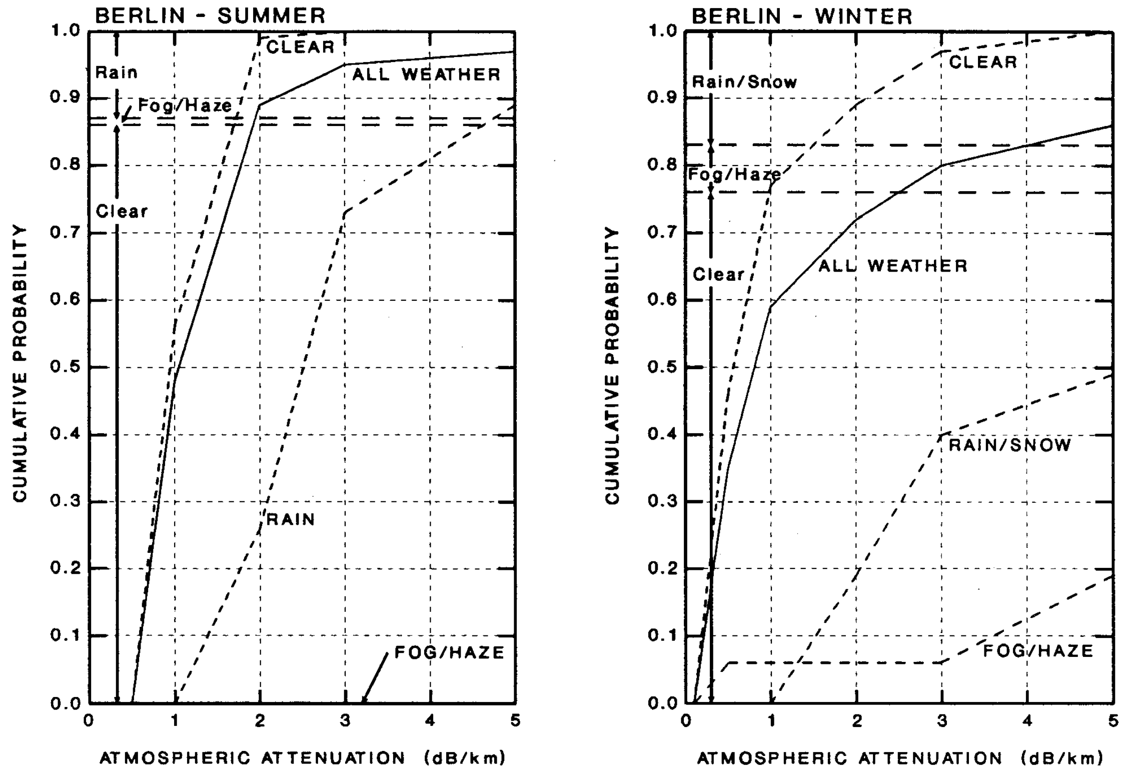
Table 3-12. R384 Attenuation coefficients for the 3-5 μm and 8-12 μm bands at various percentiles.[36]

	3-5 μm		8-12 μm	
PERCENTILE	α	β	α	β
5	0.45039	0.58359	0.12151	0.84706
10	0.48351	0.58219	0.14316	0.85943
25	0.56137	0.57989	0.21926	0.88006
50	0.67340	0.57156	0.41487	0.88614
70	0.72825	0.56431	0.52054	0.88492
80	0.75743	0.57004	0.57288	0.88366
85	0.77748	0.57151	0.59310	0.88554
90	0.81578	0.58742	0.63084	0.88640
95	0.98757	0.69721	0.72255	0.92409

In the mid-1970's, Kleiman and Modica [37] at M.I.T. Lincoln Laboratory used data from the Rand Weather Data Bank [38] to perform a similar probabilistic analysis. In 1974, this database had weather reports from 83 locations around the world, with an average of 12 years of continuous reporting from each site (during the 1950's and 1960's), mostly at hourly or 3-hourly intervals. Kleiman and Modica used all of the data from 9 of the 83 weather stations in the archive and performed a LOWTRAN3 calculation for each weather observation. From these calculations they were able to estimate the cumulative extinction probability curves for several different sensor systems at different times of the year and undertaking different assumption. The calculated data for Berlin during the summer and winter seasons at 10.6 μm wavelength are shown in Figure 3-39. Please note that LOWTRAN has been modified many times since 1975 to considerably improve its accuracy. The data in Figure 3-39 should be assumed to display trends and relative values, but are not quantitatively accurate.

There are several kinds of information summarized in these curves. The solid curve is the overall "all-weather" cumulative probability curve. For example, we can use the Berlin-Winter curve to determine that 80% of time during the winter, the extinction was less than 3 dB/km. The horizontal dashed lines divide 100% probability into three weather conditions: rain/snow, fog/haze,

Figure 3-39. Cumulative atmospheric extinction probability curves for Berlin during the summer and the winter.[36]



and clear. The height of each region tells the percentage of the time that each of the three conditions occurred. The dashed curves are the cumulative probabilities conditional on the stated weather condition existing. For example, in the summer, if it was raining, then 73% of the time the attenuation was less than 3 dB/km and it rained 12% of the time. If it was foggy or hazy in the summer (a 1-2% occurrence), then the attenuation was never less than 5 dB/km. Summer fogs are certified 10 μm sensor killers.

Cumulative extinction probability curves are useful for determining when conditions of "diminishing returns" exist. Consider the summertime curve. A sensor capable of working in 3 dB/km attenuation will work 95% of the time. In order to boost the performance percentage to 97%, the sensor must be designed to work in 5 dB/km attenuation. If the nominal operating range is 10 km, then the difference between 3 dB/km and 5 dB/km is 20 dB for a passive sensor and 40 dB for an active sensor. A sensor designed to handle 20-40 dB higher attenuation will of necessity be bigger, consume more power, weigh more, and cost considerably more. Seldom is the cost involved worth the benefit of only 2% more operational utility. The "knee" of the all-weather attenuation curve occurs at 3 dB/km. It is common that the knee of a cumulative probability curve represents a point of diminishing returns. Sensor systems designers need to be aware of this to avoid costly overdesign.

Such cumulative probability analyses are so valuable, that it is strange that they are not done more often. When Kleiman and Modica did their study in 1975, the calculations required hundreds, if not thousands of hours of mainframe time. Today, the same calculations could be done in a few hours on a laptop computer. The computational resources are now ubiquitous, the codes are still available, and the weather data is available from one source or another. This author recommends that this type of probabilistic analysis be performed on every new sensor design.

There are many sources of cumulative probability data that can be used by the sensor designer and the systems engineer to evaluate performance. Hundreds of measurement programs have been conducted by government agencies, universities, and private companies to facilitate analysis of their own sensor developments. Unfortunately, it is difficult to find all of this information. Some of it is published as adjunct technical data supporting a system, some of it is published in formal technical reports, much of it is published in short course notes presented at technical society meetings by the experts that collected the data, some is published in short notes to trade magazines, and some of it is only "published" in internal memoranda or viewgraph briefing packages (that sometimes find sizable external distribution). Most successful sensor design engineers and systems engineers actively seek out and acquire a substantial amount of this data during their careers. New entrants into these professions would do well to acquire as much of this data as is possible from the older professionals they work with. It is often these compilations that let the "expert" professionals generate the quantity of creative syntheses and analyses that earned them the title of expert in the first place. The author cannot reproduce every such data set in his own collection. He has included a few that seem most likely to be of benefit. Figure 3-40 shows the cumulative probability of attenuation in the 8-12 μm thermal imaging band obtained during the winter months at Meppen, Germany during 1976-77 by Project OPAQUE.[39] Figure 3-41 compares 10.6 μm and 1.06 μm attenuation data collected by Plessey Radar Ltd. at their Isle of Wight facility during October 1974 to May 1975.

Figure 3-40. Cumulative probability of 8-12 μm attenuation obtained during Project OPAQUE. Data site was Meppen, Germany and the collection period was the Winter of 1976-77. [39]

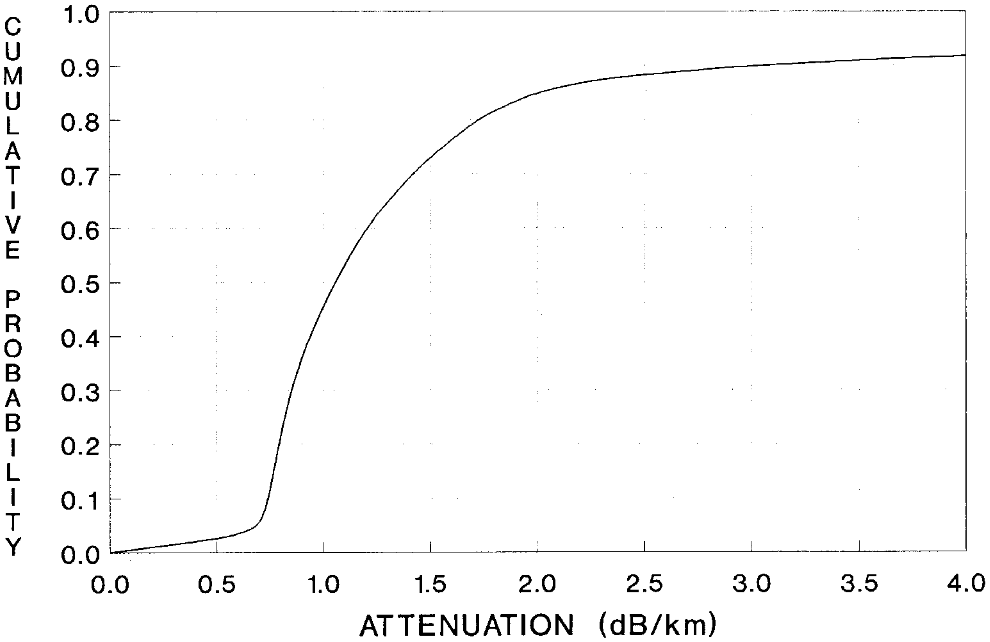
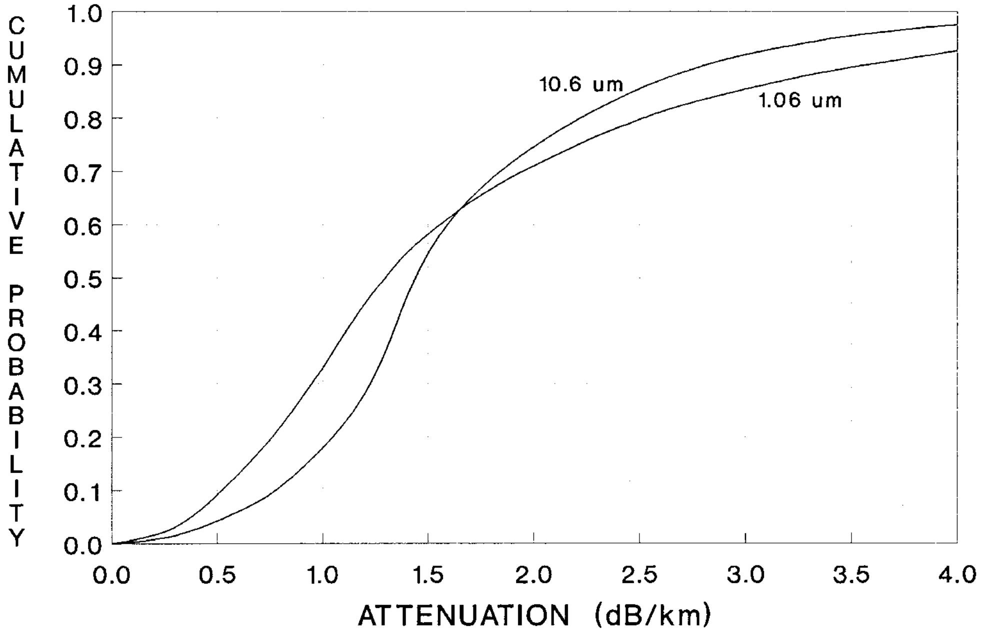


Figure 3-41. Cumulative probability of laser line attenuation obtained by Plessey Radar Ltd. Data site was the Isle of Wight, UK and the collection period was October 1974 to May 1975.



The attenuation due to any clouds (except cirrus) is usually so high in the visible and infrared that clouds are considered opaque. The utility of visible and infrared systems looking up (or down) through the atmosphere depends on the ability to find a cloud-free line of sight. Kleiman and Modica analyzed this problem and came up with a table which relates the probability of obtaining a cloud-free line of sight (CFLOS) as a function of line of sight elevation angle and the meteorological cloud cover.[37] Cloud cover is often reported as the number of eighths of the sky that is obscured by clouds. A four (eighths) rating implies that 4/8 = 50% of the sky is obscured. The results of the analysis are summarized in Table 3-13. It should be noted that even at zero (eighths) – nominally cloudless – there is still a small possibility of having a cloud pass through the line of sight. At eight (eighths) – nominally total overcast – there is a small possibility of finding a small hole in the cloud cover.

Table 3-13. Probability of cloud-free line of sight as a function of partial cloud cover (eighths) and line of sight elevation angle.[37]

ELEV.			PROBABILITY OF CLOUD-FREE LINE OF SIGHT						
ANGLE			CLOUD COVER (Eighths)						
(Deg)	0	1	2	3	4	5	6	7	8
0	0.962	0.640	0.490	0.370	0.275	0.195	0.095	0.040	0.020
5	0.965	0.730	0.600	0.490	0.380	0.280	0.180	0.090	0.025
10	0.970	0.820	0.695	0.590	0.485	0.360	0.250	0.130	0.030
15	0.973	0.875	0.750	0.645	0.540	0.420	0.300	0.190	0.035
20	0.975	0.895	0.805	0.700	0.600	0.490	0.365	0.225	0.040
25	0.978	0.910	0.825	0.730	0.645	0.525	0.400	0.250	0.045
30	0.980	0.920	0.850	0.775	0.680	0.575	0.430	0.275	0.050
35	0.983	0.930	0.860	0.790	0.695	0.595	0.460	0.290	0.055
40	0.985	0.940	0.875	0.800	0.710	0.610	0.480	0.300	0.060
45	0.988	0.945	0.880	0.815	0.725	0.630	0.500	0.320	0.065
50	0.990	0.950	0.890	0.825	0.740	0.650	0.510	0.325	0.070
55	0.993	0.952	0.895	0.830	0.750	0.660	0.520	0.330	0.075
60	0.995	0.954	0.898	0.840	0.765	0.675	0.525	0.340	0.080
90	1.000	0.960	0.900	0.850	0.780	0.680	0.545	0.350	0.085

Another useful set of statistics is the cumulative probability of rainfall rate as a function of geographic location. Barring explicit data sets such as those in Figure 3-34 there are several other ways to obtain rain rate estimates. Figure 3-42 shows the worldwide rain climate zones used at one time by the Consultative Commission on International Radio of the ITU.[31] Table 3-14 shows data on the percentage of time each year that the rain rate exceeds specified levels (this is cumulative probability data presented in a slightly different format. Similar plots and tables have been generated by Crane.[40] The existence of local microclimates (small regions a few square kilometers in size where terrain and windflow are much more or much less favorable to precipitation) make it hazardous to rely on such global estimation techniques. It should be noted that MIL-STD-210C gives data on climatic extremes (and occasionally percentage exceedance) values for a number of meteorological variables of importance to system design.[41] For comparison with the data in Table 3-14, MIL-STD-210C gives:

* highest 1-min rainfall = 31.2 mm (= 1807.2 mm/hr)
* highest 1-hr rainfall = 305 mm
* highest 12-hr rainfall = 1350 mm (= 112.5 mm/hr)
* highest 24-hr rainfall = 1880 mm (= 78.3 mm/hr)
* highest 5-day rainfall = 4300 mm (= 35.8 mm/hr).

Figure 3-42. CCIR Rain Climate Zones.[31]

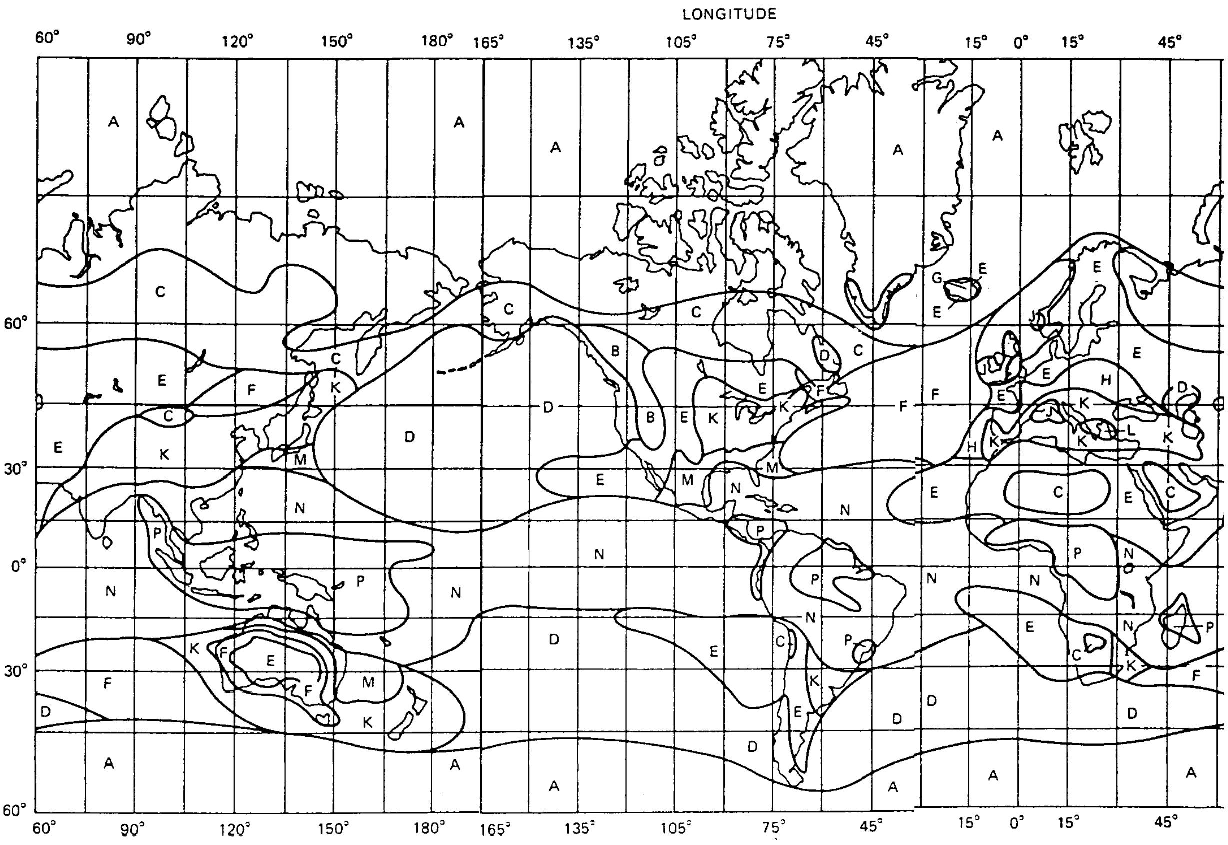


Table 3-14. Rain Rate Distributions for the CCIR Rain Climate Zones. [31]

	RAIN RATE DISTRIBUTION VALUES (mm/hr)						
	PERCENTAGE OF TIME (%) & (TIME PER YEAR) EXCEEDED						
CLIMATE ZONE	1.0 (87.7h)	0.3 (26.3h)	0.1 (8.77h)	0.03 (158m)	0.01 (52.6m)	0.003 (15.8m)	0.001 (5.26m)
A	–	1	2	5	8	14	22
B	1	2	3	6	12	21	32
C	–	3	5	9	15	26	42
D	3	5	8	13	19	29	42
E	1	3	6	12	22	41	70
F	2	4	8	15	28	54	78
G	–	7	12	20	30	45	65
H	–	4	10	18	32	55	83
J	–	13	20	28	35	45	55
K	2	6	12	23	42	70	100
L	–	7	15	33	60	105	150
M	4	11	22	40	63	95	120
N	5	15	35	65	95	140	200
P	12	34	65	105	145	200	250

References

[1] Koschmieder, *Beitrage Phys. freien Atmos.,* 12, 33-53, 171-181 (1924).

[2] Hulburt, E. O., "Optics of Atmospheric Haze", *J. Opt. Soc. Amer.*, 31, 467-476 (1941).

[3] McClatchey, R. A., Fenn, R. W., Selby, J. E. A., Volz, F. E., and Garing, J. S., "Optical Properties of the Atmosphere (3rd Edition)", USAF Cambridge Research Laboratories, Report AFCRL-72-0497 (1972). AD 753 075.

[4] Selby, J. E. A. and McClatchey, R. A., "Atmospheric Transmittance From 0.25 to 28.5 μm: Computer Code LOWTRAN 2", USAF Cambridge Research Laboratories, Report AFCRL-72-0745 (19 December 1972). AD 763 721.

[5] Selby, J. E. A. and McClatchey, R. A., "Atmospheric Transmittance From 0.25 to 28.5 μm: Computer Code LOWTRAN 3", USAF Cambridge Research Laboratories, Report AFCRL-72-0745 (1975). AD A017 734.

[6] Selby, J. E. A., Shettle, E. P., and McClatchey, R. A., "Atmospheric Transmittance From 0.25 to 28.5 μm: Supplement LOWTRAN 3B", USAF Cambridge Research Laboratories, Report AFGL-TR-76-0258 (1976). AD A040 701.

[7] Selby, J. E. A., Kneizys, F. X., Chetwynd, J. H., Jr., and McClatchey, R. A., "Atmospheric Transmittance/Radiance: Computer Code LOWTRAN 4", USAF Geophysics Laboratory, AFGL-TR-78-0053 (1978). AD A058 643.

[8] Kneizys, F. X., Shettle, E. P., Gallery, W. O., Chetwynd, J. H., Abreu, L. W., Selby, J. E. A., Fenn, R. W., and McClatchey, R. A., "Atmospheric Transmittance/Radiance: Computer Code LOWTRAN 5", USAF Geophysics Laboratory, AFGL-TR-80-0067 (1980). AD A088 215.

[9] Kneizys, F. X., Shettle, E. P., Gallery, W. O., Chetwynd, J. H., Abreu, L. W., Selby, J. E. A., Clough, S. A., and Fenn, R. W., "Atmospheric Transmittance/Radiance: Computer Code LOWTRAN 6", USAF Geophysics Laboratory, AFGL-TR-83-0187 (1983). AD A137 796.

[10] Kneizys, F. X., Shettle, E. P., Abreu, L. W., Chetwynd, J. H., Anderson, G. P., Gallery, W. O., Selby, J. E. A., and Clough, S. A.,"User's Guide to LOWTRAN 7", USAF Geophysics Laboratory, AFGL-TR-88-0177 (1988).

[11] Kneizys, F. X., Shettle, E. P., Anderson, G. P., Abreu, L. W., Chetwynd, J. H., Selby, J. E. A., Clough, S. A., and Gallery, W. O., "LOWTRAN 7 Source Code", Revision 3.6, Air Force Geophysics Laboratory (30 Jan 1989). Available from ONTAR Corporation, Brookline MA.

[12] Flight Standards Service, Aviation Weather (Federal Aviation Administration, Washington DC, 1975), pp.9-10, 47-50.

[13] Crane, Robert K. and Blood, David W., "Handbook for the estimation of microwave propagation effects – link calculations for earth-space paths (Path Loss and Noise Estimation)", Environmental Research & Technology, Report P-7376-TR1 (June 1979).

[14] Coordinating Committee on International Radio, "Attenuation by Atmospheric Gases", Report 719, Recommendations and Reports of the CCIR,1978, Vol. 5, Propagation in Non-Ionized Media, (International Telecommunications Union, Geneva, 1978), pp. 97-102.

[15] Crane, R. K., "An Algorithm to Retrieve Water Vapor Information from Satellite Measurements", NEPRF Tech. Report 7-76 (ERT), Final Report No. 1423, Environmental Research & Technology, Inc. (November 1976).

[16] Liebe, H. J., "A Nemesis for Millimeter Wave Propagation" in Atmospheric Water Vapor edited by Deepak, A., Wilkerson, T. D., and Ruhnke, L. H. (Academic Press, New York NY, 1980), p. 168.

[17] Berk, A., Bernstein, L. S., and Robertson, D. C., "MODTRAN: A Moderate Resolution Model for LOWTRAN 7", USAF Geophysical Laboratory, Report AFGL-TR-89-0122 (1989).

[18] ONTAR Corp., "PCMODWin" (1996). Available from ONTAR Corp., Brookline MA.

[19] Rothman, L. S., Gamache, R. R., Goldman, A., Brown, L. R., Toth, R. A., Pickett, H. M., Poynter, R. L., Flaud, J.-M., Camy-Peyret, C., Barbe, A., Husson N., Rinsland, C. P., and Smith, M. A. H., "The HITRAN Database: 1986 Edition", *Applied Optics*, 26, 4058-4097 (1987).

[20] Clough, S. A., Kneizys, F. X., Shettle, E. P., and Anderson, G. P., "Atmospheric Radiance and Transmittance: FASCOD2", Proc. Sixth Conference on Atmospheric Radiation, Williamsburg VA, (American Meteorological Society, Boston MA, 1986), pp. 141-144.

[21] University of South Florida, "USF HITRAN-PC" (1992). Available from ONTAR Corp., Brookline MA.

[22] ONTAR Corp., "PCLnTRAN3P" (1994). Available from ONTAR Corp., Brookline MA.

[23] Duncan, Louis D., et al, "EOSAEL 82", U. S. Army Atmospheric Sciences Laboratory, Report ASL-TR-0122 (1982).

[24] Warner, John and Bichard, V. M., "The Statistics of Atmospheric Extinction at the CO_2 Laser Wavelength Derived from Meteorological Office Records", *Infrared Physics*, 19, 13-18 (1979).

[25] Meteorological Office, Elementary Meteorology 2nd Ed. (Her Majesty's Stationery Office, London UK, 1981).

[26] World Meteorological Organization, International Cloud Atlas Vol. 1, (World Meteorological Organization, Geneva Switzerland, 1975), pp. 16-17.

[27] Thomas, Michael E. and Duncan, Donald D., "Atmospheric Transmission" in Fred G. Smith (Ed.), Atmospheric Propagation of Radiation, Vol. 2 of The Infrared and Electro-Optical Systems Handbook (SPIE Optical Engineering Press, Bellingham WA, 1993).

[28] Liebe, H. J., "An Atmospheric Millimeter Wave Propagation Model", National Telecommunications and Information Administration, NTIA Report 83-137 (December 1983).

[29] ITU Radiocommunication Assembly, "Attenuation Due to Clouds and Fog", Recommendation ITU-R P838-1 (1999).

[30] ITU Radiocommunication Assembly, "Specific Attenuation Model for Rain for use in Prediction Methods", Recommendation ITU-R P838-1 (1999).

[31] Ippolito, Louis J., Jr., Radiowave Propagation in Satellite Communications (Van Nostrand Reinhold, New York NY, 1986).p.46,88.

[32] Department of Defense, "Quantitative Description of Obscuration Factors for Electro-Optical and Millimeter Wave Systems", DoD-HDBK-178 (25 July 1986).

[33] Bisyarin, V. P., Bisyarina, I. P., Rubash, V. K., and Sokolov, A. V., "Attenuation of 10.6 and 0.63 μm Laser Radiation in Atmospheric Precipitation", *Radio Engineering and Electronic Physics*, 16, 1594-1597 (October 1971).

[34] Bogush, A. J., Radar and the Weather, (Artech House, Norwood MA, 1989).

[35] Anonymous, "Weather Effects on EO/IR/MMW Sensors", Wall Chart, Dynetics, Inc. (1990).

[36] Naval Surface Warfare Center, unpublished memorandum (1994).

[37] Modica, A. P., and Kleiman, H., """Statistics of Global IR Transmission", M.I.T. Lincoln Laboratory, Project Report TT-7 (1976).

[38] Rodriguez, Esperanza and Huschke, Ralph E., "The Rand Weather Data Bank (RAWDAB): An Evolving Base of Accessible Weather Data," Rand Corporation Report R-1269-PR (March 1974).

[39] Shettle, Eric P., Fenn, Robert W., Toolin, Robert B., and Turner, Vernon D., "OPAQUE – Data I, From the US/FRG Station, Meppen, Federal Republic of Germany, Winter 1976/77", Air Force Geophysics Laboratory, Report AFGL-TR-79-0065 (16 July 1979).

[40] Crane, Robert K., Electromagnetic Wave Propagation Through Rain (John Wiley & Sons, New York NY, 1996).

[41] Department of Defense, "Climatic Information to Determine Design and Test Requirements for Military Systems and Equipment", MIL-STD-210C (9 January 1987).

Problems

3-1. Identify the major potential contributors to atmospheric extinction and qualitatively describe the altitude, wavelength, and weather regimes in which each is a major factor.

3-2. What is an "atmospheric window"? Name or describe by frequency or wavelength at least five major atmospheric windows.

3-3. Describe the relation between extinction, absorption, and scattering.

3-4. Smoke from a nearby fire reduces the visibility in an area to 500 meters. Without any other information what would you estimate the scattering coefficient at 550 nm wavelength to be?

3-5. A chemical pollutant causes only absorption (no scattering) at 10 μm wavelength. The total extinction at this wavelength is estimated to be 1.0 km^{-1}. Normal atmospheric constituents are expected to contribute to produce a background absorption coefficient of 0.5 km^{-1} and a scattering coefficient of 0.3 km^{-1}. What is the absorption coefficient of the pollutant? If the atmospheric extinction contributions increased two-fold and the pollutant concentration increased ten-fold, what total extinction coefficient would result?

3-6. What meteorological parameter defines the vertical structure of the atmosphere?

3-7. Sketch the behavior of temperature, pressure, humidity, & aerosol density vs. altitude. Use the same vertical scale for each sketch.

3-8. A temperature gradient of 8 degrees per 1000 meters exists in an air mass. Is this air mass considered stable or unstable?

3-9. What atmospheric constituents are almost entirely absent at altitudes above 15000 meters? What atmospheric constituent is stronger at 15000 meters than at sea level? Do not consider clouds.

3-10. Given an atmosphere containing O_2, N_2, CO_2, H_2O, and O_3:
a. Which species produce ultraviolet absorption between 200 and 300 nm?
b. Which species have strong absorption lines between 10 and 100 GHz?
c. Which species have strong absorption bands between 1 and 10 μm?

3-11. If the ambient temperature is 305 K, the absolute humidity is 30 g/m^3, and the rain rate is 25 mm/hr, estimate the sea level microwave absorption coefficient at 15 GHz.

3-12. For this problem assume that the "signal-to-noise ratio" of a radar scales as $SNR \propto R^{-4} e^{-2\alpha R}$. At a particular sea-level radar test range the average absolute humidity is 7.5 g-m^{-3}. At this range one model of a 17 GHz surface search radar has a detection range of 20km against a standard target. Ignoring all other possible phenomena that can affect range, plot the detection range against the standard target of this radar as a function of absolute humidity ranging from 0 g-m^{-3} (Arctic conditions) to 30 g-m^{-3} (tropical conditions). Six to ten well-chosen data points are sufficient to be plotted.

3-13. If the dewpoint is 293.15 K, estimate the saturation humidity.

3-14. For this problem assume that the "signal-to-noise ratio" of a radar scales as $SNR \propto R^{-4} e^{-2\alpha R}$. At the same test range and for the same standard target, two different radars give the same detection range of 50 km. One radar operates at 1 GHz; the second radar operates at 25 GHz (near a peak absorption of water vapor). A third radar operates at 60 GHz (at the peak of an oxygen absorption) and has a detection range of 1 km. Ignoring all other phenomena that can affect range, plot the variation of detection range with altitude over the interval from 0 km altitude to 30 km altitude. Six to ten well-chosen data points are sufficient to be plotted.

3-15. If the aerosol concentrations of four model atmospheres (that vary only in aerosol type and concentration) are adjusted such that the extinction coefficients at 10 μm of the four major LOWTRAN aerosol models (rural, urban, maritime, and tropospheric) are equal, which model will have higher extinction at 1.0 μm?

3-16. If the visibility is 10 km, estimate the absorption coefficient of a rural aerosol at 5.0 μm.

3-17. A millimeter-wave radar (at 98 GHz) is proposed for an anti-aircraft seeker. At maximum range (10 km) the seeker has 10 dB margin for weather penetration. Will clouds cause this seeker to fail?

3-18. Water vapor causes a 5 dB/km extinction at a wavelength in the near infrared. What rain rate will produce an extinction of equal magnitude?

3.19. Using the data in Figures 3-34 through 3-37 estimate the rain rate that should be used to calculate extinction if greater than 95% probability of operation is desired in any geographic locale. Estimate the visibility that should be assumed for the equivalent aerosol extinction calculations.

3-20. You are responsible for developing a new low-altitude (< 100 m max. alt.), subsonic, land attack cruise missile. The Program Executive Officer (PEO) received his training in the 70's. He demands that use an "all-weather" altimeter (implying an existing 17 GHz radar altimeter). You know that this device can be detected, localized, identified, and tracked by a new generation of threat-owned, netted, radio-frequency interferometers. As a result most of the missiles can be easily downed. In addition to tactfully explaining the above, what can you say to the PEO about the existence of alternatives and the true meaning of "all weather"?

3-21. You are responsible for picking the sensors for an "air vehicle" to be used during the first manned mission to Mars. Assume that the Martian atmosphere is almost completely composed of nitrogen and carbon dioxide with no oxygen, ozone, or water vapor, and that the average surface temperature is 210K.

(1) Which of the following sensors are likely to experience a significant problem from atmospheric attenuation (ignoring the effects of Martian aerosols):

a) a 60 GHz terrain mapping radar?

b) a high-resolution thermal imager operating in the 4.0-4.5 μm wavelength band and designed for surveying land travel routes?

c) a general-purpose thermal imager operating in the 8-12 μm band?

d) an image intensified optical viewer operating in the 0.9-1.0 μm band?

e) a "solar-blind" ultraviolet sensor operating in the 0.21-0.29 μm band and designed to detect the low-intensity ultraviolet beacons on previous unmanned landers?

(2) Which of the above sensors will be significantly affected by the Martian dust storms (assume that all dust particles have diameters less than 1 μm)? Which of these sensors will be marginally affected?

(3) Do any of the above sensors have serious problems other than molecular attenuation or aerosol attenuation? Which sensors have what problems? Consider things like signature limitations or noise sources.

(4) Which of these sensors will be difficult to flight test here on Earth? Why? (be specific)

3-22. A laser communication link is required for ship-to-ship covert communication over distances less than 20 km. Two strong candidates present themselves: a GaAs diode laser link operating at 0.89 μm and a CO_2 laser link operating at 10.59 μm. Considering only the effects of the atmosphere on system performance, compare the two links. List all atmospheric propagation phenomena which are likely to adversely affect the two systems and rate each phenomenon as to whether it: 1) negatively affects the GaAs link much more than it does the CO_2 link, 2) negatively affects the GaAs link somewhat more than the CO_2 link, 3) affects both links equally, 4) negatively affects the CO_2 link somewhat more than the GaAs link, or 5) negatively affects the CO_2 link much more than it does the GaAs link. Present your results in the form of a table. If the relative impact of any phenomenon can be quantitatively evaluated without use of computer codes, give a numerical number for that relative impact. If any phenomenon affects either the GaAs or CO_2 links, but does not significantly affect the other, please indicate these unilateral impacts. NOTE: multipath is not a problem for these two systems.

3-23. A rangefinder is required to provide accurate tracking of ballistic missiles at ranges up to 500 km for an airborne high energy laser defense system. Both the airborne laser and the targets will be at altitudes above 10 km. Serious candidates exist incorporating YAG lasers at 1.06 μm and CO_2 lasers at 10.59 μm. Perform a comparative analysis of these two candidates considering only atmospheric propagation effects.

CHAPTER 4

PROPAGATION OF ELECTROMAGNETIC RADIATION. III. – REFRACTIVE EFFECTS

Horizon Extension

The atmosphere has a refractive index that is close to unity, but nevertheless not equal to unity. This refractive index is a function of temperature, pressure, and humidity, as well as wavelength. As a consequence, there is considerable spatial and temporal inhomogeneity in the atmospheric refractive index. This innate variability allows the atmospheric refractive index to produce a number of interesting and significant effects on the propagation of electromagnetic radiation.

One characteristic of the atmospheric refractive index is its relatively predictable variation with altitude. In the visible and infrared regions of the spectrum, the refractive index of air is given with reasonable accuracy by an expression due to Edlen [1]. As mentioned in the preceding chapter, the refractive index *n* of air is more commonly expressed in terms of the refractivity *N*

$$N = (n-1) \times 10^6 \tag{4.1}$$

According to Edlen, the refractivity of air may be determined from the expression

$$\begin{aligned} N = &\left[83.43 + \frac{185.08}{1-(0.1603/\lambda)^2} + \frac{4.11}{1-(0.0877/\lambda)^2} \right] \\ &\times \left[\frac{(p_A - p_{H_2O})}{1013.25} \frac{296.15}{T} \right] \\ &+ \left[43.49 - (0.0588/\lambda)^2 \right] \frac{p_{H_2O}}{1013.25} \end{aligned} \tag{4.2}$$

where λ = wavelength, P_A = total pressure (mbar), P_{H2O} = partial pressure of water vapor (mbar), and T = temperature (K). The partial pressure of water vapor may be calculated from the relative humidity and the temperature using the expression

$$p_{H_2O} = \frac{RH}{100} \times 6.105 \exp\left[25.22 \frac{T-273.15}{T} - 5.31 \ln\left(\frac{T}{273.15}\right)\right]. \tag{4.3}$$

In the microwave spectral region, the refractivity may be determined from the expression [2]

$$N = \frac{77.6 p_A}{T} + \frac{3.73 \times 10^5 p_{H_2O}}{T^2} \tag{4.4}$$

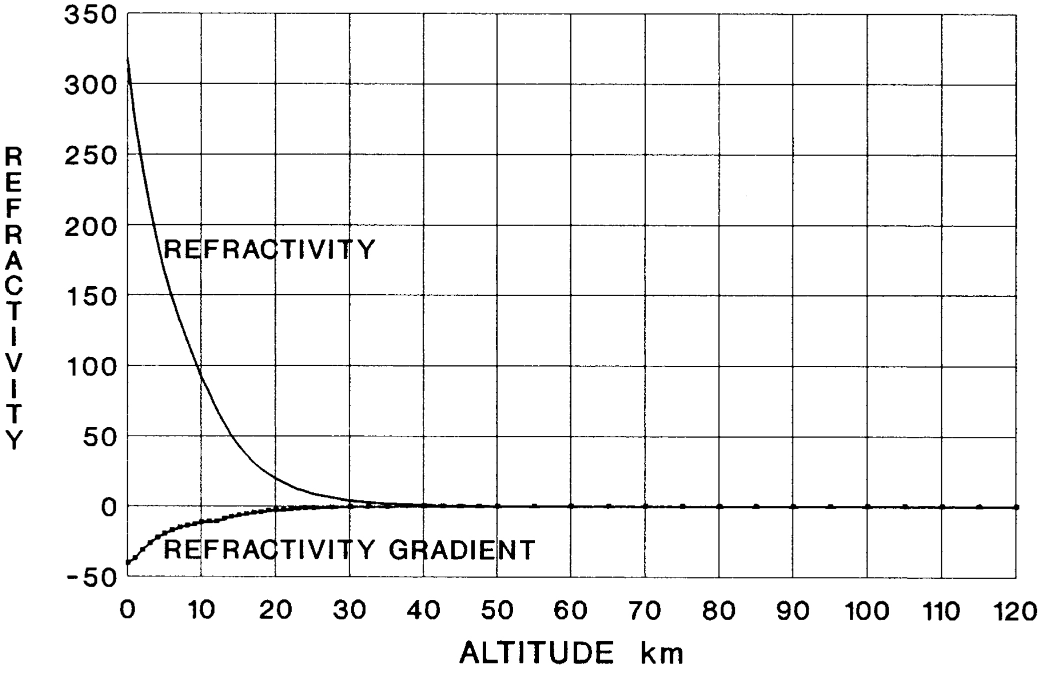
Microwave propagation specialists often define a quantity called the modified refractivity M

$$M = N + 0.157 h \tag{4.5}$$

where h is the height above the ground in meters.

Figure 4-1 shows the microwave refractivity as a function of altitude for the U. S. Standard atmosphere along with the gradient (per km) of that refractivity. Note that for this "standard" condition, there is strong negative gradient This also means that as the refractivity decreases with increasing altitude, the speed of light increases with increasing altitude. Remembering Snell's law

Figure 4-1. Refractivity and its gradient (per km) of the U. S. Standard atmosphere.



(Figure 4-2),

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 = \text{constant for a given ray path}\,, \qquad (4.6)$$

we would expect upward traveling rays to bend away from the vertical as they propagated. To see this, consider an atmosphere with a vertical gradient that has been divided into layers of a finite thickness δ. Each successively higher layer has a slightly larger speed of light (lower index of refraction). This is essentially a finite element approximation to the real atmosphere. If δ were to become infinitesimally small and the number of layers were to become infinite, the model would approach reality (which is continuous). Part of such a model is shown is shown in Figure 4-3. As a ray moves upward, propagating from one layer to the next, it undergoes successive Snell's law refractions that cause it to bend more and more away from vertical propagation. Such bending will occur in a continuous atmosphere with a gradient, although the bending will take the form of a continuous curvature rather than discrete bends.

Figure 4-2. Snell's law geometry.

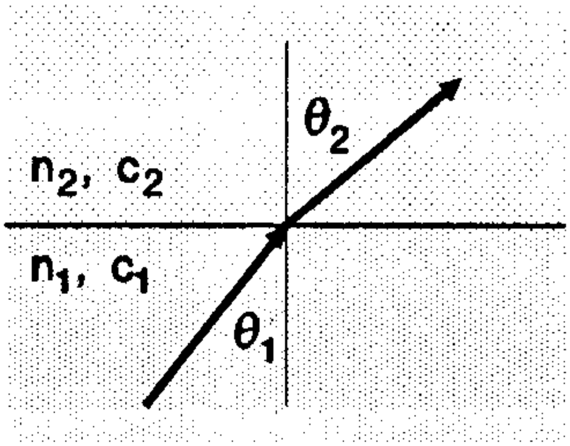
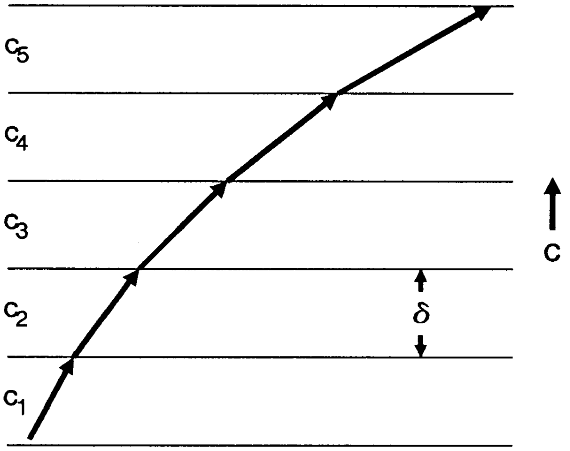


Figure 4-3. Ray curvature in an atmosphere with a negative refractivity gradient.



The radius of curvature is related to the gradient.[2],[3] That this is so can be easily demonstrated geometrically. Consider the stratified geometry of Figure 4-4. For a small enough angle $\delta\alpha$, the radius of curvature r can be determined from

$$x = r\,\delta\alpha \qquad (4.7)$$

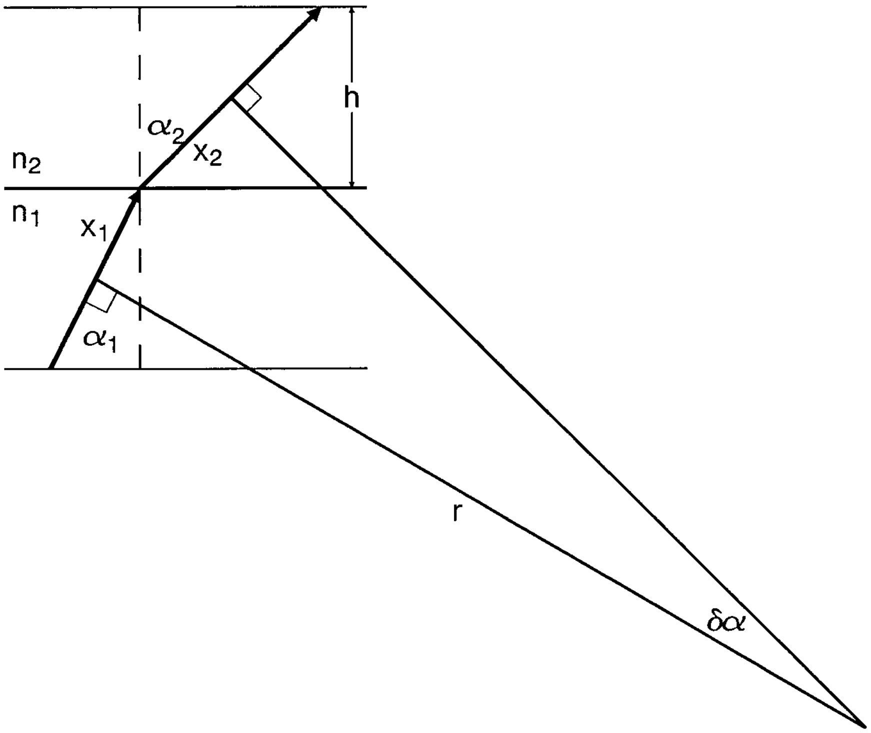
where

$$\begin{aligned} x &= x_1 + x_2 \\ &= \delta h / 2\cos\alpha_1 + \delta h / 2\cos\alpha_2 \\ &\approx \delta h / \cos\alpha \qquad (4.8) \end{aligned}$$

Combining these two results yields the following relation for the radius of curvature

$$r = \delta h / \left(\delta\alpha \, \cos\alpha\right) \quad . \qquad (4.9)$$

Figure 4-4. Radius of curvature resulting from a refractive index gradient.



If we now differentiate Snell's Law we obtain

$$\delta n \sin\alpha + n\cos\alpha \; \delta\alpha = 0 \tag{4.10}$$

or

$$\delta\alpha \,/\, \delta n = -(1/n)\tan\alpha \; . \tag{4.11}$$

Combining Eqs. (4.9) and (4.11) yields

$$r = -n \; \delta h \,/\, \delta n \; \sin\alpha \; . \tag{4.12}$$

For nearly horizontal propagation in air, $n \sim 1$, and $\alpha \sim 90°$, and

$$r = -\delta h \,/\, \delta n = -1 \bigg/ \frac{dn}{dh} \, . \tag{4.13}$$

A positive radius means the center of curvature lies below the ray. The "curvature" ρ of the ray is the inverse of the radius of curvature.

$$\rho = \frac{d\alpha}{dx} = \frac{1}{r} = -\frac{dn}{dh} \tag{4.14}$$

If the propagation path now lies above a curved earth (curvature = $1/R_e$, where R_e is the radius of the Earth) the curvature of the ray relative to the Earth's surface is

$$\rho' = \frac{1}{R_e} + \frac{dn}{dh} \tag{4.15}$$

with an effective radius R

$$\frac{1}{R} = \frac{1}{R_e} + \frac{dn}{dh}. \tag{4.16}$$

Multiplying by the true Earth radius gives

$$\frac{R_e}{R} = 1 + R_e\frac{dn}{dh} = \frac{1}{k} \tag{4.17}$$

where k is the ratio of the effective radius to the true radius. The curvature of the ray relative to the surface of an earth with effective radius $R = kR_e$ is zero. The ray will act as if it follows a straight line path in this effective earth. A nominal average value of $dn/dh = -3.9 \times 10^{-8}$ m^{-1} in the microwave region of the spectrum. [2] Given a true Earth radius of approximately 6.37×10^6 m, we find that a typical value of k is roughly 4/3.

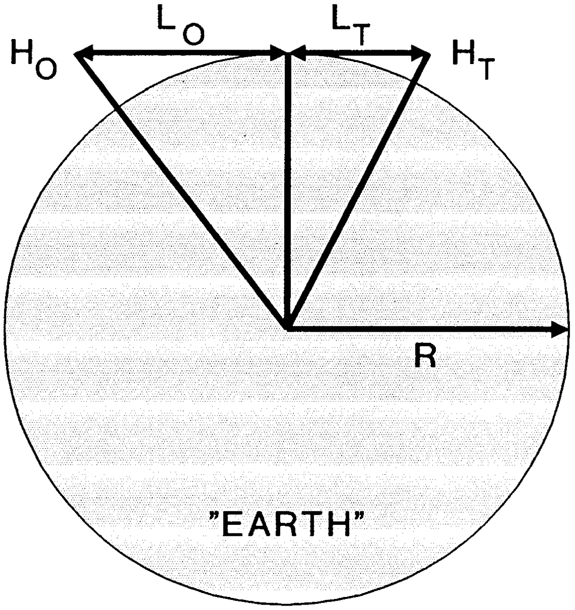
Because of this, "normal" refraction will affect the distance at which objects near the horizon may be detected. Assuming an effective earth radius allows refraction to be neglected and any ray to be treated as a straight line. Consider a spherical earth of arbitrary radius R, as shown in Figure 4-5. The distance L to the horizon (tangent point) for an object at height H above the surface can be calculated from the expression

$$\left(R + H\right)^2 = R^2 + L^2 \tag{4.18}$$

For any H small compared to the radius R, it is possible to use the approximate result

$$L = \left(H^2 + 2HR\right)^{1/2} \approx \left(2HR\right)^{1/2} \ . \tag{4.19}$$

Figure 4-5. Geometry of determining the distance to the horizon.



The true average radius of the Earth is

$$R_e = 6371.0 \text{ km} = 3440.065 \text{ nmi}. \tag{4.20}$$

Thus, the geometric distance to the horizon [$R = R_e$] on the Earth is

$$L = 3.570\ H^{1/2} \quad \text{km} \qquad (H \text{ in meters}) \tag{4.21}$$

$$L = 1.064\ H^{1/2} \quad \text{nmi} \qquad (H \text{ in feet}). \tag{4.22}$$

In the microwave region of the electromagnetic spectrum we may assume that the effective radius of the Earth in the microwave region is 4/3 times the true Earth radius, as we previously calculated using Eq. (4.17). This validates our use of the straight line equation (4.19). Substituting $R = (4/3)R_e$ into Eq. (4.19) leads to the following simple expressions for the radar horizon

$$L = 4.122\ H^{1/2} \quad \text{km} \qquad (H \text{ in meters}) \tag{4.23}$$

$$L = 1.229\ H^{1/2} \quad \text{nmi} \qquad (H \text{ in feet}). \tag{4.24}$$

The increase in radar horizon given by these expressions is reasonable under normal conditions. If superrefractive or ducting conditions exist (see below), then the effective horizon may exceed these estimates.

In the visual and infrared regions of the spectrum, there is also refraction. In general this refraction is not as strong as that in the microwave region. The visual horizon may be estimated using a 7/6 Earth radius.[4] Substituting $R = (7/6)R_e$ into Eq. (4.19) leads to the following simple expressions for the visual horizon

$$L = 3.856\ H^{1/2} \quad \text{km} \qquad (H \text{ in meters}) \qquad (4.25)$$

$$L = 1.149\ H^{1/2} \quad \text{nmi} \qquad (H \text{ in feet}). \qquad (4.26)$$

The easy-to-remember fraction 7/6 (=1.1667) is the nearest simple fraction to the true mean value of 1.1662. Slightly better horizon distances could be found from the equations

$$L = 3.839\ H^{1/2} \quad \text{km} \qquad (H \text{ in meters}) \qquad (4.25\text{A})$$

$$L = 1.144\ H^{1/2} \quad \text{nmi} \qquad (H \text{ in feet}). \qquad (4.26\text{A})$$

if desired. The alternate set of expressions (Eq. (4.25A) & (4.26A)) are more accurate and preferred.

The total distance to an object that appears right on the horizon is the sum of the distance from the object to the horizon and the distance from the horizon to the observer

$$L_{TOTAL} = L_T + L_O = A\left(H_T^{1/2} + H_O^{1/2}\right), \qquad (4.27)$$

where A is one of the numerical coefficients from the equations above.

It is interesting to note that if we select ρ' equal to zero, then the ray curvature relative to the Earth's surface is zero, a ray initially traveling parallel to the surface remains parallel even as it travels completely around the world. This condition is

$$\frac{dn}{dh} = -\frac{1}{R_e} = -1.57 \times 10^{-7}\ \text{m}^{-1} \qquad (4.28)$$

or in terms of refractivity

$$\frac{dN}{dh} = -0.157\ \text{m}^{-1}. \qquad (4.29)$$

This is the origin of the $0.157h$ term in the conversion between N and M (in Eq. (4.5)). If the refractivity gradient is more negative than this then every ray (in an infinitely high atmosphere) has enough curvature to ultimately bend back and intercept the earth. The rays may be said to be "trapped".

Ducting and Mirages

As explicitly demonstrated in Eq. (4.4), the refractivity is a function of pressure, temperature, and water vapor concentration. Except in severe storms with strong vertical winds, total pressure almost always falls off smoothly with increasing altitude. However, temperature anomalies (variations in lapse rate, possibly including inversions) and vertical variations in humidity (variations in gradient due to excessively low or high surface humidity or even layers of abnormally low or high humidity) can exist in the atmosphere. For these reasons, the actual bending of ray paths need not follow the simple relations presented above (and corresponding to "**normal refraction**").[5]

Consider Figure 4-6. If the refractive gradient is less negative (or even positive) then the curvature is less than "normal" (radius of curvature is larger than normal) or curved in the opposite direction (away from the surface). This atmospheric condition is called "**subrefractive**". If the refractive gradient is more negative than normal then the curvature is larger than normal (radius of curvature is smaller than normal). This atmospheric condition is called "**superrefractive**". In extreme conditions, the curvature may become so large that the rays will bend more quickly than the earth curvature. Even rays initially directed away from the surface will curve until they strike the earth's surface and reflect. This is called a "**trapping**" condition, because the rays cannot escape.

The refractivity gradients that correspond to these conditions are described in Table 4-1. Trapping occurs when the gradient in the modified refractivity becomes negative. "Normal" refraction is assumed to occur whenever the modified refractivity is between 79 and 157 km^{-1}.

Figure 4-6. Four modes of propagation in a refractive atmosphere.

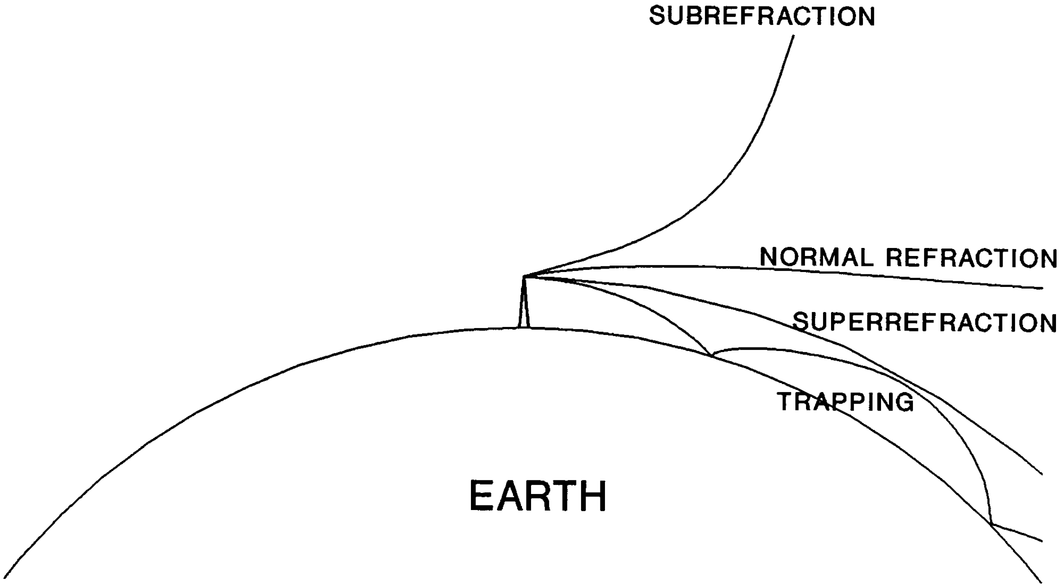


Table 4-1. Relation of refractivity gradient to refractive behavior.

BEHAVIOR	N-GRADIENT (N/km)	M-GRADIENT (M/km)
TRAPPING	< -157	< 0
SUPERREFRACTIVE	- 157 to -79	0 to 79
NORMAL	- 79 to 0	79 to 157
SUBREFRACTIVE	> 0	> 157

Trapping is such an anomalous behavior, when it occurs, that it must have profound impacts on sensor performance. For this reason it will be examined in more detail. The first question is what conditions lead to a negative M-gradient? We will begin in the microwave region of the electromagnetic spectrum. Combining Eqs. (4.4) and (4.5) yields the expression

$$M = \frac{77.6 p_A}{T} + \frac{3.73 \times 10^5 p_{H_2O}}{T^2} + 0.157h \tag{4.30}$$

The only factors in this expression which are capable of very anomalous behavior are the water vapor density and the temperature. A negative gradient requires either a very rapid decrease in water vapor concentration or a very rapid increase in temperature. As we have seen in Figure 4-1 (add 157 to the N-gradient shown to get the M-gradient), neither the rapid falloff of water vapor in the normal troposphere nor the temperature inversion at the tropopause are adequate to make the M-gradient negative. First, consider dry air at sea level. In the absence of water vapor the refractivity becomes

$$M\Big|_{p_A = 1013\text{ mbar};\ p_{H_2O}=0} = \frac{78609}{T} + 0.157h \tag{4.31}$$

Taking the derivative with respect to height gives the gradient

$$\left.\frac{dM}{dh}\right|_{p_A = 1013\text{ mbar};\ p_{H_2O}=0} = 0.157 - \frac{78609}{T^2}\frac{dT}{dh} \tag{4.32}$$

or

$$\left.\frac{dM}{dh}\right|_{p_A = 1013\text{ mbar};\ p_{H_2O}=0;\ T=273.15\text{ K}} = 0.157 - 1.054\frac{dT}{dh} \tag{4.33}$$

where the last expression has been explicitly evaluated at a temperature of 273.15 K. The M-gradient becomes negative if the temperature gradient exceeds

$$\left.\frac{dT}{dh}\right|_{p_A=1013\text{ mbar};\ p_{H_2O}=0;\ T=273.15\text{ K}} > 0.149\ \ \text{K/m} \qquad (4.34)$$

This is a substantial temperature gradient but one which could easily occur in a situation where warm air moved over a cold surface or cold air mass.

Now consider moist air. If we once again evaluate the modified refractivity expression at a temperature of 273.15 K, we find

$$\left.M\right|_{p_A=1013\text{ mbar};\ T=273.15\text{ K}} = 287.79 + 5.000 p_{H_2O} + 0.157h \qquad (4.35)$$

The M-gradient becomes

$$\frac{dM}{dh} = 0.157 + 5.000\frac{dp_{H_2O}}{dh} \qquad (4.36)$$

which in turn becomes negative if the water vapor pressure gradient is sufficiently negative, i.e.,

$$\frac{dp_{H_2O}}{dh} < -0.0314\ \ \text{mbar / m}. \qquad (4.37)$$

At a temperature of 273.15 K, the water vapor pressure can be calculated from the relative humidity from the relation

$$\left.p_{H_2O}\right|_{T=273.15\text{ K}} = \frac{RH}{100}\times 6.105 \quad \text{mbar} \qquad (4.38)$$

Given partial pressures of several mbar, a gradient of a few hundredths of a mbar/meter seems relatively easy to achieve.

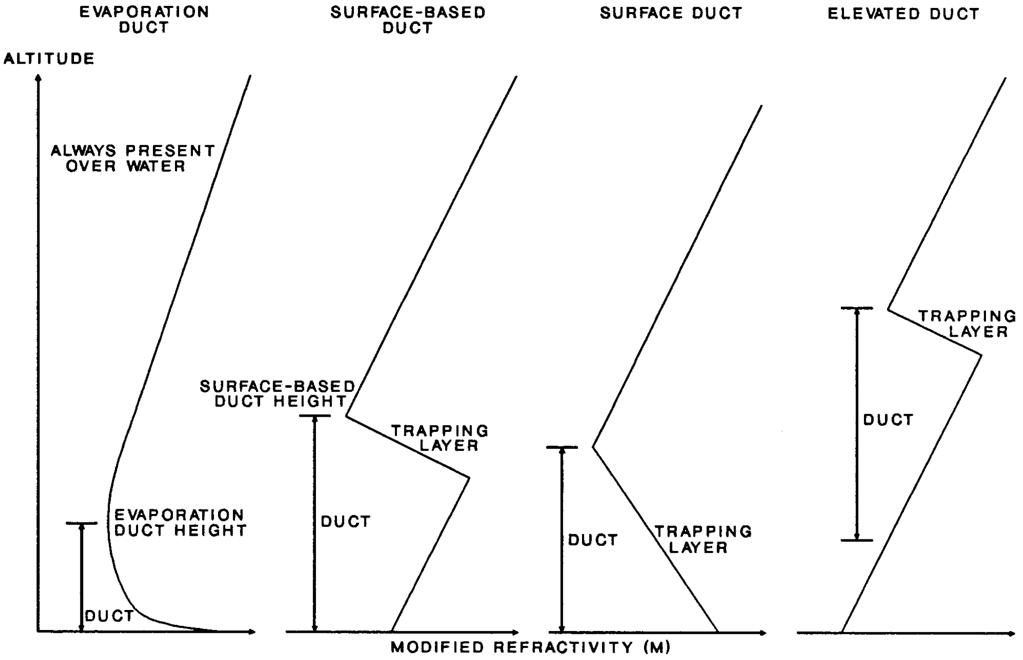
Indeed we find such gradients over virtually every body of water. The ocean is the primary source of water in the overall hydrologic cycle (ocean water evaporates; ocean air moves over land causing precipitation; water returns to ocean through streams and rivers). The air in intimate contact with the surface of the water will be fully saturated with water vapor at the water temperature. As the saturated air moves away from the surface, it will mix with drier air causing a negative gradient. The gradient will diminish with increasing distance until at a height of a few meters to a few tens of meter the negative gradient disappears and more normal refractivity behavior resumes. The entire height region from the ground to the point at which a positive M-gradient resumes is called an **evaporation duct (EVD)**. The height profile of a typical evaporation duct is

$$M(h) = M_{SURFACE} + \frac{h}{8} - \frac{\delta}{8}\ln\left(\frac{h+0.00015}{0.00015}\right) \quad . \tag{4.39}$$

where δ is the height of the top of the duct. This is shown schematically in Figure 4-7.

Occasionally conditions arise where layers of moist air are forced to move over a layer of less moist air. These can lead to the three other kinds of ducts as shown in Figure 4-7. If the trapping layer is more extended than an evaporation duct, but touches the surface, we have a **surface duct (SD)**. The trapping layer is that range of heights where the M-gradient is negative (i.e., the refractivity indicates trapping behavior). If the trapping layer does not extend to the surface, but the minimum modified refractivity in the layer is smaller than the modified refractivity at the surface, we have a **surface-based duct (SBD)**. Although the trapping layer is elevated, the duct it produces (see below) extends to the ground. If the trapping layer is sufficiently elevated, the minimum modified refractivity will not be smaller than the value at the surface. In this case we have an **elevated duct (ED)**. The duct will extend from the top of the trapping layer down to the height at which modified refractivity below the trapping layer equals the minimum modified refractivity at the top of the trapping layer. The position and thickness of the duct regions and the trapping layers are indicated for all four kinds of ducts. As mentioned earlier, evaporation ducts are inevitable over water. The other kinds of ducts are known to occur occasionally if not frequently over land as well as over water, but data is limited.

Figure 4-7. Types of ducts resulting from trapping layers in the microwave.

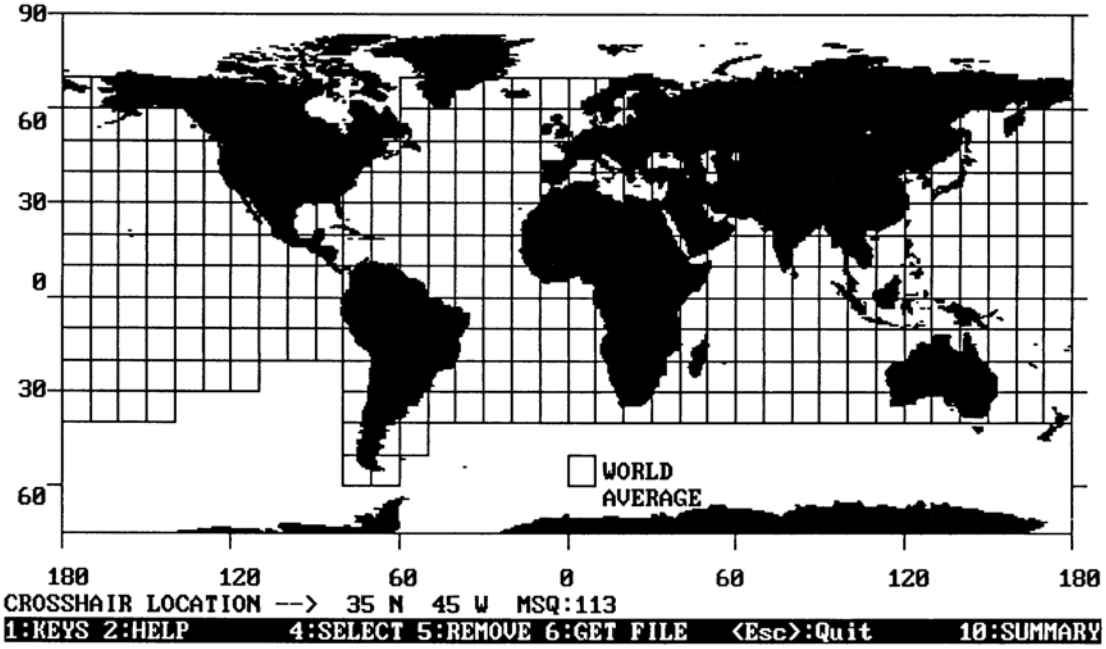


Statistics on some ducts in maritime environments has been summarized in a portion of a computer code called Engineer's Refractive Effects Prediction System (EREPS).[5] A component of this code called SDS produces statistical outputs on demand. Some of the available data is summarized in Table 4-2. Using the SDS/EREPS data we have calculated the evaporation duct heights at the 10 percentile, 50 percentile, and 90 percentile cumulative probability levels for 26 maritime locations around the world and for the world as a whole. We have also listed the average value of k, the refractivity at the surface $N_{SURFACE}$, the probability of occurrence of surface-based ducts (SBD), and the average SBD duct height for each location. Data is available for most of the ocean surface averaged over each Marsden square. Each Marsden square is a 10 degree x 10 degree region of the ocean surface. Figure 4-8 shows the Marsden squares for which duct data is available. Figure 4-9 shows an example of the output. Data are not available for the Arctic and Antarctic regions, nor are they available over land (however, ducts over the water tend to extend over adjacent land areas although may elevate as they move farther from the coasts). Data are not available for surface ducts or elevated ducts.

Table 4-2. Maritime duct statistics obtained from the SDS portion of the EREPS code.

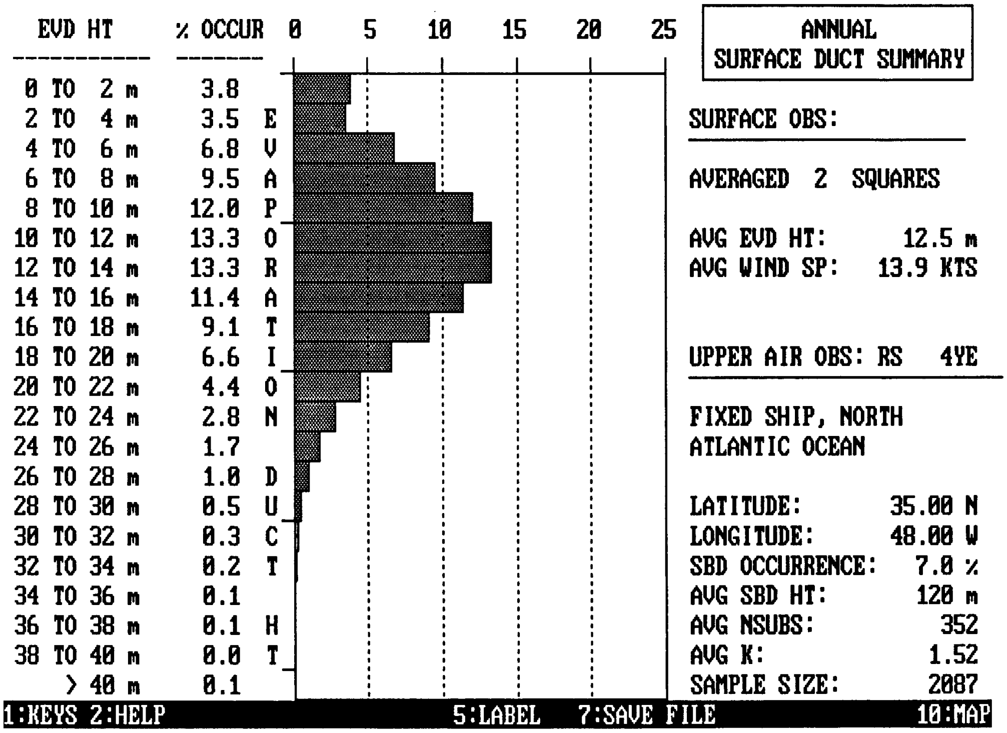
	EVD HEIGHT (m)						SBD
GEOGRAPHIC	PERCENTILES			AVERAGE		SBD	AV. HT.
LOCATION	10	50	90	K	$N_{SURFACE}$	(%)	(m)
Worldwide Average	4	13.1	22	1.45	339	8.0	85
California Coast	2	8.9	16	1.62	328	12.0	117
Gulf of Alaska	1	5.3	10	1.38	321	2.0	48
Hawaiian Islands	9	16.8	24	1.51	362	20.0	114
Rapa, French Oceania	3	10.6	18	1.44	355	2.0	154
Guam	10	17.7	25	1.71	380	10.0	107
Coral Sea	8	16.2	25	1.48	353	10.0	90
Tasman Sea	5	13.5	22	1.41	329	21.3	78
Yellow Sea	3	11.9	21	1.41	322	14.3	83
Paracel Islands	8	16.3	25	1.59	374	14.0	79
North Australian Basin	8	17.2	26	1.69	342	19.0	147
Gulf of Thailand	7	14.9	22	1.68	381	14.4	83
Bay of Bengal	8	16.0	24	1.77	378	17.5	109
Diego Garcia	8	15.0	22	1.60	382	22.0	80
Persian Gulf	6	14.7	25	2.08	357	45.5	202
Central Mediterranean	5	11.9	20	1.54	337	21.8	122
Cape of Good Hope	5	13.1	22	1.41	334	2.0	90
Ascension Island	10	16.9	24	1.37	322	1.7	69
Mauritanian Coast	8	15.2	22	1.78	345	37.0	148
North Sea	1	6.4	12	1.37	322	1.7	69
Greenland-Iceland Gap	2	5.9	10	1.34	316	3.5	45
N. Atlantic (44N,41W)	2	9.7	17	1.41	336	5.0	68
Grand Banks	1	6.6	15	1.34	315	1.7	68
Bermuda	6	14.7	24	1.48	347	2.0	200
Bahamas	8	16.4	25	1.59	370	1.0	114
Caribbean Sea	10	17.6	25	1.60	371	19.0	122
Cape Horn	2	6.4	14	1.31	312	3.5	48

Figure 4-8. Availability of duct data by Marsden squares.



```
CROSSHAIR LOCATION --> 35 N  45 W  MSQ:113
```

Figure 4-9. Example of SDS/EREPS duct data output.



```
 EVD HT        % OCCUR          ANNUAL
-----------    -------     SURFACE DUCT SUMMARY
 0 TO  2 m      3.8
 2 TO  4 m      3.5  E     SURFACE OBS:
 4 TO  6 m      6.8  V
 6 TO  8 m      9.5  A     AVERAGED  2  SQUARES
 8 TO 10 m     12.0  P
10 TO 12 m     13.3  O     AVG EVD HT:      12.5 m
12 TO 14 m     13.3  R     AVG WIND SP:   13.9 KTS
14 TO 16 m     11.4  A
16 TO 18 m      9.1  T
18 TO 20 m      6.6  I     UPPER AIR OBS: RS   4YE
20 TO 22 m      4.4  O
22 TO 24 m      2.8  N     FIXED SHIP, NORTH
24 TO 26 m      1.7        ATLANTIC OCEAN
26 TO 28 m      1.0  D
28 TO 30 m      0.5  U     LATITUDE:         35.00 N
30 TO 32 m      0.3  C     LONGITUDE:        48.00 W
32 TO 34 m      0.2  T     SBD OCCURRENCE:   7.0 %
34 TO 36 m      0.1        AVG SBD HT:       120 m
36 TO 38 m      0.1  H     AVG NSUBS:          352
38 TO 40 m      0.0  T     AVG K:             1.52
    > 40 m      0.1        SAMPLE SIZE:       2087
```

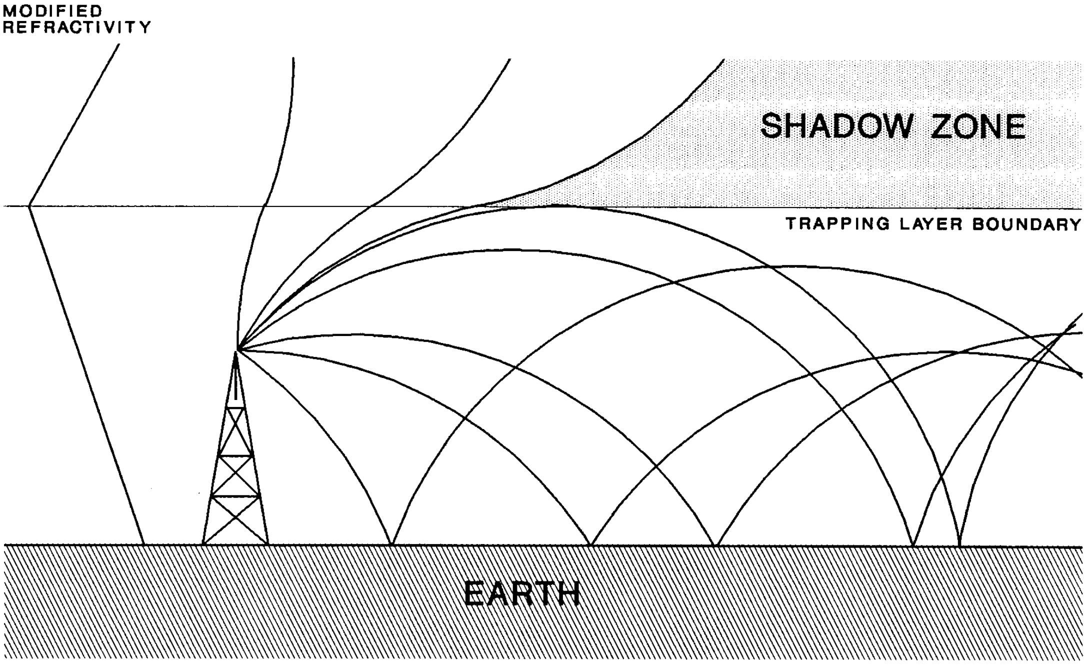
It is worthwhile to summarize the data in Table 4-2. Evaporation ducts occur all of the time over large bodies of water. Duct heights can range from a low as one meter to as high as 40 meters but 5 to 25 meters is most probable with 13 meters as an average. Yearly average values of *N* at the water surface can range from roughly 310 ppm to over 380 ppm depending on location. Surface-based ducts occur only 8% of the time world wide, but probabilities of occurrence can be as low as 1% and as high as 45% depending on location. Hot & humid locations tend to have high probabilities of SBD occurrence, while cold regions tend to have low probabilities of SBD occurrence. SBD heights average 85 meters worldwide but can range from 50 meters to over 200 meters depending on location.

We have described the anomalous propagation regions resulting from trapping layers as ducts although we have not yet justified this description. Consider a trapping layer that extends to the surface and assume that the refractivity gradient is a constant negative value, as shown in Figure 4-10. We will use a flat earth illustration, but curved earth equations in the computations. For a constant negative gradient, the ray curvature is constant, curved downwards toward the surface, and given by

$$\rho' = \frac{1}{R_e} + \frac{dn}{dh} \tag{4.40}$$

relative to the Earth's surface. For many rays this curvature is sufficient to cause them to strike the

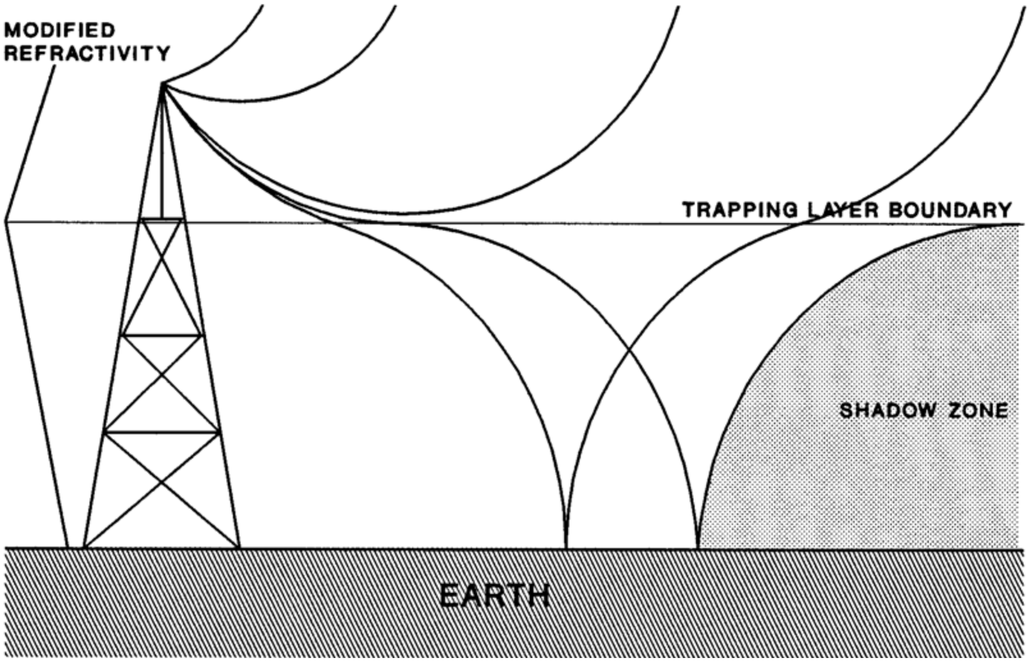
Figure 4-10. Trapping in a negative linear gradient extending up from the surface.



surface, reflect from that surface, and repeat the process of curving back towards the surface until it strikes the surface again (and again ...). This distance traveled before striking the surface depends on the transmitter height, strength of the gradient, and the initial propagation direction. If the height of the trapping layer minus the transmitter height is less than the radius of curvature of the rays, then some of the rays will reach the top of the trapping layer. When this happens, the curvature abruptly changes and the ray diverges from the surface and escapes to space. Unless the relative curvature above the trapping layer is zero, there will be a wedge-shaped region in which it is not possible to propagate any ray originating from the transmitter. This "**shadow zone**" is indicated in the figure. The existence of a shadow zone implies a region in which a target can exist in which there is no illumination by the transmitter. In a radar system, no illumination implies no detection. The confinement of radiation in the "duct" will often lead to lower propagation losses than expected from a $1/R^2$ spreading.

A different situation exists if the transmitter (or receiver) is above the surface duct (see Figure 4-11). Radiation approaching the top of the duct at too shallow an angle will be refracted away before it reaches the top of the trapping layer. Radiation approaching at steeper angles will penetrate the duct. These rays are then rapidly bent down until they strike the surface. After reflection they will not be refracted sharply enough to prevent them from leaving the duct at which point they are rapidly refracted out to space. In this example it is apparent that there is a shadow zone inside the duct just as there is a shadow zone outside the duct for a transmitter within the duct. Any object in the shadow zone is incapable of transmitting to or receiving a signal from the transmitter/receiver.

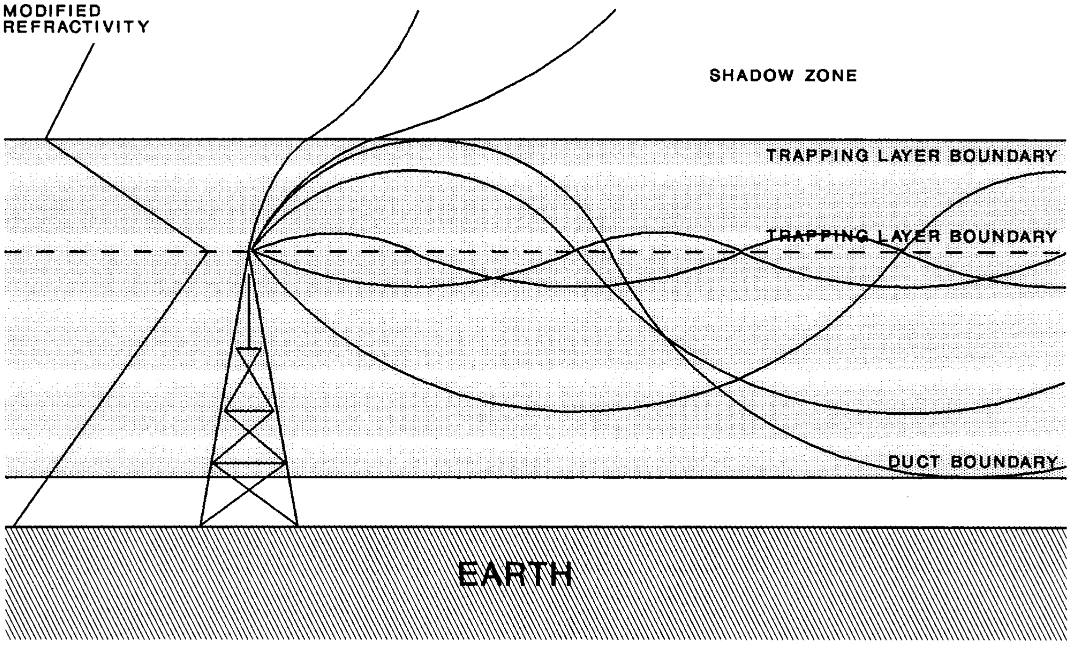
Figure 4-11. Penetration of a surface duct by a transmitter above the duct.



The ducts in Figures 4-10 and 4-11 act via a combination of reflection and refraction. If the trapping layer is elevated, as shown in Figure 4-12, then reflection no longer plays a role. Nevertheless, a duct can be formed using only refraction. In the trapping layer there is strong downward curvature. Just below the trapping layer there is upward curvature. The magnitude of the downward curvature is independent of the magnitude of the upward curvature. Consider an emitter within the duct and near the bottom of the trapping layer. Rays emitted upward are curved downwards by the trapping layer; rays emitted downward are curved upward by the "normal" layer below the trapping layer. As soon as an upward traveling ray is bent back toward the surface enough that it exits the bottom of the trapping layer, it experiences a different curvature and starts to bend upward away from the surface. When it reenters the trapping layer, the curvature reverses and the cycle begins again. The analogous process happens to most downward traveling rays. Once again we have a situation in which periodic oscillation occurs with minimal spreading loss in the vertical direction.

If an initially upward traveling ray exits the trapping layer before it can be completely redirected downward, then that ray will escape to infinity. If an initially downward traveling ray has enough downward motion to carry it beyond the lower boundary of the duct (defined by a refractivity equal to that of the minimum value at the top of the duct), it will either strike the surface and reflect or it will reverse direction and curve upward. In either case, the ray has enough upward direction that it will go past the top of the trapping layer and escape. A shadow zone exists in elevated ducts as well as surface ducts.

Figure 4-12. Trapping by an elevated negative-gradient layer.



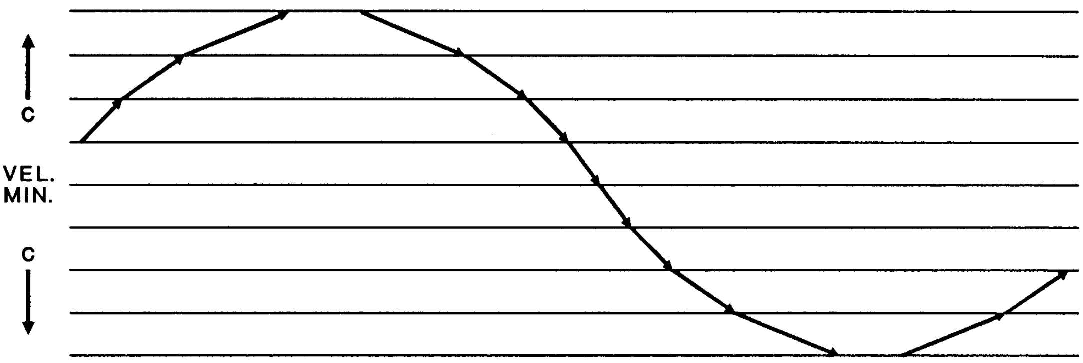
Figures 4-10 through 4-12 are drawn assuming only linear gradients. In this case all curvatures are constant and it is easily to trace the rays. However, as evidenced by the profile of an evaporation duct, Eq. (4.39), linear gradients are the exception not the rule. Although the instantaneous curvature is always given by Eq. (4.40), if the gradient is a function of height, then the curvature will be a function of height. Tracing of ray paths will be more difficult. There are a number of approaches to determining the flow of radiation in a ducting environment. One is to analytically solve the differential equations. This is practical only for certain idealized profiles. A second approach is to break the atmosphere into many layers assigning a discrete refractive index to each layer. A ray may then be traced by applying Snell's law to each interface it encounters. When the angle of incidence of the ray approaches 90° to within some very tight tolerance, then the ray is reflected. This "**finite element**" or "**finite difference**" approach is schematically illustrated in Figure 4-13. In this figure we have highlighted the fact that propagation velocity is inversely proportional to refractive index. Thus rather than having a local maximum of refractive index (or refractivity) define the duct, a local minimum of propagation velocity is used instead. This should give the reader a hint that we will see this material again (when we discuss propagation of acoustic radiation). Light, like sound, is "lazy" and seeks to move at the minimum allowable velocity. Wherever possible, sound or light rays oscillate about the velocity minimum and remain in the duct. A third approach is also a finite element or finite difference approach. In this approach the atmosphere is layered, but each layer is assigned a finite curvature. Each ray is propagated from layer to layer along piecewise circular paths.

Although the first method can only be used occasionally, there is one profile for which exact results can be obtained: the parabolic index profile

$$n(r) = n_0 - 0.5 n_2 r^2 \tag{4.41}$$

Assume that this profile is symmetric about the z-axis. If we restrict the radial distance to small

Figure 4-13. Finite element (finite difference) approach to ray path calculation.



values, we have the paraxial approximation. The position of any ray can be expressed as

$$\vec{r} = r(z)\hat{\rho} + \vec{z} \ , \tag{4.42}$$

where $\hat{\rho}$ is a unit vector perpendicular to the z-axis. The equation describing a curved ray can be shown to be [6],[7]

$$\frac{d}{ds}\left(n\frac{d\vec{r}}{ds}\right) = \nabla n \tag{4.43}$$

where *s* is the distance along the ray. In the paraxial approximation this becomes

$$\frac{d^2r}{dz^2} = \frac{1}{n(r)}\frac{dn(r)}{dr} \ . \tag{4.44}$$

Now from the parabolic index profile we can extract

$$\frac{dn}{dr} = -n_2 r \ . \tag{4.45}$$

Substituting this expression into Eq. (4.44) we have

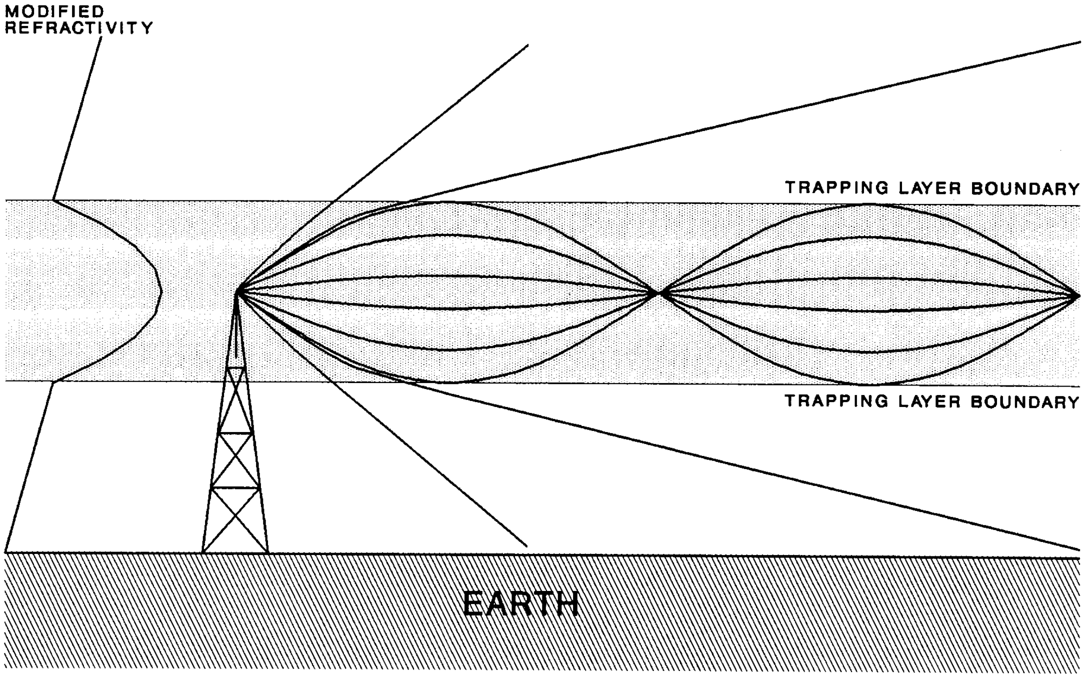
$$\frac{d^2r}{dz^2} = -\frac{n_2}{n_0} r \tag{4.46}$$

which has sine and cosine solutions. Specifically, if r_0 is the initial value of the radial displacement and r_0' is the initial slope of the ray, then the solutions can be written as

$$r(z) = r_0 \cos\left[(n_2 / n_0)^{1/2} z\right] + r_0'(n_0 / n_2)^{1/2} \sin\left[(n_2 / n_0)^{1/2} z\right]. \tag{4.47}$$

The parabolic index profile is often a better approximation to a real profile than is the constant gradient profile assumed in our earlier analyses. It also finds real application in gradient index fiber optics. Figure 4-14 shows an elevated duct with a hypothetical parabolic index profile. All rays have the same frequency so the pattern alternately focuses and defocuses as it propagates down the duct. We will see such refocusing behavior again when we discuss convergence zones in acoustic

Figure 4-14. Trapping in an elevated duct with a parabolic index profile.



propagation. The spatial repeat frequency in this instance is

$$k = (n_2 / n_0)^{1/2} . \qquad (4.48)$$

There are a limited number of computer codes available to assist in calculating the effects of refractive propagation on sensor systems. One of the more widely available codes is EREPS (Engineer's Refractive Effects Prediction System).[5] In addition to the duct database (SDS) contained in the code, EREPS has several parts: a ray tracing program (RAYS), path propagation loss programs (PROPR and PROPH), and a graphical program (COVER) for predicting spatial detection coverage. EREPS was intended to assist users in assessing the impact of lower atmosphere propagation effects on radars, electronic warfare systems, and communications links. It is applicable over the spectral range from 100 MHz to 20 GHz. The models account for multipath (see Chapter 5), diffraction, refraction, evaporation and surface-based ducts, water vapor absorption, and sea clutter for horizontally-homogeneous atmospheric conditions. The ray tracing program (RAYS) takes a user-defined refractivity profile (consisting of up to 14 regions of linear refractivity gradient) and uses repeated application of Snell's law to project the ray path.

All of the preceding effects occur in the microwave region of the spectrum. In the visible and infrared region of the spectrum, the same ducting and other refractive phenomena occur. However, there are some differences. Most of these have to do with the sources of refractive anomalies. In the microwave region we found that water vapor is the driving force behind anomalous refraction. In the visible and infrared we shall see that water vapor variations are much less significant a driver than are temperature variations.

Close examination of Eq. (4.2) and (4.3) shows that in the visible and infrared, the sensitivity of N to temperature changes (ΔN per Kelvin) is roughly four times the sensitivity to water vapor (ΔN per mbar H_2O). For example

$$\left.\frac{dN}{dT}\right|_{p_A=1013\,\text{mbar};\; p_{H_2O}=5\,\text{mbar};\; \lambda=1\,\mu\text{m}} \approx \frac{-82190}{T^2} \tag{4.49}$$

versus

$$\left.\frac{dN}{dp_{H_2O}}\right|_{p_A=1013\,\text{mbar};\; T=273.15\,\text{K};\; \lambda=1\,\mu\text{m}} \approx -0.254\,. \tag{4.50}$$

This is a serious difference between the visible and the microwave region. Thus, refractivity gradients are much more likely to be driven by temperature gradients than by water vapor gradients. Subrefractive and superrefractive (even trapping) layers will still exist but they will be caused by different circumstances.

A negative temperature gradient is commonly produced over hot surfaces (the surface is hotter than the air temperature) and will produce a strongly positive refractivity gradient. This produces subrefractive behavior. A subrefractive atmosphere will produce inferior mirages.[8] In an inferior mirage, an elevated object appears lower than normal. An inferior mirage in which the sky is viewed as being “lower” than the surface of a highway (and is invariably perceived as water on the road) is perhaps the most commonly observed manifestation of an inferior mirage. Other manifestations can appear as terrain features being “reflected” by a flat surface such as water. The refractive behavior leading to an inferior mirage is illustrated in Figure 4-15.

A positive temperature gradient (a temperature inversion) can be produced over cold surfaces. If the gradient is strong enough then superrefractive or even trapping behavior occurs. Super-refractive conditions produce superior mirages.[8] Less common than inferior mirages, superior mirages can be much more striking. Objects can appear to float in the air. Multiple images of an object can appear; some of which are inverted or grossly distorted. Sometimes small surface height variations are magnified into huge pillars that can be mistaken for structures or forbidding terrain features. These “Fata Morgana” are awe-inspiring examples of superior mirages. The refractive conditions that can produce a simple superior mirage are shown in Figure 4-16. More complicated mirages result when several superrefractive and normal or subrefractive layers occur at different altitudes.

Figure 4-15. Refractive paths in an inferior mirage with inverted "reflection".

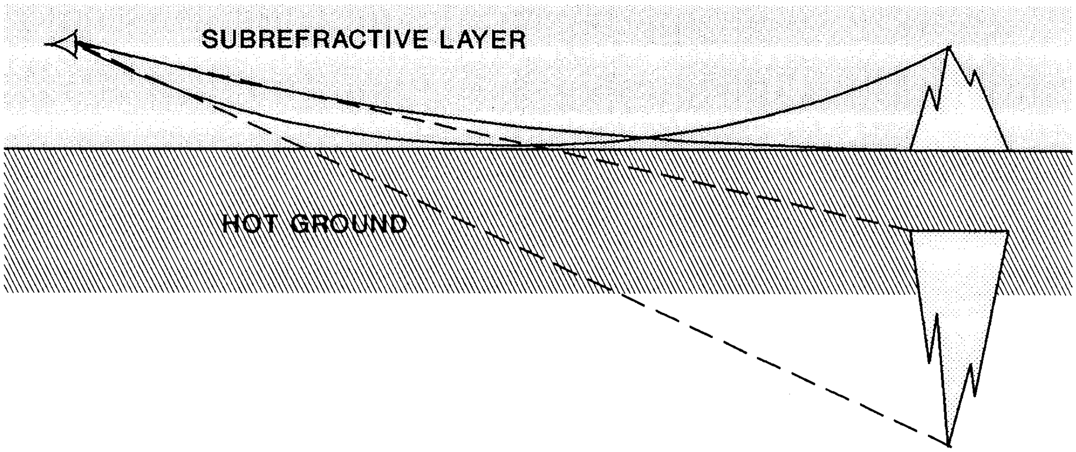
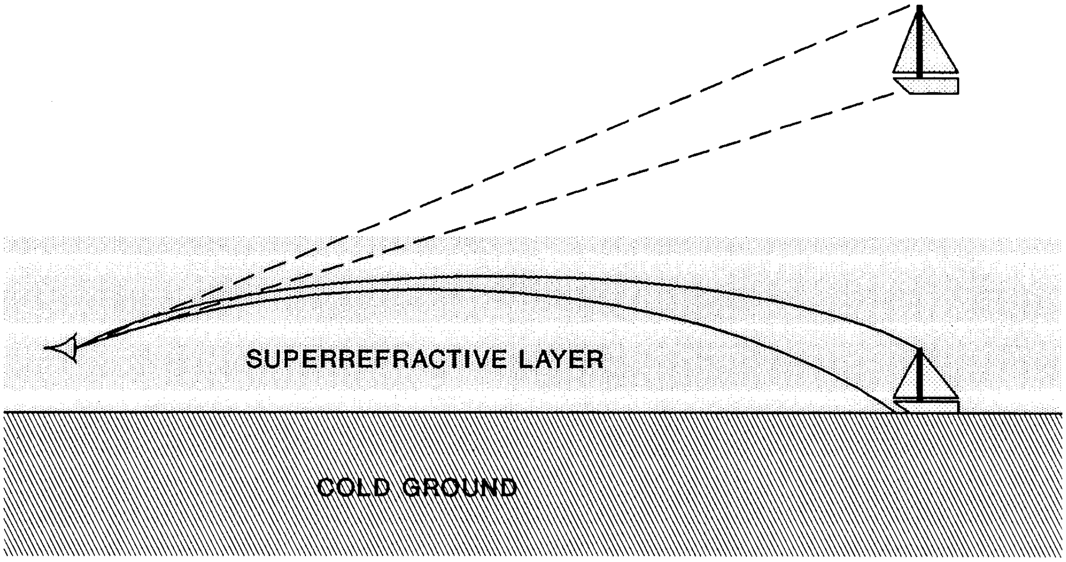


Figure 4-16. Refractive paths in a superior mirage with upright image.



Atmospheric Turbulence

Atmospheric turbulence has two different but somewhat related definitions. The first definition relates to medium-scale atmospheric motions more properly called wind shears. "Turbulence" produced by wind shear is the effect felt by airline passengers as their aircraft encounters spatially adjacent updrafts and downdrafts. The rapidly changing accelerations produced by the winds are felt as bumps, lurches, and/or violent shaking. The atmospheric motions may be produced by airflows deflected by mountains, waves at the boundaries of jet streams, circulation associated with storms, or rising columns of solar heated air interspersed with the descending currents of colder air required to maintain rough atmospheric pressure equilibrium. These columnar motions are the largest scales of the phenomenon that affects sensor systems. The second definition of atmospheric turbulence, which is the one we are concerned with in this section, is the turbulent mixing of eddies of air with slight temperature differences (typically $\Delta T \sim 1$ K). The temperature changes produce slight refractive index variations ($\Delta n \sim 10^{-6}$) which in turn produce random lensing and interference effects in electromagnetic radiation.

Atmospheric turbulence is produced by several effects. However, the strongest driver is solar heating of the ground. Sunlight heats the ground to temperatures that may be considerably above the ambient air temperature. It is not uncommon for the ground temperature to exceed 60 °C and possible for it to exceed 90 °C under conditions of the strongest solar irradiance. The hot surface heats the air in immediate contact. This superheated air rises rapidly. Consider air that is heated only 30 °C hotter than the ambient air temperature. If we consider average lapse rate conditions (6 °C per km), the adiabatic cooling of a rising column of air that was originally 30 °C above ambient will reach the same temperature as the surrounding air at an altitude of 10 km. Thus we see that the solar heated air can easily rise up to the tropopause. If the heated air rises, some amount of colder air must fall to replace the rising hot air and prevent formation of a vacuum. The cold air cannot diffuse through the rising hot air so it tends to form descending columns. Similarly the hot air will be channeled into rising columns. These columns will be initially be spatially distinct and adjacent to each other.

However, there is a hydrodynamic phenomenon known as the two-stream instability that rapidly affects both the rising and falling columns. The two-stream instability occurs when a stream of fluid with one density passes through a fluid with a different density. In this instance, any perturbation on the surface of one stream with grow exponentially in time. The growth of these perturbations ultimately causes the stream to break up into globules. The most common visual manifestation of two-stream instability is found in a common garden hose. When a strong stream of water is emitted from the hose, it is initially a smooth cylinder of water. However, after only a few feet, the smooth cylinder breaks up into dozens of droplets of varying sizes. This is the result of the two-stream instability (the water is one stream, the air constitutes the other). The same effect causes the rising hot air columns and the falling cold air columns to break up into irregularly shaped "turbules" of hot air and cold air. Some of these turbules interact with each other and break up into smaller and smaller turbules. The process of interaction imposes a degree of random motion onto the dominant streaming motion, while winds will impart lateral motions to the turbules.

Any condition which produces air hotter than the ambient temperature will produce turbulence. Direct solar heating produces strong turbulence. However, even at night, the thermal inertia of the ground may keep it warmer than the ambient air. This will also produce turbulence. However, the strength of the turbulence depends on the maximum ΔT that is produced. Nighttime turbulence is almost never as strong as that produced by solar heating. By the same token, if the sun temporarily goes behind a cloud, the direct solar heating is reduced and the turbulence strength will fall dramatically. On overcast days, turbulence strength may not exceed nighttime levels. Fires or other strong heat sources will produce intense turbulence although only above the region that is burning or has recently burnt. Turbulence over water is usually much weaker than turbulence over the land. Ground absorbs almost all of the solar radiation right at the surface; but sunlight can penetrate meters into water. The sea surface is not strongly heated by sunlight, so there is no strong temperature difference produced. Turbulence over water is typically less than nighttime turbulence over land.

The interaction of wind with terrain, vegetation, and structures will also produce atmospheric turbulence. Dynamic (wind-driven) turbulence is typically weaker and less persistent than solar-driven turbulence. The eddies produced by dynamic turbulence do not have significant temperature differences, just pressure differences. The pressure disturbances will equalize more quickly (decay times are typically less than a minute for disturbances produced by structures smaller than a few tens of meters) than the temperature differences. For this reason dynamic turbulence is not significant at altitudes much above the altitude of the highest up-wind obstructions. It is also not significant if the optical paths of interest lie more than a few hundred meters downwind from vegetation sources of turbulence (forested areas). High-speed flow around airborne platforms can become turbulent. The initially laminar flow along the airframe can become unstable and breakup. Such boundary layer turbulence is limited in thickness, but its location right at sensor apertures on board the platforms can cause problems. Because of the complexity of the problem and its analysis we will not address boundary-layer turbulence beyond the mention of its potential impact.

Turbulent eddies will range in size from millimeters (the so-called “inner scale”) to tens or hundreds of meters (the so-called “outer scale”). The larger scales only become apparent at heights comparable to their diameters. Convective mixing coupled with adiabatic lapse causes the turbulence to diminish in strength as the altitude increases. We will return to the altitude dependence of turbulence in a few paragraphs. Drift dominates the evolution of the atmospheric eddies. Convective rise or wind produces this drift. As a consequence, the coherence time of the atmosphere as viewed by a typical sensor is of the order of milliseconds. That is, the ensemble of turbules viewed by a sensor at one instant of time will be completely different from the ensemble of turbules viewed by that sensor a few milliseconds earlier or a few milliseconds later.

Atmospheric turbulence affects electromagnetic fields during the propagation of those fields. The Huygens-Fresnel equation (Eq. (2.58)) describes the propagation of electromagnetic fields under the influence of diffraction. To describe the effects of turbulence on propagation, we need to modify (extend) the Huygens-Fresnel equation. The unmodified equation has the form

$$\xi\left(x, y, z_0 + L\right) = \frac{ie^{-ikL}}{L\lambda} \iint\limits_{\substack{\text{input}\\ \text{plane}}} dx_0 dy_0 \; \Big\{ \xi\left(x_0, y_0, z_0\right) \times \exp\left[\frac{-ik}{2L}\left[\left(x - x_0\right)^2 + \left(y - y_0\right)^2\right]\right]\Big\} \tag{4.51}$$

To accommodate the effects of turbulence we add another term to the equation and obtain the Extended Huygens-Fresnel Equation

$$\xi\left(x, y, z_0 + L\right) = \frac{ie^{-ikL}}{L\lambda} \iint\limits_{\substack{\text{input}\\ \text{plane}}} dx_0 dy_0 \; \Big\{ \xi\left(x_0, y_0, z_0\right) \times \exp\left[\frac{-ik}{2L}\left[\left(x - x_0\right)^2 + \left(y - y_0\right)^2\right]\right] \times \exp\left[\chi\left(x_0, y_0, x, y\right) + i\phi\left(x_0, y_0, x, y\right)\right]\Big\} \; . \tag{4.52}$$

This added term has a real part and an imaginary part. The argument of the real part χ is called the log-amplitude (e^{χ} is an amplitude factor, so χ is the log of the amplitude – the log-amplitude). The imaginary part acts like any other phase term (typically an aberration or distortion) in the diffraction equation. The effects of the two different parts of the turbulence extension will become apparent shortly.

Atmospheric turbulence can affect sensors in a number of ways. The potential effects of atmospheric turbulence on sensor performance are listed in Table 4-3. Of these effects, scintillation

Table 4-3. Potential effects of atmospheric turbulence on sensor performance.

SCINTILLATION

* TRANSMITTER-TO-TARGET PATH
* TARGET-TO-RECEIVER PATH

TRANSMIT BEAM SPREAD

* INSTANTANEOUS BEAM BREAKUP
* BEAM WANDER

RECEIVER ANGULAR SPREAD

* FIELD OF VIEW MISMATCH
* IMAGE SPECKLE
* IMAGE WANDER

RECEIVER COHERENCE LOSS

is related to the log-amplitude. Beam spread, angular spread, and coherence loss are associated with the turbulent phase term.

One of the most visible manifestations of turbulence is scintillation (fluctuations in intensity of a source). The twinkling of stars in the early evening is the effect of scintillation on the human visual sensor. Scintillation can occur on any propagation path through the atmosphere. If scintillation occurs on the transmitter-to-target path then the illumination may fluctuate strongly from point to point on the target. If the target has a few strong scatterers, the turbulent fluctuations in the illumination of those scatterers may overwhelm the fluctuations induced by the target alone. If the target is composed of many small scatterers, there may be some averaging of the turbulent fluctuations by the target. If the fluctuations occur on the return path, they will add to the target fluctuations and cannot be averaged over the aperture.

Phase perturbations of the radiation field cause beam spread, receiver angular spread, and coherence loss. A coherent beam of radiation can undergo spreading due to turbulence that is greater than that produced by diffraction. The total beam spreading will have three components: a component due to diffraction, an instantaneous breakup of the beam into high-field and low-field regions, and a slower beam wander. Beam wander and beam breakup will be roughly comparable in magnitude to each other. That is, the range of angles through the beam wanders will be roughly equal to the turbulent angular spread observed in the broken up beam. The breakup pattern will change on tens of milliseconds time scales, the same time scale as the beam wander.

Beam spreading results from angular phase perturbation imposed on a transmitted beam by turbulence acting near the source of the beam. Receiver angular spread results from angular phase perturbations imposed on a received beam by turbulence acting near the receiver. The high spatial frequency components (which correspond to those causing beam breakup on transmit) act to smear the received signal over a large image area into a random pattern of bright areas and dark areas – a speckle pattern. The low spatial frequency components act to cause the speckle pattern to wander about the image plane (the received signal analog of transmit beam wander). Both effects can cause some of the signal to miss the area delineated by the detectors. This effect called field of view mismatch results in reduced signal strengths.

A final phase perturbation effect can occur in coherent detection (heterodyne) systems. The random phase variations imposed on the signal will reduce the spatial coherence of that signal and reduce the efficiency with which a smooth local oscillator can mix with the signal. It is not possible to predict the phase variations in advance, so it is not possible to match the local oscillator to the signal, which is necessary for efficient heterodyne conversion.

Both the phase and amplitude perturbations can be characterized as Gaussian processes whose magnitude depends on a quantity C_n^2, the refractive index structure coefficient. The refractive index structure coefficient describes the spatial variability of the refractive index. Because the refractive index variations are caused by temperature variations, measurements are more common made of the temperature structure coefficient C_T^2 defined by

$$C_T^2 = \frac{\left\langle T(\vec{r}_1) - T(\vec{r}_2)\right\rangle^2}{\left|\vec{r}_2 - \vec{r}_1\right|^{2/3}} \qquad \text{with units of K}^2 \text{ / m}^{2/3} \qquad (4.53)$$

where $T(r_i)$ is the temperature at position r_i and the < > brackets denote an ensemble average. C_T^2 is determined just as it is defined, by making simultaneously temperature measurements at two points and carrying out the averaging over time. The temperature structure coefficient is related to the refractive index structure coefficient by the relation

$$C_n^2 = \left|\frac{\partial n(\lambda)}{\partial T}\right|^2 C_T^2 \qquad \text{units of m}^{-2/3}. \qquad (4.54)$$

Upon examination of Eq. (4.2) for the refractive index it can be shown that the temperature derivative takes the form

$$\frac{\partial n(\lambda)}{\partial T} = \frac{A(\lambda)\times 10^{-6}}{T^2} \qquad \text{units of K}^{-1} \qquad (4.55)$$

where $A(\lambda)$ varies from approximately 86 at a visible wavelength of 500 nm to somewhat less than 80 at an infrared wavelength of 10 μm. The dispersion of refractive index in air can almost always be considered negligible and a single value of A (a choice A = 83 is no better or no worse than any other choice in the range of 80-86) selected.

Limited data has been obtained on the temporal and spatial variability of the refractive index structure coefficient, although there is enough to make working generalizations. Table 4-4 shows data collected at 3m height above ground in upstate New York during the months of February to May.[9] A number of samples were obtained each hour of every day. Average daytime turbulence levels (as characterized by C_n^2) are seen to be around 10^{-13} $m^{-2/3}$. Average nighttime values are roughly 10^{-14} $m^{-2/3}$. The units of the structure coefficient appear strange. Nevertheless, if the values associated with these units are inserted into any of the equations that follow, then the calculated results will have the correct units. The values shown in Table 4-4 are consistent with other measurements, summarized in Reference 10.

Theoretical work has shown that solar-driven turbulence should have an altitude dependence that allows C_n^2 to decrease as the altitude to the -4/3 power.[11] In the absence of the solar driver, the turbulence should rapidly dissipate to lower overall strengths and with a less pronounced falloff with altitude. The data in Table 4-4 support this conclusion. Type I statistics are fair weather, clear sky data – days where a prominent solar driver is anticipated. The all weather statistics includes overcast days and days with rain and/or snow. Note that the all-weather values are closer to nighttime levels than to the Type I daytime levels.

Table 4-4. Monthly-averaged turbulence statistics collected at 3m height.

		LOG C_n^2 STATISTICS (TYPE I)		LOG C_n^2 STATISTICS (ALL WEATHER)	
		MEAN	STD. DEV.	MEAN	STD. DEV.
FEB	NIGHT	-14.13	0.58	-14.12	0.68
	DAY	-12.76	0.41	-13.63	0.76
MAR	NIGHT	-14.33	0.60	-14.05	0.75
	DAY	-12.70	0.52	-13.12	0.82
APR	NIGHT	-13.67	0.43	-13.64	0.51
	DAY	-12.79	0.62	-12.82	0.63
MAY	NIGHT	-13.49	0.54	-13.54	0.57
	DAY	-12.56	0.37	-12.85	0.63

TYPE I: CLEAR SKY; FAIR WEATHER

In the majority of his work on turbulent effects the author has used the following general values for estimating turbulence strength:

DAY	3m Height, Average Terrain, Clear Weather	C_n^2 (3m) = 10^{-13} $m^{-2/3}$
DAY	3m Height, Desert, Clear Weather	C_n^2 (3m) = 10^{-12} $m^{-2/3}$
DAY	3m Height, Any Terrain, Adverse Weather	C_n^2 (3m) = 10^{-14} $m^{-2/3}$
DAY	3m Height, Over Water, Any Weather	C_n^2 (3m) = 10^{-15} $m^{-2/3}$
NIGHT	3m Height, Any Terrain, Any Weather	C_n^2 (3m) = 10^{-14} $m^{-2/3}$
NIGHT	3m Height, Over Water, Any Weather	C_n^2 (3m) = 10^{-15} $m^{-2/3}$

The author also uses the following altitude dependences below 100 m:

DAY	Any Terrain, Clear Weather	C_n^2 (h) = C_n^2 (3m) $(h/3)^{-4/3}$
DAY	Any Terrain, Adverse Weather	C_n^2 (h) = C_n^2 (3m) $(h/3)^{-2/3}$
DAY	Over Water, Any Weather	C_n^2 (h) = C_n^2 (3m) $(h/3)^{-2/3}$
NIGHT	Any Terrain, Any Weather	C_n^2 (h) = C_n^2 (3m) $(h/3)^{-2/3}$

and the following altitude dependence above 100 m:[12]

$$C_n^2(h) = 8.2 \times 10^{-56} U^2 h^{10} e^{-h/1000} + 2.7 \times 10^{-16} e^{-h/1500} \tag{4.56}$$

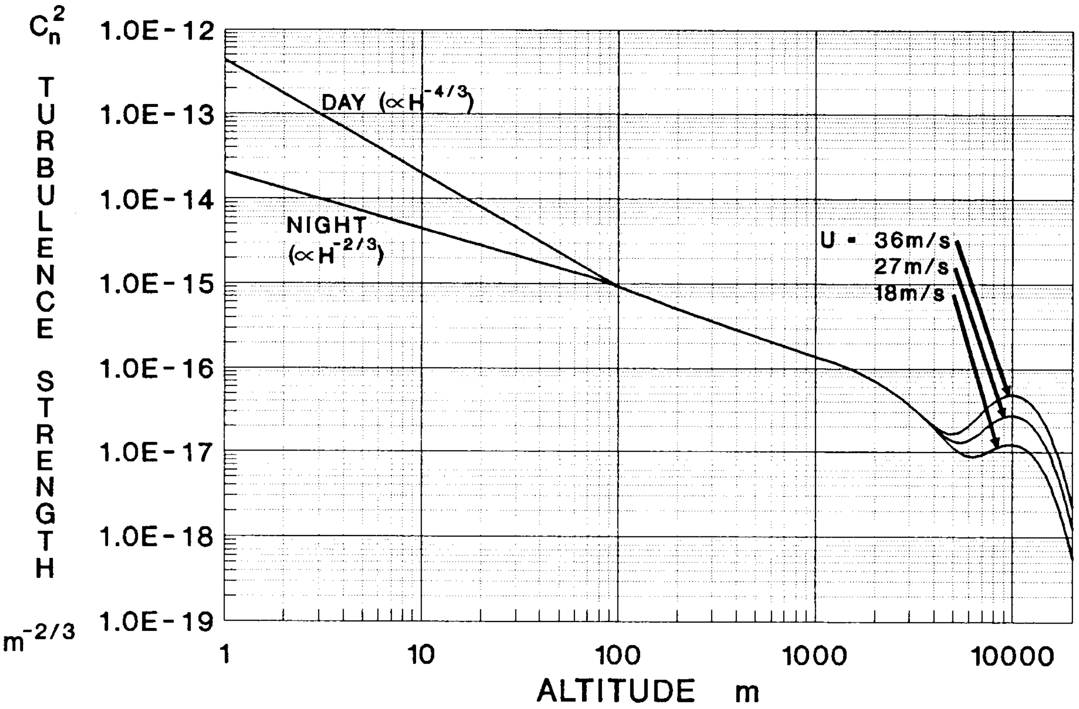
where U is the root-mean-squared wind speed averaged over the 5000-20000 m altitude regime. U has a mean value of 27 m/s and a standard deviation of 9 m/s. The altitude dependence of Eq. (4.56) is technically only valid above 1000 m or so, however, extrapolation to lower altitudes leads to a result almost equal to the extrapolations of the lower altitude profiles to higher altitudes. In the 100-1000 m altitude regime, the differences between the different extrapolations are relatively small. The combined dependence is shown in Figure 4-17. With this knowledge of the values of C_n^2 and its spatial dependence, we can proceed to look at specific turbulence effects.

The log-amplitude χ (after exponentiation) modifies the amplitude of the electric field and consequently the magnitude of the intensity

$$I = I_0 e^{2\chi} \tag{4.57}$$

Fluctuations in the log-amplitude translate into intensity fluctuations. These fluctuations are called

Figure 4-17. Generic altitude dependence of atmospheric turbulence.



scintillation. Figure 4-18 shows the effect of scintillation on a laser beam courtesy of J. Shapiro at M.I.T. An initially uniform intensity beam was transmitted at night from the roof of one building to the roof of another building several kilometers distant. The turbulence amplitude modulated the beam with a random pattern of more intense regions and less intense regions. The peak-to-valley range of the fluctuations increases as propagation distance increases and as the turbulence strength (C_n^2) increases. Note that the scintillation is apparent in both the focused and unfocused beams. The beams in Figure 4-18 have also propagated far enough to show significant beam spreading. The focal spot of the unspread beam would have been only slightly larger than the size of an individual "speckle".

Turbulence tends to produce fluctuations in the log-amplitude that are Gaussianly distributed

$$p_\chi(\chi) = \left(2\pi\sigma_\chi^2\right)^{-1/2} e^{-\left(\chi+\sigma_\chi^2\right)^2 / 2\sigma_\chi^2} \tag{4.58}$$

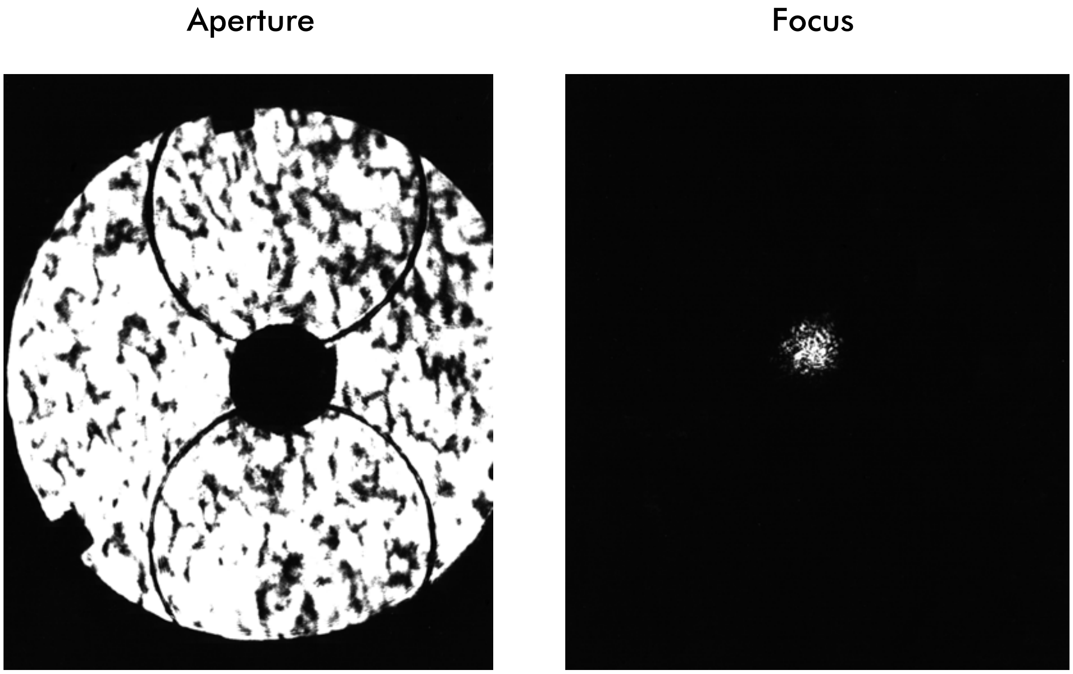
with mean value

$$\langle \chi \rangle = -\sigma_\chi^2 \tag{4.59}$$

and variance

$$\mathrm{var}(\chi) = \sigma_\chi^2 \tag{4.60}$$

Figure 4-18. Photographs of scintillation on a laser beam taken at the aperture of a collecting telescope and at the focal plane of that telescope.



The key quantity σ_χ^2 is called the log-amplitude variance. It can be calculated using the equation

$$\sigma_\chi^2 = 0.56 k^{7/6} R^{11/6} \int_0^1 d\zeta \, C_n^2(\zeta) \, \zeta^{5/6} (1-\zeta)^{5/6} \tag{4.61}$$

where

$$k = 2\pi / \lambda \tag{4.62}$$

In this equation we have defined a normalized path length

$$\zeta = z / R = \text{Normalized distance along path} \tag{4.63}$$

This permitted all dimensioned quantities except the refractive index structure coefficient to be brought outside the integral. The wavelength and path length dependence become explicit. This normalization will be used for all turbulence calculation. If the refractive index structure coefficient is constant over the path, then the equation for the log-amplitude variance simplifies to

$$\sigma_\chi^2 = 0.124 \; C_n^2 \; k^{7/6} \; R^{11/6} \tag{4.64}$$

When scintillation fluctuations become large enough they undergo a phenomenon called saturation of scintillation, that is, even though C_n^2 continues to grow, σ_χ^2 stops growing and levels off at a value of roughly 0.5. However, by the time saturation of scintillation has been reached, turbulence effects are highly significant and sensor performance is usually severely degraded. An analysis of the effects of scintillation on laser radar detection performance is presented in Chapter 11. The degrading effects of scintillation on laser radar imaging performance are analyzed in Chapter 14. In light of these results, it is a worthwhile objective to try to keep the log-amplitude variance at least an order of magnitude below saturation. If scintillation is not saturated, then the spatial intensity correlation distance (scale size) of the scintillation light and dark patches is given by

$$d_C = (\lambda R)^{1/2} \, . \tag{4.65}$$

Figure 4-19 shows the turbulence strength dependence of the log-amplitude variance at 10.6 μm over a horizontal path (constant C_n^2). Figures 4-20 and 4-21 show the wavelength dependence of the log-amplitude variance for the typical daytime and nighttime values of turbulence. It should be noted that under daytime conditions it is relatively easy for the log-amplitude variance to become saturated over horizontal path lengths of tactical interest to sensor systems. At night, the reduced turbulence strength permits propagation to considerably longer distances before saturation occurs.

Beam spread and angle of arrival fluctuations are characterized by the atmospheric coherence length ρ_0. The coherence length depends critically on whether the strongest turbulence is near the beginning of a propagation path or near the end. Thus, we can define one coherence length for paths

Figure 4-19. Turbulence strength dependence of the log-amplitude variance.

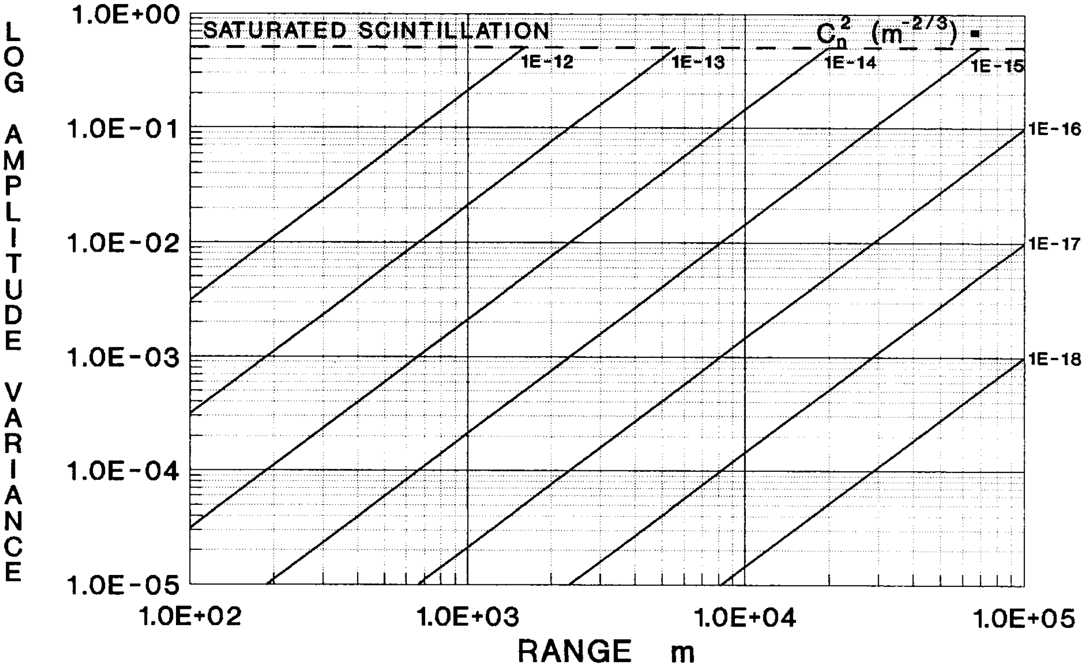


Figure 4-20. Wavelength dependence of the log-amplitude variance - typical daytime turbulence strength of $C_n^2 = 10^{-13}$ $m^{-2/3}$.

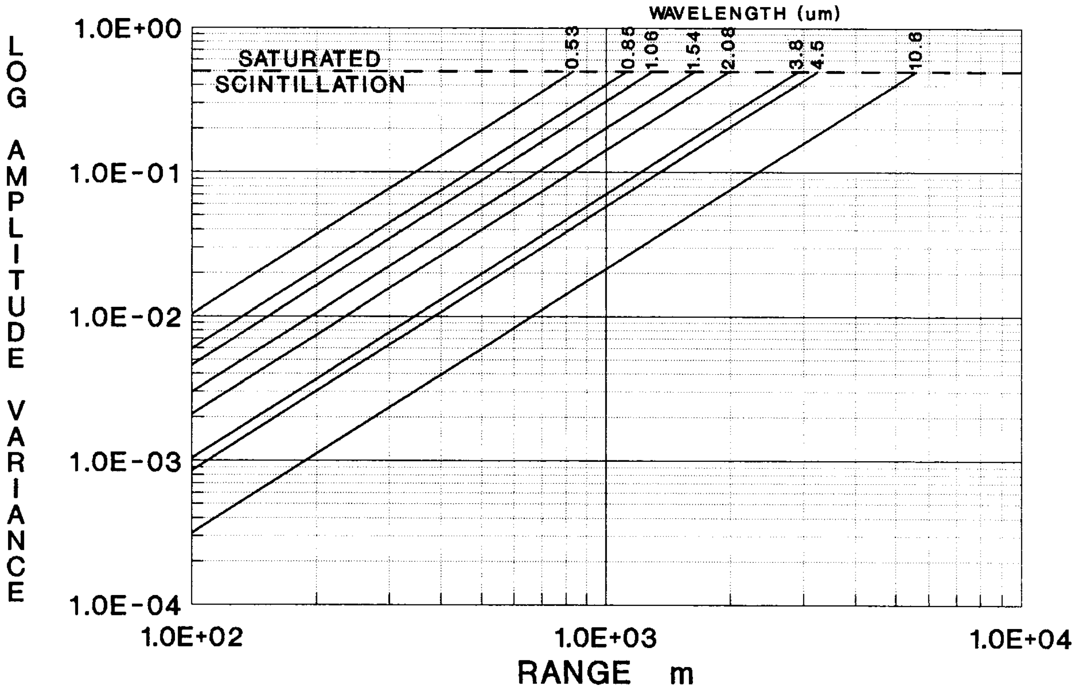
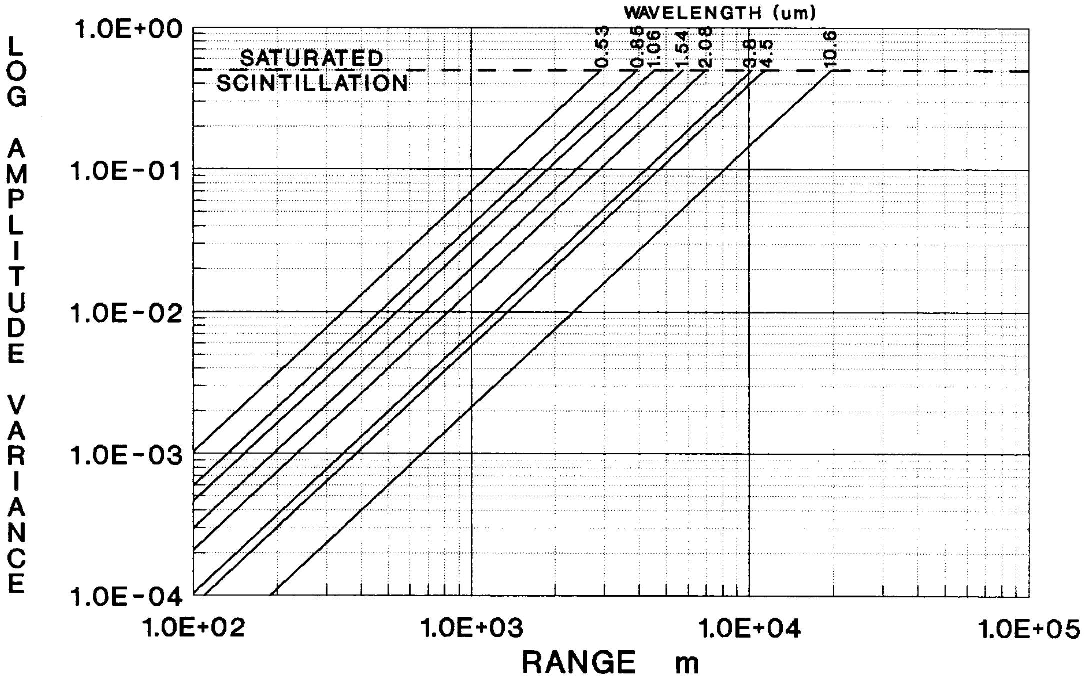


Figure 4-21. Wavelength dependence of the log-amplitude variance - typical nighttime turbulence strength of $C_n^2 = 10^{-14}$ m$^{-2/3}$.



involving transmitted signals

$$\rho_{0T} = \left[2.91 k^2 R \int_0^1 d\zeta \, C_n^2(\zeta) \left(1-\zeta\right)^{5/3} \right]^{-3/5} \tag{4.66}$$

and another for paths involving received signals

$$\rho_{0R} = \left[2.91 k^2 R \int_0^1 d\zeta \, C_n^2(\zeta) \, \zeta^{5/3} \right]^{-3/5} . \tag{4.67}$$

Note the different dependence on ζ in the integrals. Eq. (4.66) weights small values of ζ more strongly than large values, while Eq. (4.67) does the opposite. The atmospheric correlation length for a path is the transverse separation between two points on a beam of electromagnetic radiation required for the phase of the radiation to become uncorrelated. That is, the largest patch on a beam that can be considered to have a uniform phase is ρ_0.

Beam spread is association with the transmission of a beam. Thus, the degree of beam spreading will depend on the transmitted-path correlation length. The physical half-angle by which the beam will spread due to turbulence is given by

$$\theta_{BS} = 0.384\lambda / \rho_{0T} . \tag{4.68}$$

The beam will also spread by normal diffraction by a half-angle θ_D. If the aperture is circular with diameter D and uniformly illuminated, the diffraction beam spread can be found from the Airy function and has half-angle

$$\theta_D = 1.22\lambda / D . \tag{4.69}$$

The total beam spread half-angle will be given by

$$\theta = \left(\theta_{BS}^2 + \theta_D^2\right)^{1/2} \tag{4.70}$$

Turbulence beam spreading will equal diffraction beam spreading if

$$\left(1.22 / 0.384\right)\rho_{0T} = 3.181\rho_{0T} = D . \tag{4.71}$$

For this reason the quantity r_0 , defined by

$$r_0 = 3.181\rho_0 \tag{4.72}$$

where ρ_0 is either the transmitted-path or received-path coherence length, as appropriate, is called Fried's coherence diameter.[13]

On received paths, the atmospheric phase variations will cause angle of arrival variations. The magnitude of these variations can be determined from the equation

$$\theta_{AA} = 0.384\lambda / \rho_{0R} . \tag{4.73}$$

If a horizontal path (uniform C_n^2) is assumed, then the integrals in Eqs. (4.66) and (4.67) simplify and give a single result for the coherence length

$$\rho_{0T} = \rho_{0R} = \left[1.09 C_n^2 k^2 R\right]^{-3/5} = \rho_0 . \tag{4.74}$$

Figure 4-22 shows the turbulence strength dependence of the coherence length for a horizontal path and a wavelength of 10.6 μm. Figures 4-23 and 4-24 show the daytime and nighttime dependence of the coherence length on wavelength. Once again it should be noted that rather small coherence lengths (< 0.1 m) are obtained at moderately short ranges (< 10 km) in a number of cases. This means that turbulent beam spreading might adversely affect the performance of any sensor under those conditions. Beam spread may reduce the power density that illuminates the target or the power

Figure 4-22. Turbulence strength dependence of the horizontal path coherence length at 10.6 μm.

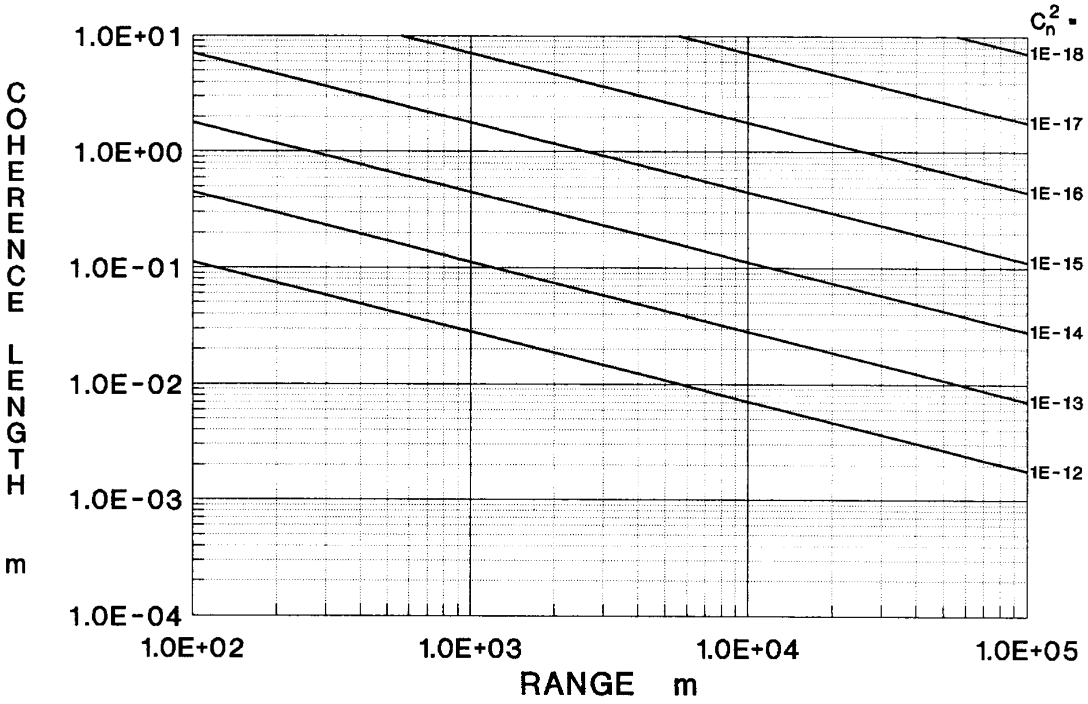


Figure 4-23. Wavelength dependence of the horizontal path atmospheric coherence length – typical daytime turbulence strength of $C_n^2 = 10^{-13}$ $m^{-2/3}$.

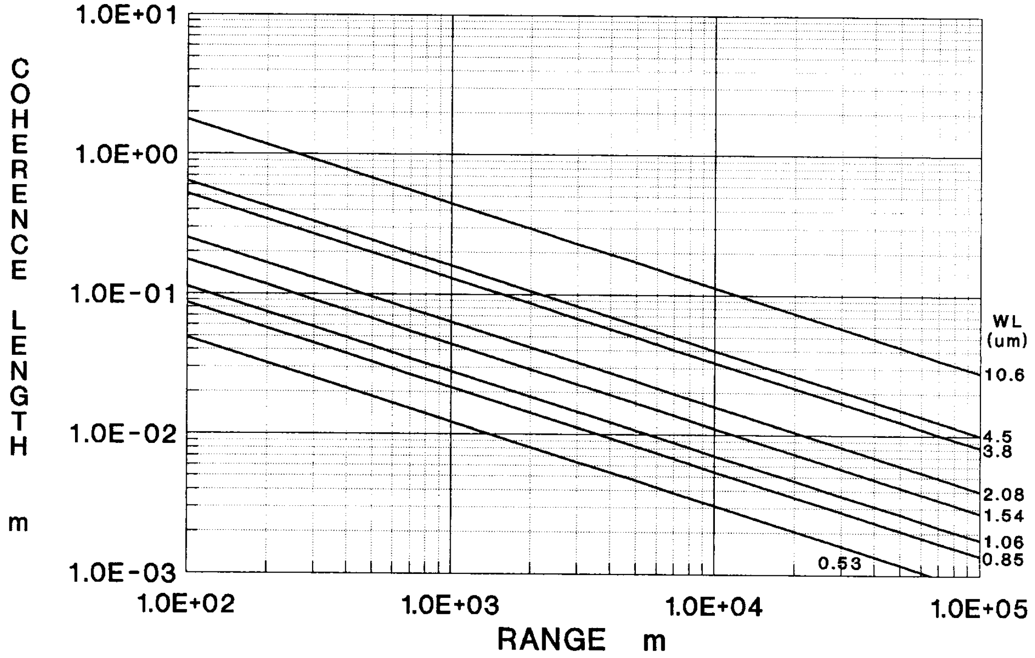
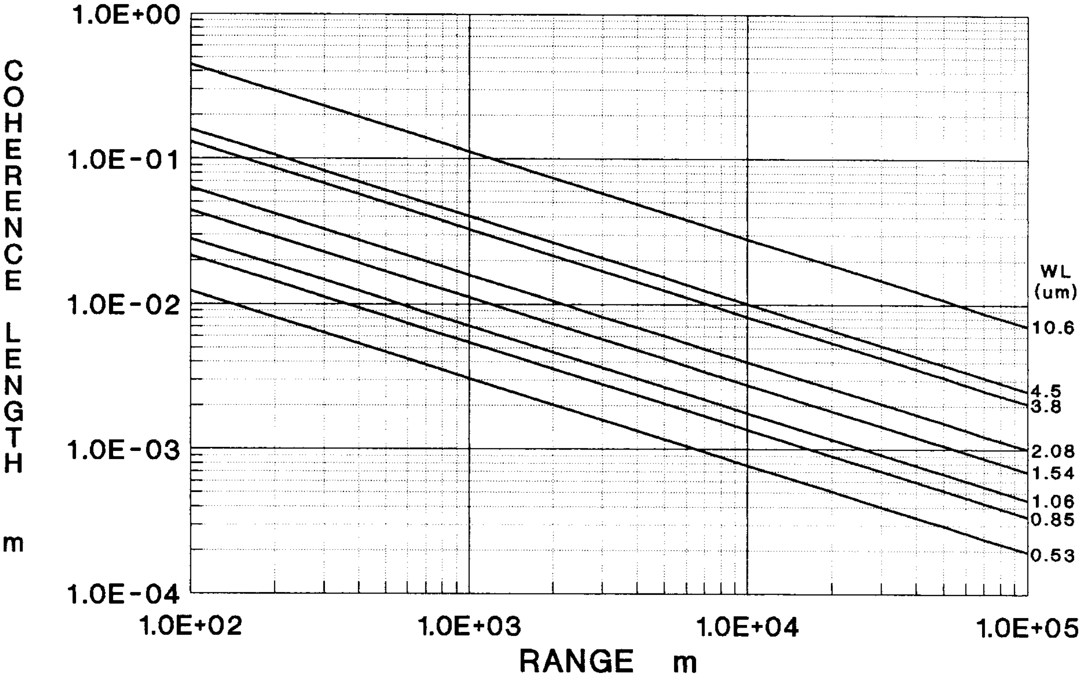


Figure 4-24. Wavelength dependence of the horizontal path atmospheric coherence length – typical nighttime turbulence strength of $C_n^2 = 10^{-14}$ m$^{-2/3}$.

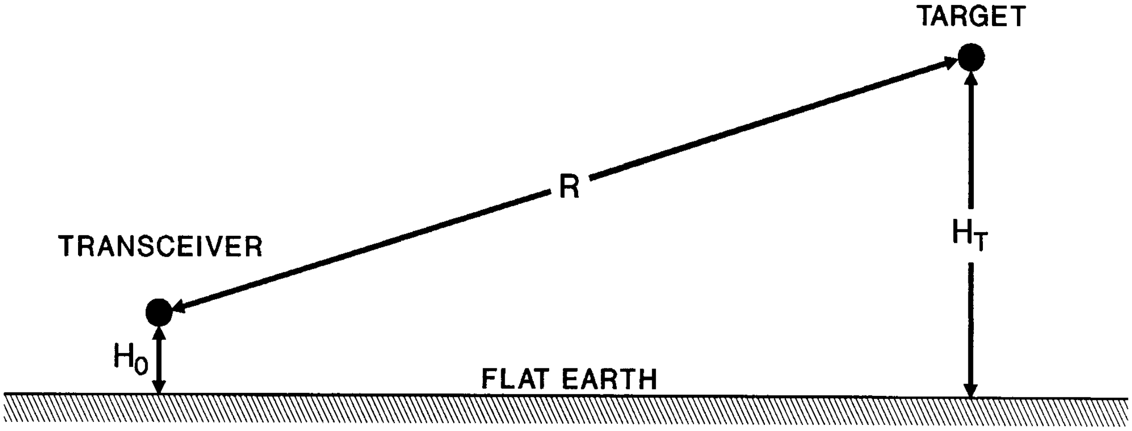


than can be focused onto a detector. It will reduce the effective angular resolution of the sensor. If the sensor uses coherent detection and the coherence length becomes smaller than the receiver aperture, then the random turbulent phase variations will significantly reduce the heterodyne mixing efficiency.

The horizontal path results close to the ground (3 m altitude) are essentially worst case predictions for real sensor systems. Because turbulence strength falls off with increasing altitude, any path characteristic that elevates part of the path (even by a few meters) reduces the effects of atmospheric turbulence. For example, few places in the world are perfectly flat. In order to have a 10 km line of sight, the observer and the target will typically have to be on the tops of local peaks. The valley(s) in between will cause the C_n^2 over those valleys to be lower than at the local peaks. This in turn will cause σ_χ^2 to be much lower and ρ_0 to be much larger than the horizontal path estimates. A common situation is when either the target or the sensor is elevated on a large structure or on an airborne platform. In such slant paths, one end of the path is high above the surface, even the other end may be near the surface. Nevertheless, slant paths provide significant reductions in turbulence effects.

The impact of slant paths can be easily calculated. Consider the geometry illustrated in Figure 4-25. If we use the previously defined normalized path distance variable ζ, then the daytime

Figure 4-25. Geometry of a simple slant path.



profile of turbulence strength with altitude can be written as

$$C_n^2(\zeta) = C_n^2(H_O)\left[1+\frac{H_T-H_O}{H_O}\zeta\right]^{-4/3} = C_n^2(H_O)[1+A\zeta]^{-4/3} \tag{4.75}$$

where we have defined a constant A which defines the slant path

$$A = \frac{H_T-H_O}{H_O}. \tag{4.76}$$

All of the path dependence is now contained in the simple quantity $[1+A\zeta]^{-4/3}$. In a similar fashion the nighttime turbulence profile can be written as

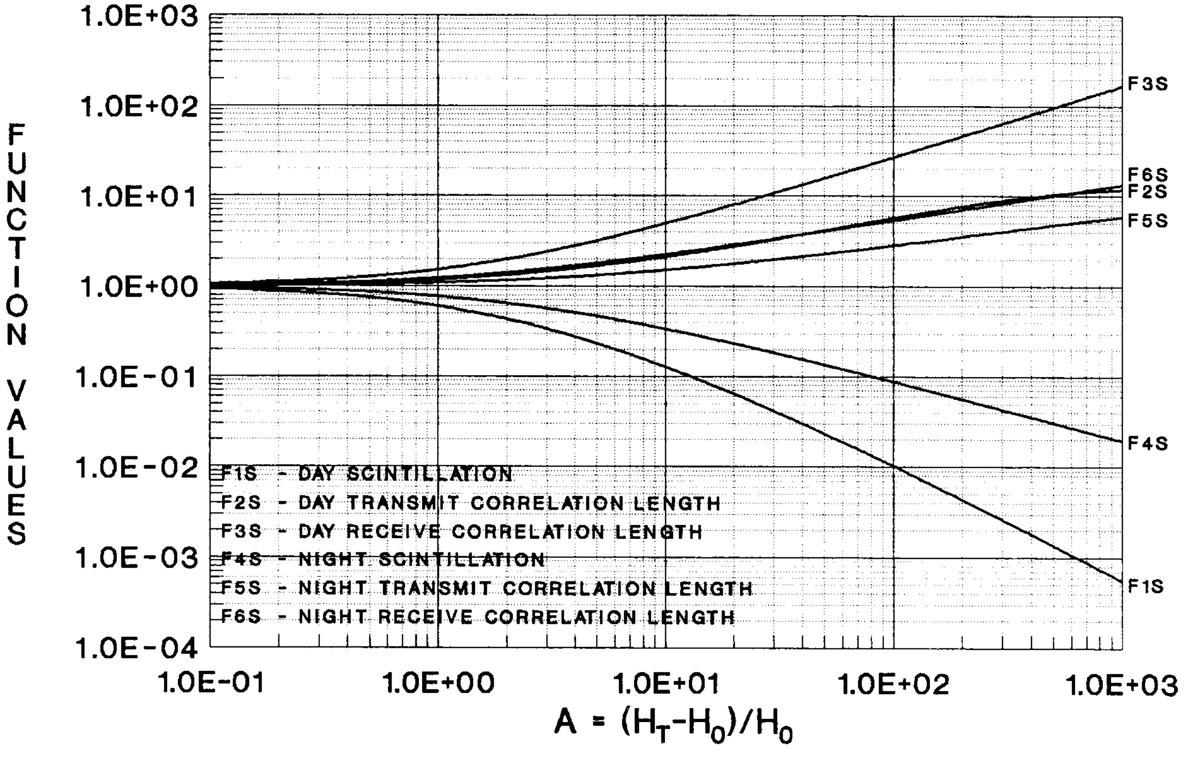
$$C_n^2(\zeta) = C_n^2(H_O)\left[1+\frac{H_T-H_O}{H_O}\zeta\right]^{-2/3} = C_n^2(H_O)[1+A\zeta]^{-2/3}. \tag{4.77}$$

Consider first daytime turbulence and the log-amplitude variance. Substituting Eq. (4.75) into Eq. (4.61) and rearranging terms we obtain

$$
\begin{aligned}
\sigma_\chi^2 &= 0.56\, C_n^2\left(H_O\right) k^{7/6} R^{11/6} \int_0^1 d\zeta\left[1+A\zeta\right]^{-4/3} \zeta^{5/6}\left(1-\zeta\right)^{5/6} \\
&= 0.124\, C_n^2\left(H_O\right) k^{7/6} R^{11/6} \times \\
&\quad \left[4.516\int_0^1 d\zeta\left[1+A\zeta\right]^{-4/3} \zeta^{5/6}\left(1-\zeta\right)^{5/6}\right] \qquad (4.78) \\
&= 0.124\, C_n^2\left(H_O\right) k^{7/6} R^{11/6} F_{1S}(A)
\end{aligned}
$$

Note that all of the slant path characteristics can be factored into a simple function $F_{1S}(A)$ which multiplies the horizontal path result. The path function F_{1S} is plotted in Figure 4-26. To determine the impact of a slant path on the log-amplitude variance, all the user needs to do is perform a horizontal path calculation, determine the relevant value of A, look up the value of $F_{1S}(A)$ in the figure, and multiply the horizontal path value by the value of F_{1S}. Similar calculations can be performed for each of the other turbulence parameters.

Figure 4-26. Multipliers used to convert horizontal path results into slant path results.



The daytime slant-path, transmitted-path atmospheric coherence length becomes

$$\begin{aligned}\rho_{0T} &= \left[2.91\, C_n^2\left(H_O\right) k^2 R \int_0^1 d\zeta \left[1+A\zeta\right]^{-4/3} \left(1-\zeta\right)^{5/3}\right]^{-3/5} \\ &= \left[1.09\, C_n^2\left(H_O\right) k^2 R\right]^{-3/5} \times \\ &\quad \left[2.670 \int_0^1 d\zeta \left[1+A\zeta\right]^{-4/3} \left(1-\zeta\right)^{5/3}\right]^{-3/5} \\ &= \left[1.09\, C_n^2\left(H_O\right) k^2 R\right]^{-3/5} F_{2S}\left(A\right)\end{aligned} \tag{4.79}$$

The daytime slant-path, received-path atmospheric coherence length becomes

$$\begin{aligned}\rho_{0R} &= \left[2.91\, C_n^2\left(H_O\right) k^2 R \int_0^1 d\zeta \left[1+A\zeta\right]^{-4/3} \zeta^{5/3}\right]^{-3/5} \\ &= \left[1.09\, C_n^2\left(H_O\right) k^2 R\right]^{-3/5} \times \\ &\quad \left[2.670 \int_0^1 d\zeta \left[1+A\zeta\right]^{-4/3} \zeta^{5/3}\right]^{-3/5} \\ &= \left[1.09\, C_n^2\left(H_O\right) k^2 R\right]^{-3/5} F_{3S}\left(A\right)\end{aligned} \tag{4.80}$$

The nighttime slant-path log-amplitude variance becomes

$$\begin{aligned}\sigma_\chi^2 &= 0.56\, C_n^2\left(H_O\right) k^{7/6} R^{11/6} \int_0^1 d\zeta \left[1+A\zeta\right]^{-2/3} \zeta^{5/6}\left(1-\zeta\right)^{5/6} \\ &= 0.124\, C_n^2\left(H_O\right) k^{7/6} R^{11/6} \times \\ &\quad \left[4.516 \int_0^1 d\zeta \left[1+A\zeta\right]^{-2/3} \zeta^{5/6}\left(1-\zeta\right)^{5/6}\right]\end{aligned} \tag{4.81}$$

$$= 0.124\ C_n^2(H_O)\, k^{7/6}\, R^{11/6}\, F_{4S}(A) \qquad (4.81)$$

The nighttime slant-path, transmitted path atmospheric coherence length becomes

$$\rho_{0T} = \left[2.91\, C_n^2(H_O)\, k^2 R \int_0^1 d\zeta \left[1 + A\zeta\right]^{-2/3} (1-\zeta)^{5/3}\right]^{-3/5}$$

$$= \left[1.09\, C_n^2(H_O)\, k^2 R\right]^{-3/5} \times \left[2.670 \int_0^1 d\zeta \left[1 + A\zeta\right]^{-2/3} (1-\zeta)^{5/3}\right]^{-3/5} \qquad (4.82)$$

$$= \left[1.09\, C_n^2(H_O)\, k^2 R\right]^{-3/5} F_{5S}(A)$$

while the nighttime slant-path, received-path atmospheric coherence length becomes

$$\rho_{0R} = \left[2.91\, C_n^2(H_O)\, k^2 R \int_0^1 d\zeta \left[1 + A\zeta\right]^{-2/3} \zeta^{5/3}\right]^{-3/5}$$

$$= \left[1.09\, C_n^2(H_O)\, k^2 R\right]^{-3/5} \times \left[2.670 \int_0^1 d\zeta \left[1 + A\zeta\right]^{-2/3} \zeta^{5/3}\right]^{-3/5}$$

$$= \left[1.09\, C_n^2(H_O)\, k^2 R\right]^{-3/5} F_{6S}(A) \qquad (4.83)$$

All six path functions F_{1S} through F_{6S} are plotted versus the path constant A in Figure 4-26. As anticipated even relatively small elevations of one end of the path can significantly reduce the effects of turbulence. The figure is valid for slant paths with $H_T > H_O$. The same expressions and curves can be used if $H_O > H_T$ by using the absolute value of A instead of A and using F_{2S} and F_{5S} for received paths and F_{3S} and F_{6S} for transmitted paths (instead of the other way around).

Figures 4-27 through 4-29 explicitly show the effects of slant paths on the log-amplitude variance, transmit-path atmospheric coherence length, and receive-path atmospheric coherence length at 10.6 μm and assuming daytime turbulence.

Figure 4-27. Effects of daytime slant paths on turbulence log-amplitude variance. Transceiver altitude is 2m; wavelength is 10.6 μm; $C_n^2(2m) = 10^{-13}$ $m^{-2/3}$.

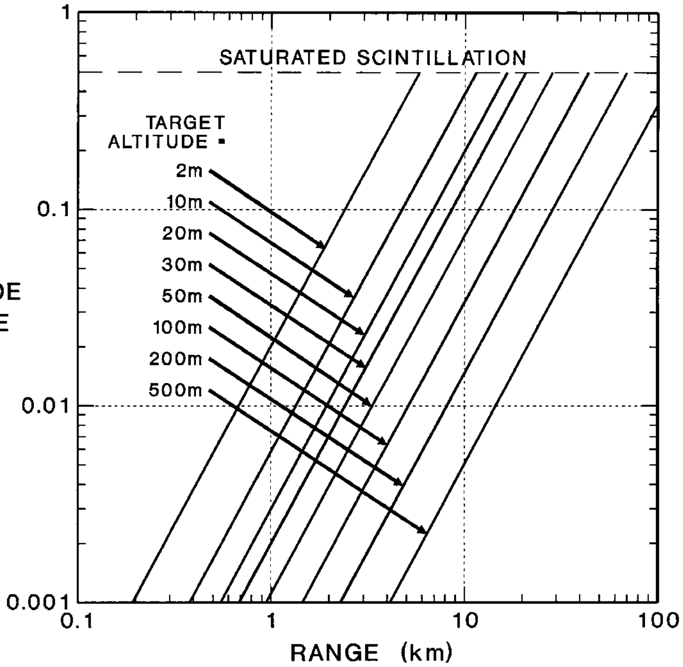


Figure 4-28. Effects of daytime slant paths on transmitted-path atmospheric coherence length. Transceiver altitude is 2m; wavelength is 10.6 μm; $C_n^2(2m) = 10^{-13}$ $m^{-2/3}$.

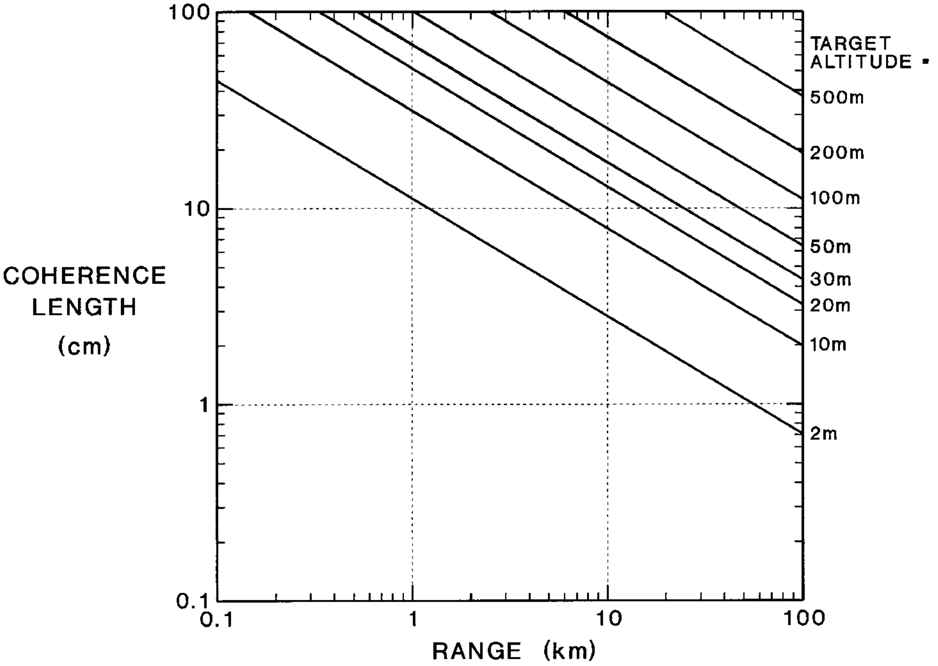
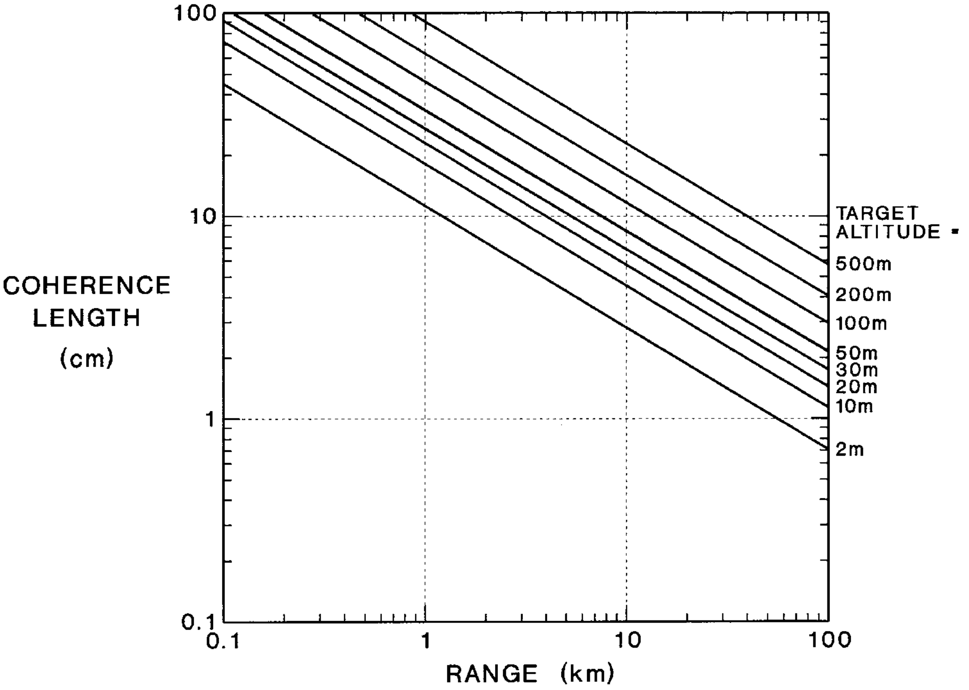


Figure 4-29. Effects of daytime slant paths on received-path atmospheric coherence length. Transceiver altitude is 2m; wavelength is 10.6 μm; $C_n^2(2m) = 10^{-13}\ m^{-2/3}$.



Ionospheric Effects

In the thermosphere, not only does the kinetic temperature of the molecules rise dramatically, the degree of ionization increases (although the ionization is not produced by the temperature rise). Solar x-radiation from solar prominences and flares and extreme ultraviolet radiation, blackbody radiation plus strong photospheric emission lines such as Lyman-α (121.5 nm) and Lyman-β (102.5 nm) from hydrogen atoms as well as lines from ionized helium in the solar photosphere, is absorbed by photoelectric emission from the oxygen and nitrogen atoms and molecules in the upper atmosphere. The resulting ionization, establishing a thick layer of the upper atmosphere called the ionosphere, significantly effects low frequency radio wave propagation.[14]-[17]

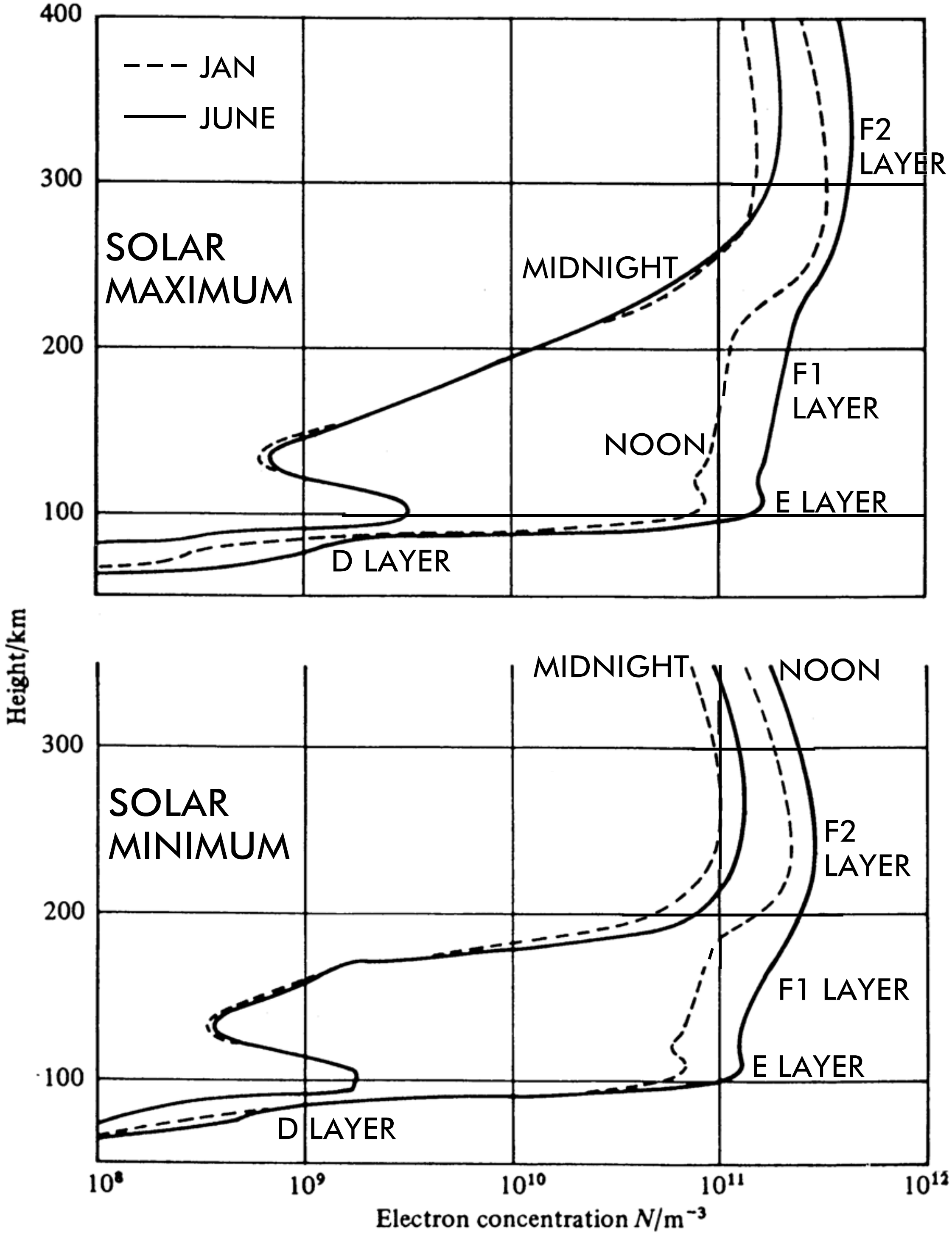
The ionosphere is characteristically divided into four significant layers: the D layer, the E layer, the F1 layer, and the F2 layer. Obviously from the nomenclature there are several other layers of minimal significance. The vertical structure of these layers is illustrated in Figure 4-30.[15] There is considerable variation of these layers with time. As the ionosphere results from direct ionization produced by sunlight, we would expect and indeed find that daytime ionization levels are higher than nighttime levels. There is also a weaker dependence on time of year (varying distance from the sun and solar elevation angle act to produce seasonal variations in ionization just as they produce similar changes in the weather in the troposhpere. Since there is a substantial contribution to the ionization from solar x-rays, this produces an expected variation of ionization with position relative to the 11-year sunspot cycle. At solar maximum (maximum sunspots and prominences) there is higher overall ionization produced than at solar minimum. All three variations (diurnal, annual, and sunspot cycle are shown in Figure 4-30.

The **E layer** is a narrowly layer at approximately 110 km altitude that appears to be dominantly produced by direct ionization of molecular oxygen by ultraviolet radiation with wavelengths below 103 nm. Secondary contributions come from ionization of nitric oxide (NO). The E layer usually nearly diminishes greatly at nighttime. It is still apparent in ionization profiles but its peak value is 1% of its daytime levels. Occasionally a nighttime E layer appears, when it is correlated with ionization produced by meteorites or electron showers.

The **D layer** lies below the E layer and is formed by x-ray and Lyman-α ionization of nitric oxide (NO). It is a weak layer that almost disappears at night. Its ionization is so low that it seldom affects propagation for producing an anomalous absorption of low-frequency radiation (below 10 MHz).

The F layers occur above the E layer and can extend to heights of several hundred kilometers. Electron production in the F layers appears to result from ultraviolet ionization of atomic oxygen although ionization of molecular and atomic nitrogen may also contribute. The **F1 layer** seldom appears as a distinct peak. It is usually a shoulder or a shelf below the dominant **F2 layer**. In the F1 layer, recombination of electrons with ions and attachment to neutrals are the mechanisms that counter the production rate (and keep the ionization finite). As altitude increases, the collisions necessary for “electron loss” become increasingly rare. This causes the F2 layer in which upward drift of electrons out of the layer balances the rate of production.

Figure 4-30. The vertical structure of the ionosphere showing diurnal, annual, and 11-year solar sunspot cycle variations in electron concentration.



The primary significance of the ionosphere on radio wave propagation comes from its refractive index. The refractive index of an electron-ion plasma is given by the equation

$$n = \left[1 - \left(\frac{\nu_p}{\nu} \right)^2 \right]^{1/2} \tag{4.84}$$

where ν is the frequency of the electromagnetic radiation and ν_p is the electron plasma frequency. The **plasma frequency** (in Hz) is given by

$$\nu_p = \left(\frac{N_e \mathrm{e}^2}{4\pi^2 m \varepsilon_0} \right)^{1/2} = 8.98 N_e^{1/2} \tag{4.85}$$

where N_e is the density (electrons per cubic meter) of free electrons in the plasma. Note that an electron density of 10^{12} m^{-3} produces a plasma frequency of roughly 9 MHz while an electron density of 10^{10} m^{-3} produces a plasma frequency of roughly 900 kHz.

At very high frequencies *n* is approximately 1 and the ionosphere can have little effect. However, as ν approaches ν_p , the refractive index starts to drop rapidly. When $\nu < \nu_p$, then *n* is pure imaginary, and the wave cannot propagate. It is completely reflected. Consider an ionosphere in which the electron density is zero at some altitude h_0 and increases steadily to a maximum value N_{max} at an altitude of h_{max}. Now consider an electromagnetic wave with a frequency less than the plasma frequency at the maximum ionization point. Let this wave be radiated at some angle to the vertical. As the wave propagates outward below the ionosphere, it experiences only the propagation phenomena normally present in the troposphere (which we will assume are negligible). As the wave encounters the ionosphere, it encounters a medium with a negative index gradient (as the electron density increases, the refractive index decreases). This is a superrefractive gradient. The wave will begin to bend away from the vertical. As the wave penetrates farther into the ionosphere, it experiences an ever-increasing superrefractive effect. Depending on its initial angle of incidence and its frequency its path may be sufficiently curved to be bent back towards the surface. It is essentially "reflected" from the ionosphere (even though reflection has no real part in the process). If it fails to be reflected by the first layer it will continue to propagate deeper into the ionosphere encountering higher layers with denser ionization and higher superrefractivity. Any ray with frequency below the plasma frequency corresponding to the peak ionization (a frequency we will call the **cut-off frequency**) will ultimately be refracted back towards the surface. The height at which it is so "reflected" will be a function of incidence angle, frequency, and ionospheric state.

If we consider a fixed ionosphere and select a frequency below the cutoff frequency and propagate rays toward the ionosphere with increasing elevation angle we observe behavior similar to that in Figure 4-31. Low elevation rays (ray 5) reflect off lower portions of the ionosphere and continue back to the surface striking the ground at great distances determined by the altitude of reflection and the real spherical geometry. The maximum distance will be approximately twice the "horizon range" as viewed from the effective altitude of the reflection (the height of the transmitter has little effect on the maximum range for realistic transmitter heights). Thus, using Eqs (4.23) and (4.24) we have

$$R_{MAX} = 260.7\ H^{1/2} \quad \text{km} \qquad\qquad (H \text{ in km}) \tag{4.86}$$

or

$$R_{MAX} = 191.7\ H^{1/2} \quad \text{nmi} \qquad\qquad (H \text{ in nmi})\ . \tag{4.87}$$

For reflection at the E layer (roughly 90 km high) we have a maximum range per bounce of about 2500 km. For reflection from the F2 layer (perhaps 200 km high) we have a maximum range per bounce of about 3700 km.

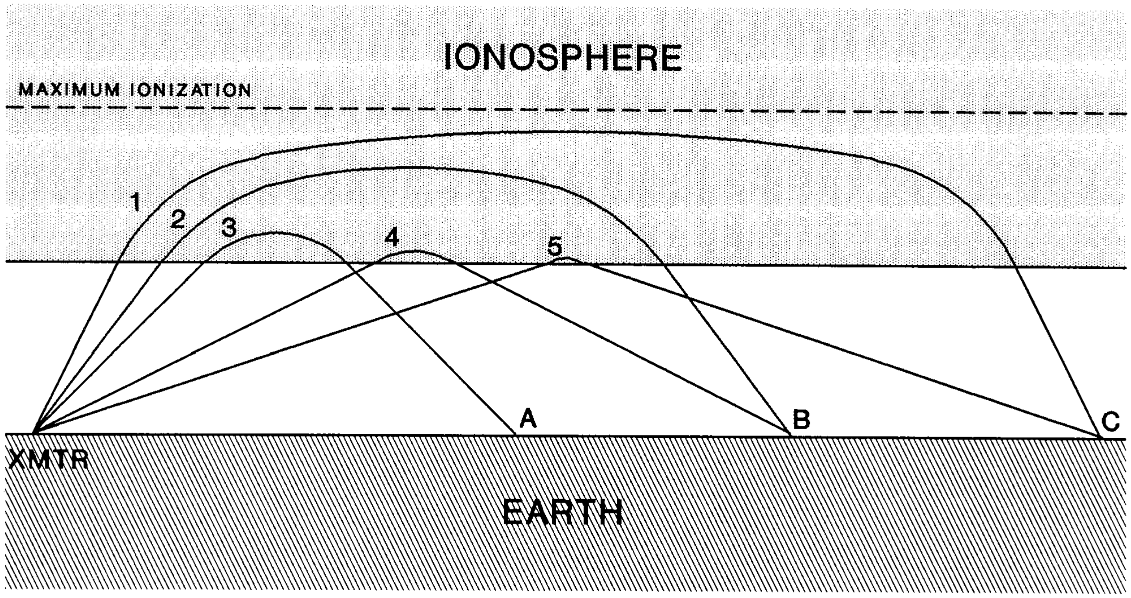
As the elevation angle of the ray increases (ray 4) the path to the ground decreases steadily. However, at some elevation angle (ray 3) corresponding to a ground bounce distance of about 800 km, the distance stops decreasing with increasing elevation angle. Rays with higher elevation angles (rays 1 and 2) will pass through the first layer and encounter regions of weaker bending that cause them to travel larger horizontal distances before ultimately reflecting off of a higher layer. Thus there is a minimum bounce distance of about 800 km. The presence of two ray paths to the same point on the ground (e.g., ray 5 and ray 1) could conceptually lead to interference, but given the wide angular separation this is not likely to be a problem for sensor performance. Only frequencies near to the cutoff frequency would not likely have a minimum bounce distance.

Frequencies above the cutoff frequency can be reflected as well but only at lower elevation angles. The higher the frequency the weaker the ionospheric refraction. Higher frequency rays traveling predominantly upward would not be sufficiently refracted before they reached the height of maximum ionization at which point the medium become subrefractive and the rays can escape. Those at lower elevation angles need only smaller changes in curvature for them to be returned to the surface. The maximum frequency at which ionospheric bounces will occur is in the range of 30-50 MHz (well above the cutoff frequency of roughly 10 MHz for normal peak ionization levels). It has been found that the "**effective plasma frequency**" for rays of low elevation angle is given by

$$\nu_{peff} = \left(\frac{N_e \mathrm{e}^2}{4\pi^2 m \varepsilon_0} \right)^{1/2} \frac{1}{\cos\alpha} = \frac{8.98 N_e^{1/2}}{\cos\alpha} \tag{4.88}$$

where α is the angle of the ray with respect to the zenith (90° minus the elevation angle). When the real frequency equals the effective plasma frequency the angled ray will be reflected.

Figure 4-31. Variation in ray paths as a function of elevation angle in ionospheric reflection.[17]



The effective height of reflection depends on the ionization profile as well as angle and frequency. As the ionosphere undergoes it diurnal variation, the effective height of reflection and thus the effective range of the bounce will vary up and down. Systems using ionospheric bounce channels are usually designed to vary the operating frequency to keep the bounce characteristics relatively independent of time of day.

In general most ionospheric effects are refractive in character. There is, however, some potential for absorption. The absorption in a weakly ionized plasma can be determined from the equation

$$\alpha = \left(\mathrm{e}^2 / 2\varepsilon_0 mc\right)\left(N_e \nu_c\right)\left[\left(2\pi\nu\right)^2 + \nu_c^2\right]^{-1} \tag{4.89}$$

where N_e is the electron density (in electrons/m^3) and ν_c is the collision frequency between electrons and neutrals (in Hz). Evaluating the constants yields

$$\begin{aligned} \alpha &= 4.6 \times 10^{-2}\left(N_e \nu_c\right)\left[\left(2\pi\nu\right)^2 + \nu_c^2\right]^{-1} \qquad \mathrm{dB/km} \\ &\approx 1.16 \times 10^{-3}\ N_e \nu_c / \nu^2 \qquad \mathrm{dB/km} \end{aligned} \tag{4.90}$$

Only in the lower layer of the ionosphere (D layer) is the collision frequency high enough to give any appreciable absorption. Typical values for the collision frequency in the 60-90 km altitude regime are 10^8 to 10^6 per second. Given electron densities of 10^8 to 10^{10} m^{-3} in this same region, we find $N_e \nu_c$ products in the range of 10^{16} to 10^{17} m^{-3}-s^{-1}. For radio frequencies in the range of 3 to 10 MHz, one can readily calculate attenuations ranging from an inconsequential 0.1 dB/km to a substantial 12.9 dB/km. Because of the inverse square dependence on frequency, absorption will be negligible at frequencies significantly greater than 10 MHz. Even though electron densities are much higher in the E and F layers, the collision frequency is many of magnitude lower (the frequency scales with atmospheric pressure which falls off roughly one order of magnitude for each 10 km increase in altitude above 70 km), preventing any substantial absorption.

A final refractive effect can occur in the ionosphere. On small scales (kilometers) the ionosphere is inhomogeneous. There are currents and eddies, just as there are currents and eddies in the atmosphere. The refractivity variations associated with these eddies produces scintillation at microwave frequencies just as atmospheric turbulence produces scintillation in the visible and infrared. Although there are no good predictive tools for ionospheric scintillation, there are a few useful observations. Scintillation appears most severe in early evening but can be noticeable at any time. It is more severe in equatorial regions than in polar regions. Scintillation is only seen at night in polar regions. Scintillation amplitude decreases with increasing frequency but is still noticeable at gigahertz frequencies. Scintillation fades can exceed 20 dB. The characteristic time scale of fading tends to be relatively slow (tens of seconds). As with turbulence, additional carrier-to-noise ratio margin may be required to overcome the effects of scintillation fades on detection probability.

References

[1] Edlen, K., “The Refractive Index of Air”, *Metrologia*, 2, 12 (1966).

[2] Eaves, Jerry L. and Reedy, Edward K., Principles of Modern Radar (Van Nostrand Reinhold, New York NY, 1987).

[3] Kerr, Donald E., Propagation of Short Radio Waves (McGraw-Hill Book Co, New York NY, 1951), pp. 41-58.

[4] McCartney, Earl J., Optics of the Atmosphere (John Wiley & Sons, New York NY, 1976) pp. 94-100.

[5] Naval Ocean Systems Center, EREPS: Engineer’s Refractive Effects Prediction System Software and User’s Manual (Artech House, Norwood MA, 1990).

[6] Davis, Christopher C., Lasers and Electro-Optics (Cambridge University Press, Cambridge UK, 1996), pp. 342-345.

[7] Born, Max and Wolf, Emil, Principles of Optics 6th Ed. (Pergamon Press, Oxford UK, 1980).

[8] Greenler, Robert, Rainbows, Halos, and Glories (Cambridge University Press, Cambridge UK, 1989).

[9] Spencer, J. L., “Long-Term Statistics of Atmospheric Turbulence Near the Ground”, RADC-TR-78-182, Rome Air Development Center (August 1978).

[10] Shapiro, J. H., Capron, B. A., and Harney, R. C., “Imaging and target detection with a heterodyne-reception optical radar”, *Applied Optics*, 20, #19, 3292-3313 (1 October 1981).

[11] Wyngaard, J. C., Izumi, Y., and Collins, Jr., S. A., “Behavior of the Refractive-Index-Structure Parameter Near the Ground”, *J. Opt. Soc. America*, 61, #12, 1646-1650 (December 1971).

[12] Hufnagel, Robert E., “Variations of Atmospheric Turbulence” in *Digest of Technical Papers,* Topical Meeting on Optical Propagation Through Atmospheric Turbulence, Optical Society of America, Washington DC, July 9-11, 1974.

[13] Fried, D. L., “Statistics of a Geometric Representation of Wavefront Distortion”, *J. Opt. Soc. Amer.*, 55, #11, 1427-1435 (November 1965).

[14] Goodman, John M., HF Communications: Science and Technology (Van Nostrand Reinhold, New York NY, 1992).

[15] Budden, K. G., The Propagation of Radio Waves (Cambridge University Press, Cambridge UK, 1985).

[16] Tascione, Thomas F., Introduction to the Space Environment (Orbit Book Co., Malabar FL, 1988).

[17] McNamara, Leo F., The Ionosphere: Communications, Surveillance, and Direction Finding (Krieger Book Co., Malabar FL, 1991).

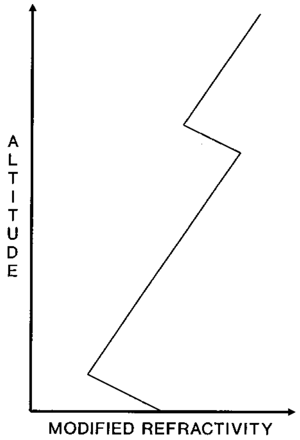
Problems

4-1. An incoming cruise missile has a closure velocity of 1.4 km/sec and an altitude of 2 m above mean sea level. An interceptor missile is to be launched to intercept the cruise missile at 5 km range. The fire control system requires 12 seconds to detect and establish track on the incoming missile. Administrative delays involving the decision to engage require at least 6 seconds. Another 2 seconds is required to prepare and actually launch the interceptor. The interceptor requires 5 seconds to fly out to the intercept range. What is the minimum acceptable height of the search radar in order to successfully accomplish this intercept?

4-2. What is ducting? What condition causes it? Note: answer desired is not specific to any type of radiation or wavelength region.

4-3. What type of duct always occurs over bodies of water? At what heights does the top of these ducts typically occur?

4-4. Given the following curve of modified refractivity vs. height identify the altitude regions having ducting. What kinds of ducts are present?



4-5. Describe the kinds of atmospheric ducts that might occur over land and outline the effects they might have on microwave sensors.

4-6. What is the dominant source of atmospheric turbulence? Describe briefly how it produces turbulence.

4-7. Quantify the two primary differences between nighttime and daytime turbulence.

4-8. Name three major effects of turbulence on sensor performance.

4-9. What three parameters determine the magnitude of all turbulence effects?

4-10. Which of the following are prominent refraction effects?
Mirages? ducting? beam wander? extinction? scintillation? beam spread?

4-11. What turbulence effects are quantified by:
- the log-amplitude variance?
- the atmospheric correlation length?

4-12. For HF radiation with a frequency of 2 MHz and using the data of Figure 4-29, estimate the heights at which reflection occurs for radiation initially radiated at 45° to the zenith. Calculate results for noon and midnight and solar maximum and solar minimum.

4-13. Using the heights calculated in the preceding problem, estimate the maximum ranges that can be obtained on a single bounce or skip.

CHAPTER 5

PROPAGATION OF ELECTROMAGNETIC RADIATION. IV. – OTHER ATMOSPHERIC AND UNDERWATER EFFECTS

Contrast Transmission

The observable quantity used by most visible and infrared imaging sensors is the target-to-background contrast. **Contrast** can be defined in several different ways:[1]

$$C = \frac{I_T - I_B}{I_B} , \tag{5.1a}$$

$$C = \frac{I_T - I_B}{I_T + I_B} , \tag{5.1b}$$

or

$$C = \frac{I_T - I_B}{I_T} , \tag{5.1c}$$

where I_T is the signal strength observed from the target and I_B is the signal strength observed from the nearby background. In the remainder of this work we will use Equation (5.1a) unless the conventions of a specific application requires otherwise. Exceptions will be explicitly identified.

The **inherent contrast** of an object is defined as the contrast of that object at zero range and is given by [1],[2]

$$C(0) = \frac{I_T(0) - I_B(0)}{I_B(0)} . \tag{5.2}$$

The signals from both target and background will be attenuated by the intervening atmosphere. They will also be augmented by a **path radiance** $I_P(R)$ that include atmospheric emission and radiation scattered into the line of sight by haze, fog, smoke, or other obscurants. Thus,

$$I_T(R) = I_T(0)e^{-\alpha R} + I_P(R) \tag{5.3}$$

and

$$I_B(R) = I_B(0)e^{-\alpha R} + I_P(R) . \tag{5.4}$$

Using these expressions the **apparent contrast** observed at a distance *R* becomes

$$
\begin{aligned}
C(R) &= \frac{I_T(R) - I_B(R)}{I_B(R)} = \frac{I_T(0)e^{-\alpha R} - I_B(0)e^{-\alpha R}}{I_B(0)e^{-\alpha R} + I_P(R)} \\
&= \frac{I_T(0) - I_B(0)}{I_B(0) + I_P(R)e^{+\alpha R}} \\
&= \frac{I_T(0) - I_B(0)}{I_B(0)} \frac{1}{1 + \left[I_P(R) / I_B(0)\right]e^{+\alpha R}} \qquad (5.5) \\
&= \frac{C(0)}{1 + \left[I_P(R) / I_B(0)\right]e^{+\alpha R}}
\end{aligned}
$$

The ratio of the apparent contrast to the inherent contrast is known as the **contrast transmittance**

$$\frac{C(R)}{C(0)} = \frac{1}{1 + \left[I_P(R) / I_B(0)\right]e^{+\alpha R}} . \qquad (5.6)$$

If the attenuation is very large, or if the path radiance is very bright (as might be caused by a highly scattering smoke cloud between the observer and the target), or both, then Eq. (5.6) can be simplified to

$$\frac{C(R)}{C(0)} = \frac{I_B(0)}{I_P(R)} e^{-\alpha R} . \qquad (5.7)$$

We will use this expression in the next section to evaluate the performance of artificial obscurants.

The path radiance becomes larger as the path increases up to a limit. At some point the attenuation of the atmosphere will start degrading the path radiance as fast as that radiance grows, and a limiting value called the **path radiance limit** I_L. The path radiance and the path radiance limit are related by the expression

$$I_P(R) = I_L\left(1 - e^{-\alpha R}\right) . \qquad (5.8)$$

Substituting for the path radiance in Eq. (5.6) yields

$$\frac{C(R)}{C(0)} = \frac{1}{1+\left[I_L / I_B(0)\right]\left[e^{+\alpha R}-1\right]} \tag{5.9}$$

where the quantity I_L/I_B is called the **sky-to-ground ratio**. Typical values of I_L/I_B are in the range of 2 to 5 [3] (although as indicated in Table 5-1 the values can be very widespread under extreme conditions [2]). Now consider the case where the total attenuation is substantial. Then $e^{\alpha R} >> 1$ and Equation (5.9) can be simplified to

$$\frac{C(R)}{C(0)} = \frac{e^{-\alpha R}}{\left[I_L / I_B(0)\right]} \tag{5.10}$$

that is, the contrast transmittance is proportional to the actual transmittance ($e^{-\alpha R}$).

Now consider the case where the target is a black target so that $C(0)$ is -1 independent of the brightness of the background and now assume that the target is viewed against the sky background. This is a fairly accurate method of estimating meteorological visibility. Given our assumptions, we have $I_L = I_B(0)$ and

$$\frac{C(R)}{C(0)} = e^{-\alpha R} \tag{5.11}$$

Now if we assume a 2% contrast is required for detection and that the range at which detection is barely possible is called the visibility V (refer back to Chapter 3). Then we may solve Eq. (5.11) to obtain Koschmieder's formula

$$\frac{C(R)}{C(0)} = 0.02 = e^{-\alpha V} \qquad \Rightarrow \qquad \alpha = \frac{3.912}{V} \ . \tag{5.12}$$

Table 5-1. Typical sky-to-ground ratios. [2]

SKY CONDITION	GROUND CONDITION	SKY-TO-GROUND RATIO
CLEAR	FRESH SNOW	0.2
CLEAR	DESERT	1.4
CLEAR	FOREST	5
OVERCAST	FRESH SNOW	1
OVERCAST	DESERT	7
OVERCAST	FOREST	25

Smoke, Dust, and Battlefield Obscurants

Obscurants are aerosols other than fog or haze, that degrade visibility to unacceptably low levels. In the form of smoke from fires and/or weapon discharges, and dust from explosions and/or vehicular traffic, obscurants have contributed to the "fog of war" ever since the beginnings of organized warfare. In the twentieth century, the development of artificial obscurants, the so-called "**military smokes**", has made it possible to affect and control the visibility almost anywhere in the battlespace.

Obscurants can affect sensor performance in four ways: absorption, scattering, turbulence, and cloud radiance. Some obscurants absorb electromagnetic radiation. The **absorption** reduces the signal strength available for detection or illumination. Since obscurants can never be perfectly uniformly distributed in space (they are generated at discrete points), the inhomogeneous absorption will lead to low frequency spatial and temporal modulation of the measured signal. All obscurants will scatter electromagnetic radiation. Just like absorption, **scattering** will reduce the signal strength available for detection or illumination, and can lead to low frequency spatial and temporal modulation of signals. The scattering also leads to spreading of electromagnetic beams and diffusion of image data. It can lead to temporal dispersion (lengthening) of short electromagnetic pulses (although this is usually an insignificant effect). Scattering can also lead to depolarization of electromagnetic radiation. Many obscurant production systems involve heat sources, such as pyrotechnic smoke canisters, spraying oil on hot engine exhaust manifolds, or incendiary grenades. As a consequence of the non-uniform introduction of thermal energy into the atmosphere, **turbulence** may be enhanced. As will be discussed in more detail later in this chapter, turbulence can cause random steering and spreading of electromagnetic beams, introduction of amplitude fluctuations (scintillation) into signals, and distortion and/or diffusion of images. The heat retained by the obscurant cloud will result in a certain amount of **cloud radiance**. Usually this radiance will be in the infrared, although some smokes (e.g., burning phosphorus) may emit significant amounts of visible radiation for a brief period of time. An additional radiance contribution may result from the scattering of ambient radiation (sunlight or earthshine) into a sensor's line of sight. The next effects of cloud radiance are reduced contrast, increased shot noise, and possible low frequency spatial and temporal modulation of background radiance.

Smoke and dust will be produced with some degree of randomness on the battlefield. Artificial military aerosols (battlefield "smokes") are dispensed with the intent to establish specific useful conditions. Their use is neither random nor unpredictable. **Screening smoke** has four primary modes of application. A **smoke blanket** is a dense smoke concentration established over a friendly area to prevent enemy observation and precision aerial bombing [4]. It has the drawback of significantly restricting friendly activity and movement within the blanket. The attenuation along any path to the exterior of the smoke blanket (including straight up) must exceed 17 dB (remember the definition of visibility in Chapter 3). This will require the visibility inside the smoke blanket to be of the order of 50-100 m, equivalent to a thick fog. The self-protection smoke produced by the smoke grenades on armored vehicles is a limited application of a smoke blanket. A **smoke haze** is a light smoke concentration established over a friendly area to reduce enemy visual observation by means other than direct overflight. Resulting visibility is typically 135-180 m, which is sufficient to permit near normal friendly movement and activity. A **smoke curtain** is a vertical screen established between friendly and enemy forces to prevent any direct observation of the activities of either side. Minimum useful one-way attenuation for a smoke curtain is 17 dB. **Obscuring smoke** is a dense smoke concentration established over enemy positions to restrict enemy observation and operations. Resulting visibility should be less than or equal to 50 m (equivalent to a dense fog), which is sufficient to slow vehicular traffic to a crawl, prevent visual coordination between adjacent units, and make it difficult to correlate position against map references.

The transmission of smoke, dust, or military aerosol clouds is given by the expression

$$T = e^{-\alpha \int_0^L dx\, C(x)} \tag{5.13}$$

where α is the mass extinction coefficient (with units of m^2/g), C is the aerosol concentration (in g/m^3), and L is the optical path length through the aerosol cloud. For simplicity of use, aerosol clouds are often characterized by a path-averaged concentration-path length product, *CL*, in which case the transmission is given by

$$T = e^{-\alpha CL} \; . \tag{5.14}$$

An αCL product of 4 yields $T=0.018$ or 17.4 dB attenuation. Thus, $\alpha CL=4$ provides an attenuation comparable to that used to define visibility. As a result this value can be used to define the attenuation of an effective smoke curtain.

Table 5-2 summarizes data on a number of common battlefield aerosols.[1],[5],[6] Included in the table are military nomenclature if applicable, the generic composition of the aerosols, typical aerosol size, and mass extinction rations at a number of wavelengths. The unfortunate lack of open source data leads to blanks in the columns of a few aerosol types. Table 5-3 summarizes the effects of relative humidity on the obscuration characteristics of two of the most important artillery delivered smokes – WP (white phosphorus) and HC (Zn/Hexachloroethane).[6] In general higher relative humidity leads to more extraction of water vapor from the air, with a higher yield ratio (mass of smoke particles per mass of initial smoke compound) and larger particle sizes. The latter favors increased infrared attenuation, but also favors faster settling of the smoke particles out of suspension in the air. Table 5-4 summarizes the obscuration characteristics of various kinds of dust and blowing snow, two natural obscurants that are common generated by the action of weapons on the soil or by the movement of vehicles or helicopters.[6]

Useful insights can be obtained by comparing the attenuation performance of several common military smokes. Figure 5-1 compares the attenuation versus concentration-path length product for FS, HC, WP, and two kinds of fog oil (A – explosively dispersed and B – generated by dripping oil onto a hot plate). The attenuations in the visible (0.63 μm) are compared with those in

Table 5-2. Extinction data on a variety of battlefield aerosols.[1],[5],[6]

TYPE	NOMEN-CLATURE	APPLICABLE MIL-SPEC	AEROSOL COMPOSITION	MEDIAN AEROSOL DIAMETER (um)	MASS EXTINCTION COEFFICIENT (m^2/g)								
					WAVELENGTH (um)								
					0.4 - 0.7	0.63	0.7 - 1.2	0.9	1.06	2.06	3 - 5	8 - 12	10.6
PHOSPHOROUS	RP	MIL-P-211	$H_3PO_4 \cdot nH_2O$	1.1	-	1.8	-	-	1.5	-	-	-	0.33
	WP	MIL-W-215		-	2.98	-	2.46	-	2.25	-	0.48	0.27	0.26
	PWP	MIL-C-00337											
HEXACHLOROETHANE/ ZnO/Al	HC	MIL-H-235	$ZnCl_2 \cdot nH_2O$	1.3	-	1.4	-	-	1.5	-	-	-	0.12
				-	1.81	-	1.98	-	1.93	-	0.31	0.07	0.09
SULFUR TRIOXIDE/ CHLOROSULFONIC ACID	FS	MIL-C-379 OBSOLETE	$H_2SO_4 \cdot nH_2O$	0.85	-	4.5	-	3.1	2.5	0.4	-	-	0.20
TITANIUM TETRACHLORIDE	FM	MIL-C-357	$Ti(OH)_4 + HCl \cdot nH_2O$	-	-	-	-	-	-	-	-	-	-
FOG OIL - TYPE SGF-2	FOG OIL	MIL-F-12070	OIL DROPLETS	-	6.85	-	4.59	-	3.48	-	0.25	0.02	0.02
- PYROTECHNIC SOURCE				0.6	-	4.2	-	4.6	3.1	1.1	-	-	0.03
- HOT PLATE SOURCE				3.4	-	2.9	-	2.5	2.2	1.4	-	-	0.20
DIESEL OIL	?	-	OIL DROPLETS	-	6.40	-	3.69	-	2.94	-	1.34	1.00	1.00
ANTHRACENE	?	?	CRYSTALS	-	6.00	-	3.50	-	2.00	-	0.23	0.06	0.05
YERSHOV COMPOSITION	?	-	$NH_4Cl \cdot nH_2O$	-	-	-	-	-	-	-	-	-	-
CARBON													
- SOOT	-	-	PARTICLES	-	1.50	-	1.46	-	1.42	-	0.75	0.32	0.30
- GRAPHITE	?	?	FLAKES	-	-	-	-	-	-	-	-	-	-
BURNING DIESEL FUEL	-	-	SOOT & HYDROCARBONS	-	6.40	-	3.69	-	2.94	-	1.34	1.00	1.00

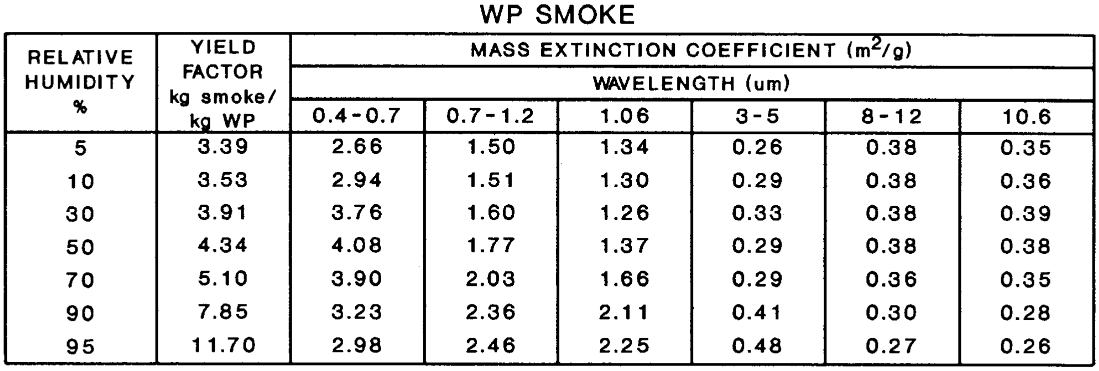
the infrared (10.6 μm). Recall that an attenuation of 17 dB ($\alpha CL = 4$) defines an effective smoke "thickness" in the visible. If a marginally visually effective smoke cloud is produced for any of these smokes, then it is readily apparent that the attenuation in the infrared will be negligible (approximately 1-2 dB). Because the mass extinction coefficients are much smaller, effective obscuration in the infrared requires *CL* products an order of magnitude larger than necessary to produce effective obscuration in the visible.

Many military obscurants have much larger scattering coefficients than absorption coefficients. These materials produce "white" smokes. However, some obscurants are dominated by absorption rather than scattering. These materials produce what are known as "black" smokes. Their effects are somewhat different. Consider the contrast transmittance

$$\frac{C(R)}{C(0)} = \frac{I_B(0)}{I_P(R)} e^{-\alpha R} \tag{5.15}$$

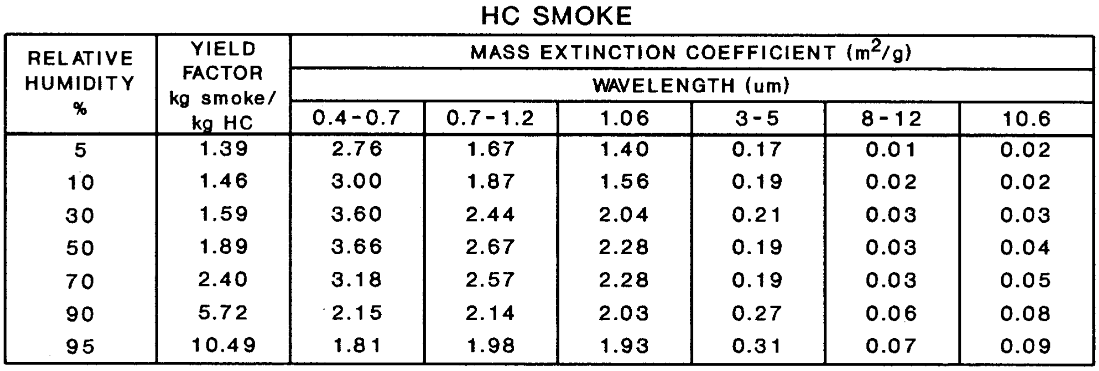
If the path radiance is due primarily to a smoke cloud between the observer and the target, we may

Table 5-3. Relative humidity effects on the extinction of WP and HC aerosols.[6]



WP SMOKE

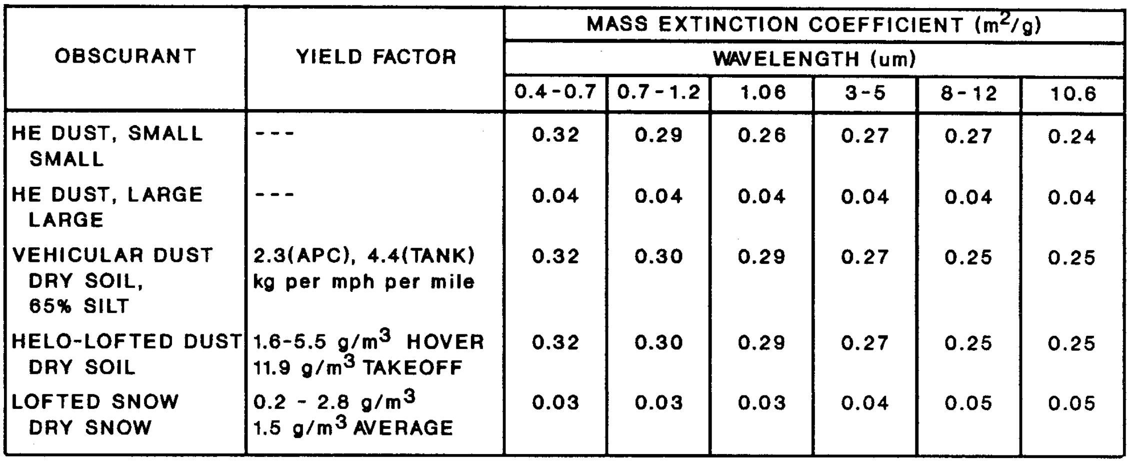
RELATIVE HUMIDITY %	YIELD FACTOR kg smoke/ kg WP	MASS EXTINCTION COEFFICIENT (m^2/g)					
		WAVELENGTH (um)					
		0.4-0.7	0.7-1.2	1.06	3-5	8-12	10.6
5	3.39	2.66	1.50	1.34	0.26	0.38	0.35
10	3.53	2.94	1.51	1.30	0.29	0.38	0.36
30	3.91	3.76	1.60	1.26	0.33	0.38	0.39
50	4.34	4.08	1.77	1.37	0.29	0.38	0.38
70	5.10	3.90	2.03	1.66	0.29	0.36	0.35
90	7.85	3.23	2.36	2.11	0.41	0.30	0.28
95	11.70	2.98	2.46	2.25	0.48	0.27	0.26



HC SMOKE

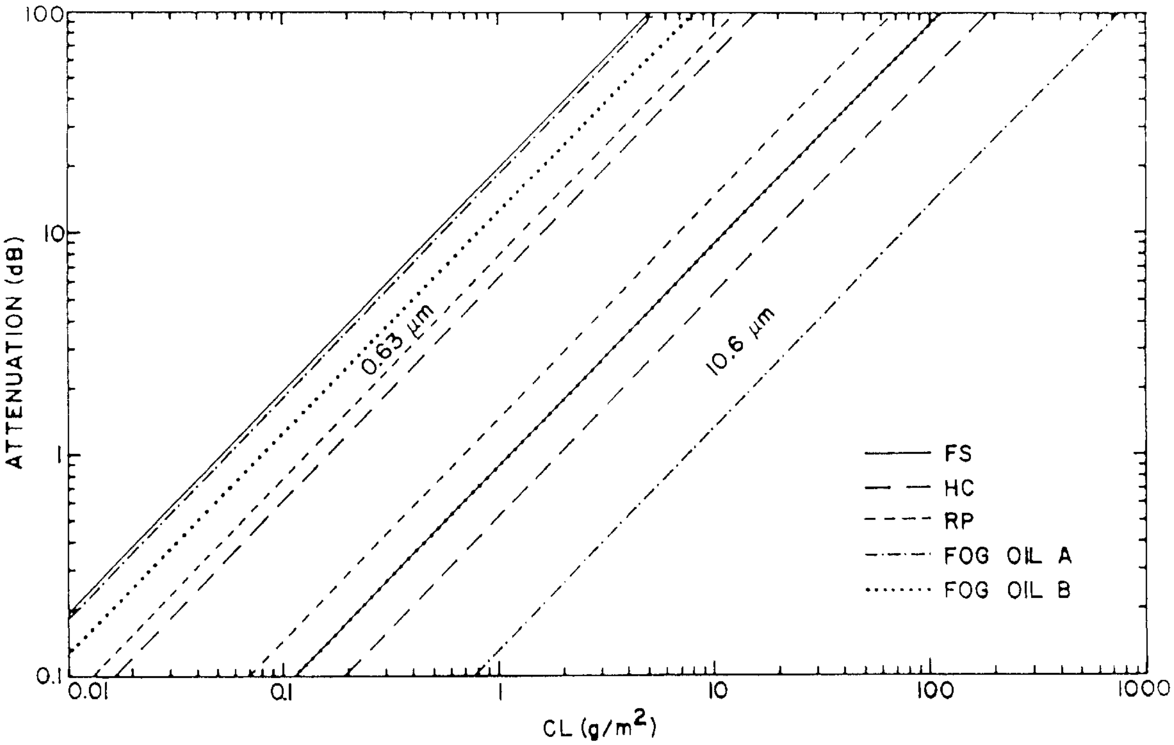
RELATIVE HUMIDITY %	YIELD FACTOR kg smoke/ kg HC	MASS EXTINCTION COEFFICIENT (m^2/g)					
		WAVELENGTH (um)					
		0.4-0.7	0.7-1.2	1.06	3-5	8-12	10.6
5	1.39	2.76	1.67	1.40	0.17	0.01	0.02
10	1.46	3.00	1.87	1.56	0.19	0.02	0.02
30	1.59	3.60	2.44	2.04	0.21	0.03	0.03
50	1.89	3.66	2.67	2.28	0.19	0.03	0.04
70	2.40	3.18	2.57	2.28	0.19	0.03	0.05
90	5.72	2.15	2.14	2.03	0.27	0.06	0.08
95	10.49	1.81	1.98	1.93	0.31	0.07	0.09

Table 5-4. Mass extinction coefficient of lofted dust and snow particles.[6]



OBSCURANT	YIELD FACTOR	MASS EXTINCTION COEFFICIENT (m^2/g)					
		WAVELENGTH (um)					
		0.4-0.7	0.7-1.2	1.06	3-5	8-12	10.6
HE DUST, SMALL SMALL	---	0.32	0.29	0.26	0.27	0.27	0.24
HE DUST, LARGE LARGE	---	0.04	0.04	0.04	0.04	0.04	0.04
VEHICULAR DUST DRY SOIL, 65% SILT	2.3(APC), 4.4(TANK) kg per mph per mile	0.32	0.30	0.29	0.27	0.25	0.25
HELO-LOFTED DUST DRY SOIL	1.6-5.5 g/m^3 HOVER 11.9 g/m^3 TAKEOFF	0.32	0.30	0.29	0.27	0.25	0.25
LOFTED SNOW DRY SNOW	0.2 - 2.8 g/m^3 1.5 g/m^3 AVERAGE	0.03	0.03	0.03	0.04	0.05	0.05

Figure 5-1. Comparison of visible and infrared attenuation produced by military smokes.



estimate the path radiance given knowledge of the external illumination. We will represent the total attenuation coefficient α by the sum of the absorption coefficient α_A and the scattering coefficient α_S. If I_S is the external illumination due to sunlight, moonlight, skyshine (scattered sunlight), earthshine (scattered sunlight and thermal emission), or other radiation source, then the fraction of that illumination scattered per unit length of obscurant is given by

$$dI_O = -dI_S = I_S \alpha_S dx \tag{5.16}$$

The obscurant nominally scatters its radiation into 4π sr. There is an angular dependence to the scattering that should be taken into account, however, we will simply denote this dependence by a generic angle dependent efficiency $f(\theta)/4\pi$. If we then sum all of the scattered contributions, we obtain the overall scattered signal as

$$I_O = I_S\left(1 - e^{-\alpha_S L}\right) f(\theta) / 4\pi \ . \tag{5.17}$$

This term can be substituted into Eq. (5.15) in place of the path radiance.

$$\frac{C(R)}{C(0)} = \frac{I_B(0)}{I_S\left(1-e^{-\alpha_S L}\right) f(\theta)/4\pi} e^{-\alpha R} \tag{5.18}$$

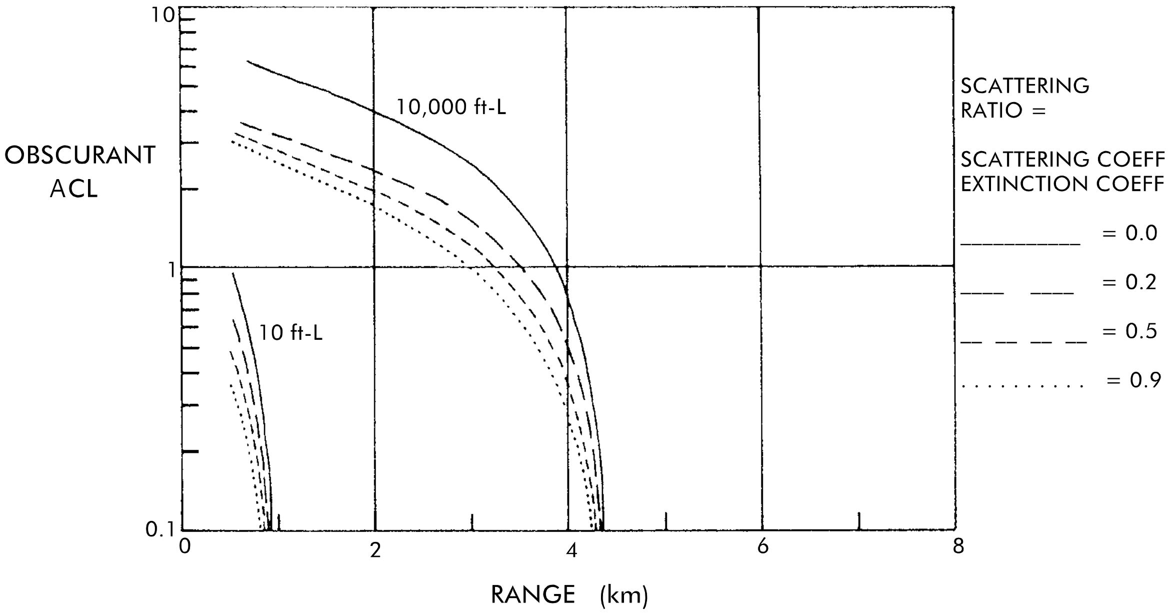
Note that if the scattering coefficient is small, the obscurant radiance will be small. Since this is the effective path radiance, then the contrast transmittance is increased. Black smokes require larger overall attenuations to produce a desired reduction in contrast. The magnitude of the effect also depends on the strength of the external illumination. The subject of white and black smokes could be discussed in considerably more detail. However, such discussion is beyond the scope of this work.

Figure 5-2 shows the effects of "gray" smokes on detection of a target by a small television sensor. The detection range is plotted versus αCL (ACL in the figure) for two different external illumination levels (daylight – 10000 ft-L and dusk – 10 ft-L). Hypothetical smokes with different scattering ratios η

$$\eta = \alpha_S / \alpha \tag{5.19}$$

ranging from 0 (pure absorption) to 0.9 (strongly scattering). Examination of the figure clearly indicates that considerably larger αCL values are needed to impact detection range when "black" smokes are employed than when "white" smokes are employed. At low light levels, then any additional reduction in contrast is magnified, although the "white smoke is better than black smoke" comparison still holds.

Figure 5-2. Comparison of white/black smokes effects on television detection performance.



Pasquill Stability and Atmospheric Diffusion

The effectiveness of employment of military obscurants or chemical and biological warfare agents depends on many atmospheric factors. One of the most important of these is the **Pasquill stability**. The Pasquill stability conditions relate to the rate of vertical diffusion of particles. The more stable the atmosphere, the less vertical transport of particulates occurs. In addition, the horizontal diffusion of particles (absent the effects of variable wind directions) tends to correlate well with the vertical diffusion. The more stable the atmosphere, the less horizontal diffusion of particulates occurs. Pasquill [7],[8] identified six levels of stability as shown in Table 5-5. A seventh category (G or 7 – Extremely Stable) was added later.

Table 5-5. Pasquill Stability Conditions[7]-[9]

CONDITION	PASQUILL DESIGNATION	ALTERNATE DESIGNATION
EXTREMELY UNSTABLE	A	1
UNSTABLE	B	2
SLIGHTLY UNSTABLE	C	3
NEUTRAL	D	4
SLIGHTLY STABLE	E	5
STABLE	F	6
EXTREMELY STABLE	G	7

Turner [9] further developed a correlation between wind speed, cloud cover, and insolation (amount of incident solar radiation) and the stability condition. Specifically he defined a net radiation index according to the following criteria:

1) Net Radiation Index = 0 if cloud cover is 10/10 and ceiling is less than 7000 feet whether day or night;

2) Nighttime (between sunset and sunrise) – a) Net Radiation Index = -2, if cloud cover is less than or equal to 4/10 and b) Net Radiation Index = -1, if cloud cover is greater than 4/10.

3) Daytime – a) Determine Insolation Class from Table 5-6;

b) Net Radiation Index = Insolation Class, if cloud cover is less than or equal to 5/10;

c) Net Radiation Index = Adjusted Insolation Class, if cloud cover is greater than 5/10.

Adjust the Insolation Class according to:

i) Subtract 2 if ceiling is less than 7000 feet;

ii) Subtract 1 if ceiling is between 7000 feet and 16,000 feet;

iii) Subtract 1 if total cloud cover is 10/10 and ceiling is above 7000 feet;

iv) If Insolation Class has not been modified by above, then assume the Adjusted Insolation Class is the same as the Insolation Class;

v) If the Adjusted Insolation Class is less than 1, set it equal to 1.

Given the Net Radiation Index and the Wind Speed, the Pasquill Stability class is determined from Table 5-7.

Table 5-6. Insolation Class Number.[9]

SOLAR ALTITUDE - A	INSOLATION	INSOLATION CLASS #
A > 60°	STRONG	4
35° < A ≤ 60°	MODERATE	3
15° < A ≤ 35°	SLIGHT	2
A ≤ 15°	WEAK	1

Table 5-7. Stability Class as a function of net radiation index and wind speed.[9]

WIND SPEED	NET RADIATION INDEX						
(knots)	4	3	2	1	0	-1	-2
0-2	1	1	2	3	4	6	7
2-4	1	2	2	3	4	6	7
4-6	1	2	3	4	4	5	6
6-7	2	2	3	4	4	5	6
7-8	2	2	3	4	4	4	5
8-10	2	3	3	4	4	4	5
10-11	3	3	4	4	4	4	5
11-12	3	3	4	4	4	4	4
≥12	3	4	4	4	4	4	4

The Stability Class correlates with the rate of vertical diffusion of particulates. Figure 5-3 illustrates the behavior of a smoke plume for nominal stable (class F(6)), neutral (class D(4)), and unstable (class B(2)) atmospheres. The stable plume reaches a stable vertical thickness (determined by the height of the inversion layer) and then does not diffuse vertically any farther. The neutral plume diffuses slowly but smoothly to ever-increasing thicknesses. The particulate concentration remains highest near the surface. The unstable plume diffuses rapidly and may actually lift from the surface. The influence on performance should be obvious.

Pasquill also estimated the amount of diffusion that could be expected for each stability class. Consider a continuous emitter of material (such as a smokestack, a smoke generator, or a venting gas tank) located at ground level and releasing *Q* kg/s of material. At a distance *d* from the source, the plume will exhibit a roughly Gaussian concentration profile in the vertical direction (see Figure 5-4) given by

$$C(z) \approx C(0)e^{-2z^2/h^2} \tag{5.20}$$

where *h* is a scale height which we have defined to be the $1/e^2$ height (Pasquill actually used the 10% height but any errors introduced by this difference will be small) and may be found from Figure 5-5 as a function of stability class and downwind distance *d*.

Figure 5-3. The effect of stability class on vertical diffusion of a plume.

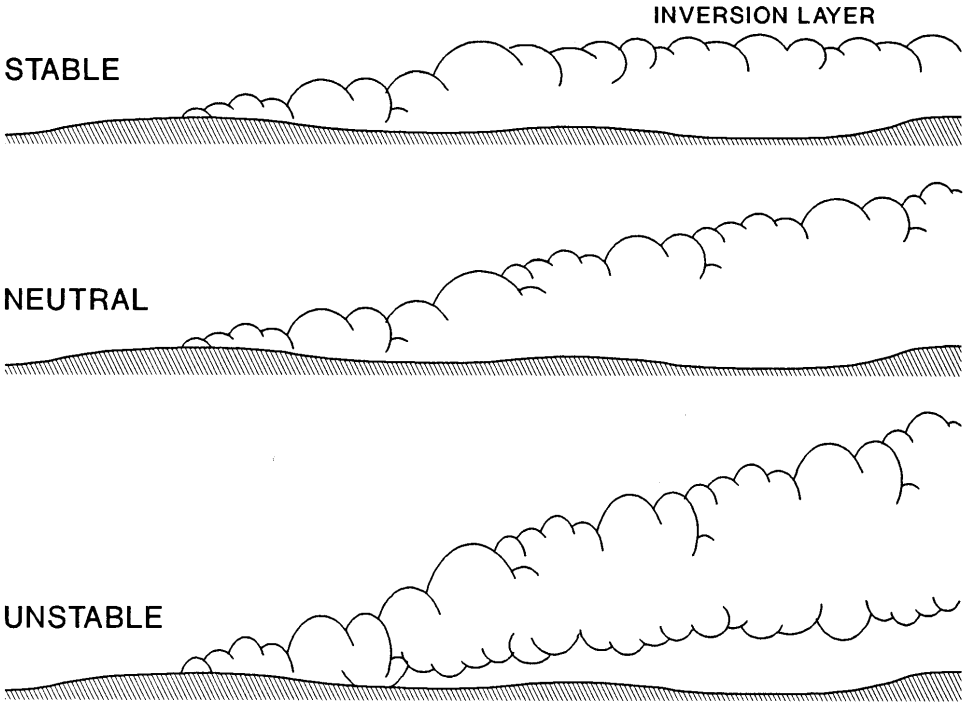


Figure 5-4. Geometry of Pasquill diffusion.

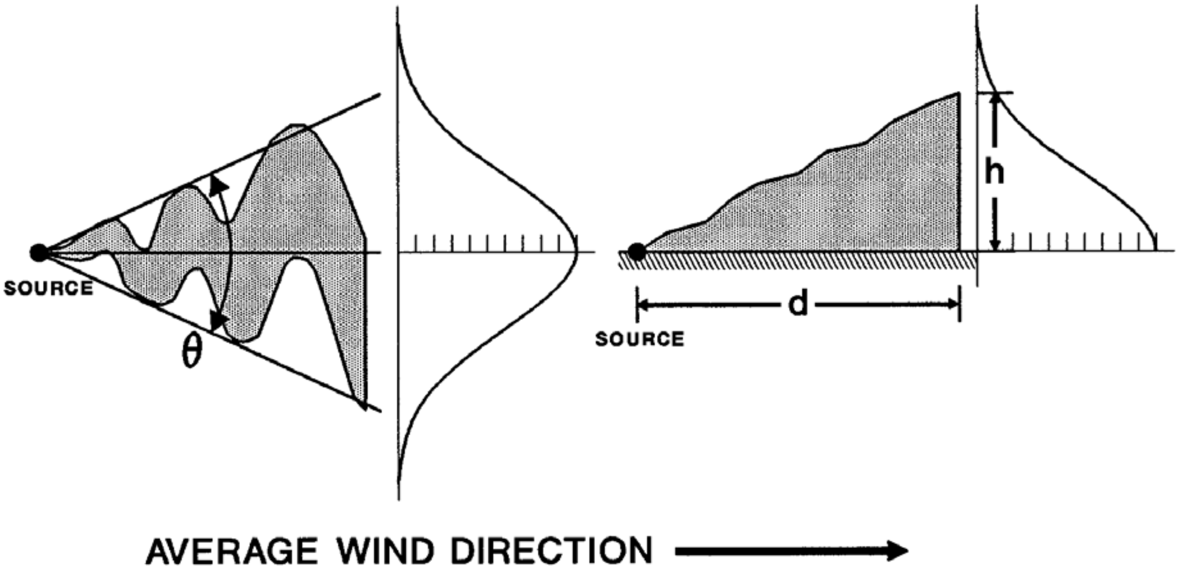
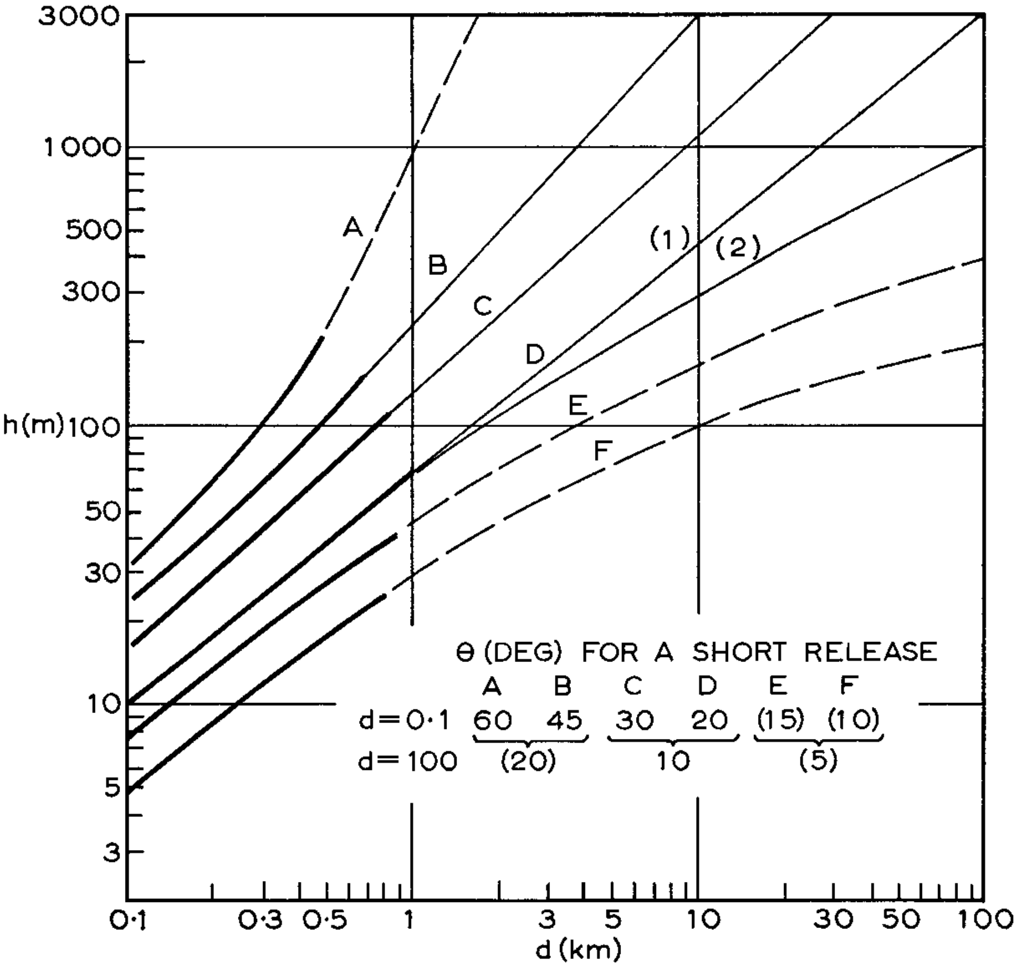


Figure 5-5. Angular diffusion rates for different Pasquill stability categories.



Note the curve for class D splits into two curves. Curve (1) should be used up to the top of the dry adiabatic layer. Above this layer, in sub-adiabatic conditions, a curve parallel to curve (2) with value at the top of the dry adiabatic layer equal to the value of curve (1) at the top of the dry adiabatic layer should be used.

Diffusion in the horizontal direction is different than diffusion in the vertical dimension. The transverse diffusion has two components. One will be a truly random diffusion that is similar (but not equal to) the diffusion in the vertical direction. Of greater significance, however, is term caused by time variations in wind direction. As the plume moves downwind, the wind will fluctuate in direction about the long-term average wind. The spatial scale of these fluctuations is relative large although the temporal scale may be as short as seconds. Large regions of the plume will temporarily move in one direction and then another. Figure 5-4 shows a "snapshot" of how the horizontal diffusion might look at one instant in time.

If one measures the concentration as a function of time and averages over a relative long time, the concentration as a function of downwind distance $x = d$ and transverse distance y from the long-term average wind vector (with origin assumed to be at the source) will also be Gaussian

$$C(y) \approx C(0)e^{-8y^2/\theta^2 d^2} \tag{5.21}$$

where θ is the total angular width of the plume (defined to be the full angle at $1/e^2$ concentration points) and has been given in Figure 5-5 for short and long ranges.

Combining the horizontal and vertical diffusion terms and normalizing the distributions to equal the source emission rate yields

$$C(d,y,z) \approx \frac{8Q}{\pi h \theta d u} e^{-8y^2/\theta^2 d^2} e^{-2z^2/h^2} \tag{5.22}$$

where C has units of kg/m^3 if Q has units of kg/s, h and d have units of meters, u is the average wind velocity in m/s, and θ is in radians.

If the source is elevated more than a few meters above the ground, it is necessary to account for downward diffusion as well as upward diffusion. The two rates may be assumed to be equal. Convection will cause the upward rate to be slightly higher, but particulate settling will tend to increase the downward rate. Except in extremely unstable conditions or if particulates are very large, neither effect should dominate the random isotropic diffusion caused by atmospheric turbulence. Assuming the downward rate is equal to the upward allows the vertical concentration equation to be modified by shifting the mean and adding a term which describes “reflection” of the diffusion from the ground. If the source is at height H then we have

$$C(d,y,z) \approx \frac{4Q}{\pi h \theta d u} e^{-8y^2/\theta^2 d^2} \left(e^{-2(z-H)^2/h^2} + e^{-2(z+H)^2/h^2} \right) \tag{5.23}$$

as the governing concentration equation.

The results in Eqs. (5.22) and (5.23) are long-term average values. Over the long term, the fluctuations in wind speed and direction producing the sinuous behavior of Figure 5-4 will average out. At some points in space and time, the equations will overestimate the actual concentrations. At other points, they will underestimate the concentrations. But over the course of many minutes, the average values will prevail.

If local real-time meteorology data is available giving actual wind speed and direction traces, then better concentration estimates can be obtained. Even better results are obtained if the point meteorological data is used as inputs to a micrometeorological prediction code that also uses real terrain data to refine the spatial variations in wind speed and direction. More detailed discussion of these codes is beyond the scope of the present work.

The preceding discussion was pertinent to a continuous, slow material release from a point source. Releases that are localized in space and time behave slightly differently. Consider the explosive release of a quantity of material. This might result from the explosion of an artillery shell filled with smoke agent. The explosion rapidly produces a cloud of material which will have substantial but limited dimensions. An artillery shell might produce a roughly spherical cloud 50 meters in radius. This **puff** as it is sometimes called will then diffuse and drift with the wind. Its small size means that it will not develop the sinuous character shown in Figure 5-4. Diffusion will be dominated more by turbulence than by wind motions. A reasonable assumption is that the horizontal diffusion (both transverse to the wind and along the wind) of the puff will be comparable to the vertical diffusion. If the puff consisted of a release of *M* kg of material at ground level into a hemispherical cloud initially characterized by radius *R*, then at a distance *d* we would have

$$C(d,y,z) \approx \frac{4\sqrt{2}M}{\pi\sqrt{\pi}\left(h^2+R^2\right)^{3/2}} \begin{bmatrix} e^{-8\left(d-uT\right)^2/\left(h^2+R^2\right)} \times \\ e^{-8y^2/\left(h^2+R^2\right)} \times \\ e^{-2z^2/\left(h^2+R^2\right)} \end{bmatrix} \tag{5.24}$$

where *uT* is the distance the cloud would drift in time *T* (the assumed time of observation) and *h* would be determined from Figure 5-5.

Eq. (5.24) gives the concentration of the puff about the center of the puff as it moves. However, the puff itself will follow a sinuous course similar to that shown in Figure 5-4. The puff will pass over some points and not pass over others. Comparing the concentration calculated from Eq. (5.24) with the concentration associated with the long term average values for the continuous-release source gives an estimate of the probability that a single puff will pass over a given point.

An expression related to Eq. (5.24) could be generated for a puff released at a height *H* above the ground. This is left as an exercise for the reader. Other source geometries can be considered by recognizing that any complex cloud geometry can be decomposed into a number of small roughly spherical puffs. For example, a reasonable approximation to the continuous emission case described earlier would be to assume a regular release of puffs such that as soon as one puff had drifted a distance *2R* from the source point another puff was produced. The **line source** is another interesting geometry. It is produced when a vehicle-mounted spray tank is used as the release mechanism. The material is released in a tight cloud that stretches over the distance traveled by the platform. In an aircraft spray situation, the material may be emitted into a source several kilometers long (e.g., 10 second emission at 300 m/s is 3 km) before any appreciable drift or diffusion can occur. The behavior of a line source can be estimated by assuming that it is produced instantaneously in a cylindrical geometry of a given radius, and then applying Gaussian diffusion in all three dimensions. The mathematics of this will not be developed here.

The Pasquill Stability class also correlates with other meteorological parameters related to the effectiveness of employing obscurants (Table 5-8). For example, the wind speed profile with height is often approximated by

$$\frac{u(z)}{u(z_R)} = \left(\frac{z}{z_R} \right)^p \tag{5.20}$$

Table 5-8. Relationship of Pasquill stability classes to other stability parameters. [6]

PASQUILL STABILITY			**A** (1)	**B** (2)	**C** (3)	**D** (4)	**E** (5)	**F** (6)
MEAN WIND SPEED (m/s)			3	4	5	8	5	3
MIXING HEIGHT (m)			1000-3000	500-2000	250-1000	100-800	50-200	30-50
ROUGHNESS FACTOR		**PARAMETER**						
z_0 (m) =	0.01	u_F	0.201	0.259	0.279	0.463	0.230	0.094
		p	0.062	0.083	0.100	0.145	0.319	0.536
		R_I	-3.58	-0.87	-0.34	0	0.33	2.66
		L	-7.8	-16.8	-51.5	∞	65.3	14.0
		μ	-157	-94.0	-33.0	0	21.5	40.9
		S	-1.24	-0.96	-0.85	0	0.84	0.69
	0.05	u_F	0.275	0.346	0.413	0.604	0.304	0.155
		p	0.086	0.119	0.144	0.189	0.384	0.588
		R_i	-3.26	-0.67	-0.21	0	0.09	0.34
		L	-9.4	-23.6	-94.4	∞	113.7	22.6
		μ	-178	-89.4	-26.7	0	18.3	41.8
		S	-2.17	-1.52	-1.09	0	2.72	1.06
	0.10	u_F	0.326	0.404	0.475	0.695	0.354	0.128
		p	0.103	0.143	0.175	0.217	0.363	0.617
		R_i	-3.12	-0.57	-0.16	0	0.14	1.86
		L	-10.4	-29.0	-137	∞	163.9	28.9
		μ	-191	-84.9	-21.1	0	13.2	27.0
		S	-1.98	-1.89	-1.23	0	4.61	6.37
	0.50	u_F	0.547	0.649	0.714	1.068	0.554	0.176
		p	0.189	0.262	0.311	0.334	0.447	0.419
		R_i	-2.66	-0.62	-0.05	0	0.07	1.23
		L	-14.1	-70.8	-568	∞	332.0	65.0
		μ	-236	-55.9	-7.7	0	10.2	16.5
		S	-8.55	-3.57	-1.21	0	3.50	1.94

where z_R is a reference height (10 m) and p is an empirical power law exponent. This equation is strictly only valid for neutral conditions. A more useful profile is given by

$$u(z) = \left(\frac{u_F}{k}\right)\left[\ln\left(\frac{z}{z_0}\right) + \psi\left(\frac{z}{L}\right)\right] \tag{5.21}$$

where u_F is the friction velocity, k is the von Karman constant (= 0.40), z_0 is a surface roughness parameter, and L the Monin-Obukhov length. The function ψ takes the form

$$\psi\left(\frac{z}{L}\right) = \begin{cases} 4.7\left(\frac{z}{L}\right) & \text{stable} \\ 0 & \text{neutral} \\ \ln\left(\frac{1+x^2}{2}\right) + 2\ln\left(\frac{1-x}{2}\right) - 2\tan^{-1}x + \frac{\pi}{2} & \text{unstable} \end{cases} \tag{5.22}$$

where

$$x = \left(1 - 15\frac{z}{L}\right)^{1/4} \tag{5.23}$$

The surface roughness parameter z_0 can be related to the mean height z_E of surrounding terrain elements, including trees, grass, hills, buildings, etc.) by

$$\ln z_0 = 1.19 \ln z_E - 2.85 \qquad \text{units of meters} \tag{5.24}$$

Table 5-8 also gives the mixing height. This is the height at which atmospheric conditions are not influenced by surface conditions. It can vary from tens of meters under stable conditions to kilometers under unstable conditions. The Richardson number R_i, the Kazanski-Monin stability parameter μ, and the static stability S are other stability related parameters that are included in the Table in the spirit of a Rosetta Stone.

Transmission Through Stratified Media

Obscurants, clouds, and fogs are typically stratified or localized. In these instances, use of the simple expression

$$T = e^{-\alpha R} \tag{5.25}$$

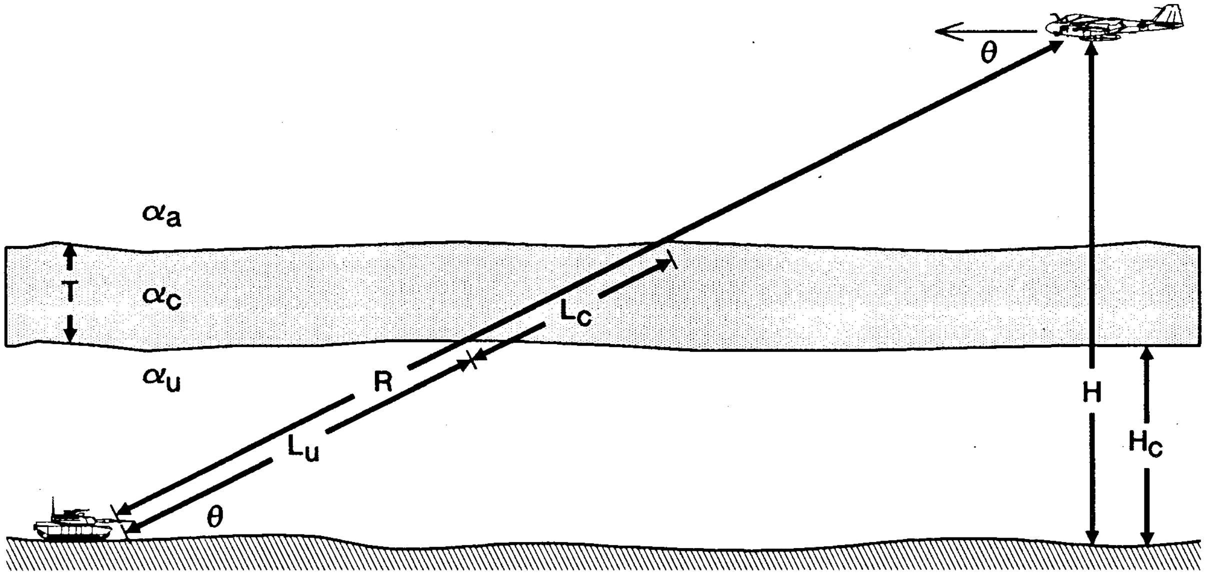
is not appropriate. If used with the obscurant attenuation coefficient for α, then Eq. (5.25) will grossly overestimate the attenuation. If the obscurants, clouds, or fogs are indeed localized, stratified, or both, then more realistic calculations can be obtained using a discrete approach.

Consider a horizontally stratified absorber such as a low cloud or a smoke plume as shown in Figure 5-6. The sine of the elevation angle (or depression angle) depending on your point of view is given by

$$\sin\theta = \frac{H}{R} = \frac{T}{L_C} = \frac{H_C}{L_U} \tag{5.26}$$

where H is the platform height, R is the total slant range, T is the absorber thickness (or the thickness of the intermediate layer if all have substantial attenuation), L_C is the slant path distance through the central absorber, H_C is the height of the base of the central absorbing layer (or the thickness of the

Figure 5-6. Geometry for calculating attenuation in a horizontally stratified environment.



bottom absorbing layer), and L_U is the slant path distance through the bottom layer. Obviously

$$L_a = R - L_C - L_U \tag{5.27}$$

is the slant path distance through the uppermost layer. The slant paths through the lower layers can be calculated from

$$L_C = T / \sin\theta = TR / H \tag{5.28}$$

and

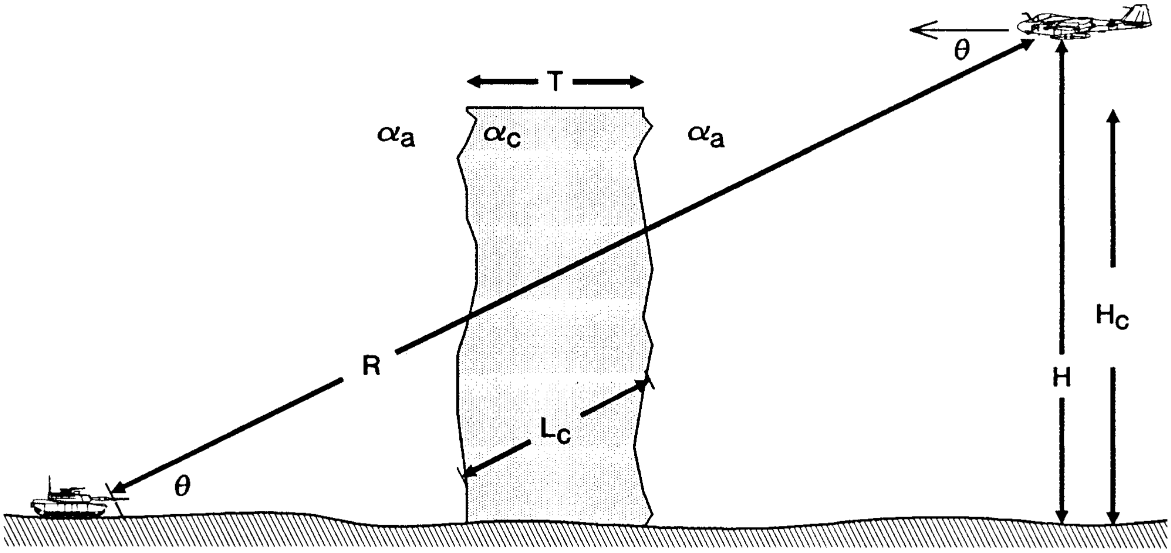
$$L_U = H_C / \sin\theta = H_C R / H \tag{5.29}$$

The actual attenuation is the sum of the slant path contributions from each layer

$$\begin{aligned} T &= e^{-\alpha_a \left(R - L_C - L_U\right) - \alpha_C L_C - \alpha_U L_U} \\ &\qquad\qquad\qquad . \\ &= e^{-\alpha_a \left(R - (T/H) - (H_C/H) \right) - \alpha_C R (T/H) - \alpha_U R (H_C/H)} \end{aligned} \tag{5.30}$$

Now consider a vertically stratified environment in which there is a single vertical intervening absorber in an otherwise homogeneous medium, as shown in Figure 5-7. The sine of the depression

Figure 5-7. Geometry for calculating attenuation in a vertically stratified environment.



angle is defined by

$$\sin\theta = H / R \tag{5.31}$$

where *H* is the platform altitude and *R* is the total slant path distance. The cosine of the depression angle can be readily determined to be

$$\cos\theta = T / L_C = \left(R^2 - H^2\right)^{1/2} / R\,. \tag{5.32}$$

The slant path distance through the smoke can now be evaluated to be

$$L_C = RT / \left(R^2 - H^2\right)^{1/2}. \tag{5.33}$$

The total path attenuation is the sum of the attenuation due to the background plus the attenuation due to the cloud

$$T = e^{-\alpha_a\left(R - L_C\right) - \alpha_C L_C} \approx e^{-\alpha_a R - \alpha_C L_C}\,. \tag{5.34}$$

If the thickness of the cloud is small compared to the slant path distance (*T*<<*R*), then the final approximation in Eq.(5.34) is a very good one.

Multipath

Radars at low altitudes above the ground often face conditions when the radar beam simultaneously illuminates both the ground and the target. In this case, the radiation can take at least four distinct paths between the transmitter to the target and back to the receiver. As shown in Figure 5-8 there is the desired direct path (A to C to A). There are two paths that bounce once off the ground (A to B to C to A and A to C to B to A) called diplane paths. There is also one path that bounces twice off the ground (A to B to C to B to A), occasionally called the mirror image path.

If we assume that the radar is pointed at the target, then the relative gain in the direction of the ground bounce is ΔG. The effective reflectance of the ground is denoted by ρ. L denotes the distance between the radar and the target along the direct path. S denotes the distance between the radar and the target along the bounce path. The relative signal power received at any point is the square of the sum of the four paths as shown below.[10]

$$\frac{P}{P_0} = \left(e^{ik2L} + 2\rho \Delta G \, e^{ik(L+S)} + \rho^2 \, \Delta G^2 \, e^{ik2S} \right)^2$$

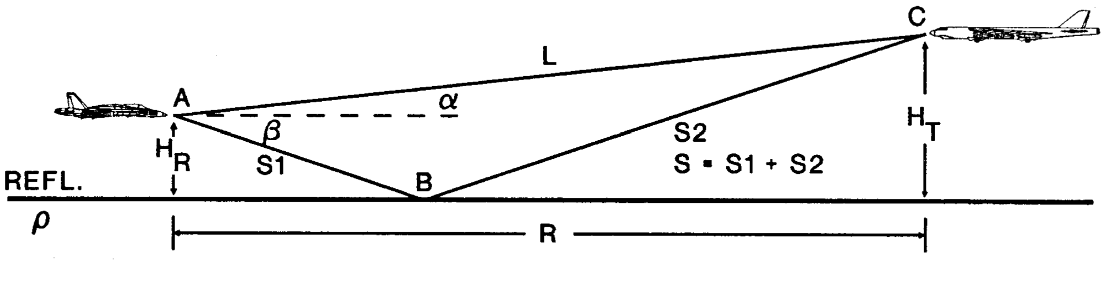
$$= \left(e^{ikL} + \rho \Delta G \, e^{ikS} \right)^4 \qquad (5.35)$$

$$= e^{i4kL} \left(1 + \rho \Delta G \, e^{ik(S-L)} \right)^4$$

Multipath of this type will occur when the angle between the target and the ground bounce point is less than the beamwidth

$$\alpha + \beta = \tan^{-1}\left[\left(H_T - H_R \right) / R \right] + \tan^{-1}\left[\left(H_T + H_R \right) / R \right] < \theta_{BW} \qquad (5.36)$$

Figure 5-8. Geometry for multipath lobing calculations.



where α is the elevation angle of the target and β is the depression angle of the ground bounce point, *R* is the ground range between target and radar, H_R is the height of the radar above the ground, and H_T is the height of the target above the ground. Nominally, this "multipath condition" is satisfied when

$$H_T < \lambda R / 2D \tag{5.37}$$

Equation (5.35) can be simplified by making the following simplifying approximations

$$S = \left[\left(H_T + H_R \right)^2 + R^2 \right]^{1/2} \approx R + \left[\left(H_T + H_R \right)^2 / 2R \right] \tag{5.38}$$

and

$$L = \left[\left(H_T - H_R \right)^2 + R^2 \right]^{1/2} \approx R + \left[\left(H_T - H_R \right)^2 / 2R \right]. \tag{5.39}$$

Using these approximations, the path difference between direct and bounce paths is given by

$$S - L \approx 2 H_T H_R / R. \tag{5.40}$$

It is also instructive to make the additional simplifying assumptions: $\Delta G = 1$ (equivalent to assuming that the target to ground bounce angle is small compared to the beamwidth) and $\rho = -1$ (equivalent to assuming that the ground is a perfect conducting surface – this results in a π phase shift which gives the minus sign to the reflectance). Using these approximations, Eq. (5.35) becomes

$$\frac{P}{P_0} \approx \left[2 \sin\left[k\left(S - L \right) / 2 \right] \right]^4 \approx 16 \sin^4 \left[k H_T H_R / R \right]. \tag{5.41}$$

This result implies that P/P_0 varies from 0 to 16 times the free space value (P_0) as H_T increases or *R* increases. The nulls in the received power occur at integer multiples (*m*) of π. That is, the nulls occur when

$$m\pi = k\, H_T\, H_R / R \tag{5.42}$$

or when

$$H_T = m \lambda R / 2 H_R \tag{5.43}$$

or

$$R = 2 H_T H_R / m\lambda \tag{5.44}$$

Multipath interference is multiplicative to the normal range dependence of the radar equation. If we define R_{DET} to be the maximum detection range of the radar (against the target of interest) in the absence of multipath, then the received power relative to the minimum power required for detection can be written as

$$\frac{P}{P_{DET}} \approx 16 \left(\frac{R_{DET}}{R} \right)^4 \sin^4 \left[k H_T H_R / R \right] . \tag{5.45}$$

If P/P_{DET} is greater than 1, then the target will be detectable; if it less than 1, it will not be detectable.

In Figures 5-9 and 5-10 we have plotted P/P_{DET} versus range for two generic radars with realistic parameters. Figure 5-9 shows the effects of multipath on a search radar. The target is assumed to be a cruise missile at 5 m altitude. The primary effect is a potential reduction of the maximum detection range from 20 km to less than 14 km. Because the reflecting surface is assumed to be conducting, the π phase per ground reflection leads to destructive interference at zero path difference ($m = 0$). This causes performance at long ranges to be worse than predicted by the $1/R^4$ dependence of the radar equation. Such a reduction in maximum detection range could have a catastrophic effect on the system timeline. If the missile velocity is 300 m/s, then the available reaction time is reduced from a maximum of 67 seconds to less than 47 seconds. The other major nulls occurs at ranges too short to be of significance. Note, that if different values of target height, radar height, wavelength, or nominal detection range had been chosen, then the $m=0$ interference might have occurred at ranges too long to be of significance.

Figure 5-9. Effects of multipath interference on the performance of a search radar.

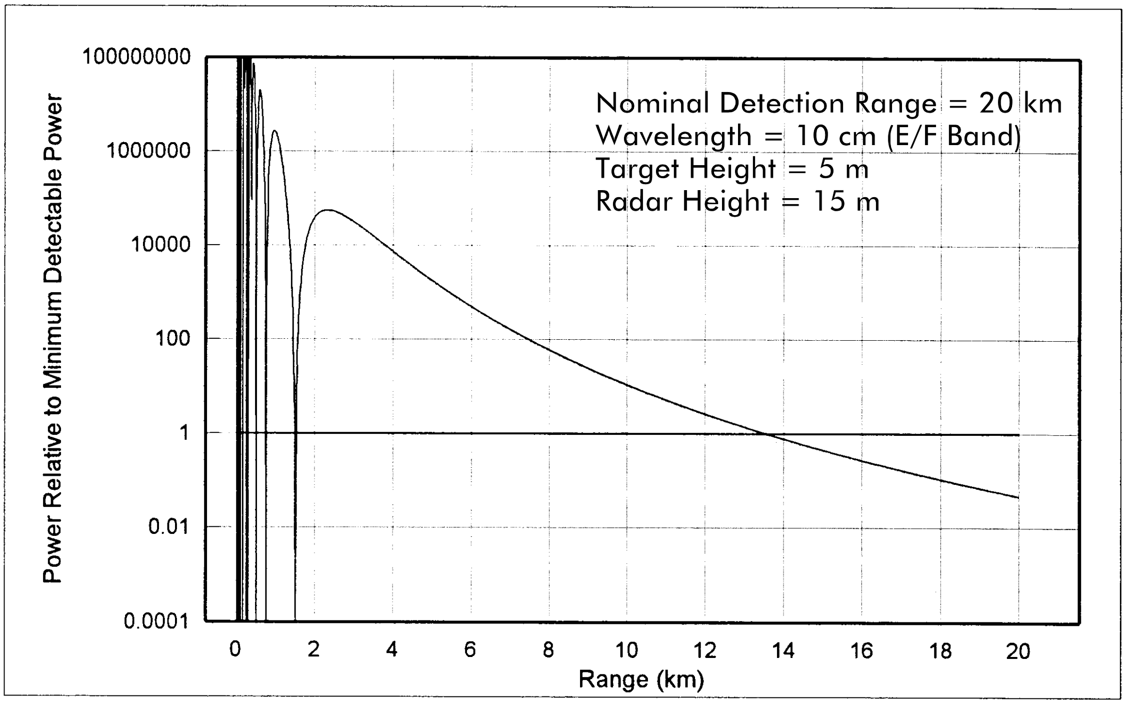


Figure 5-10. Effects of multipath interference on the performance of a tracking radar.

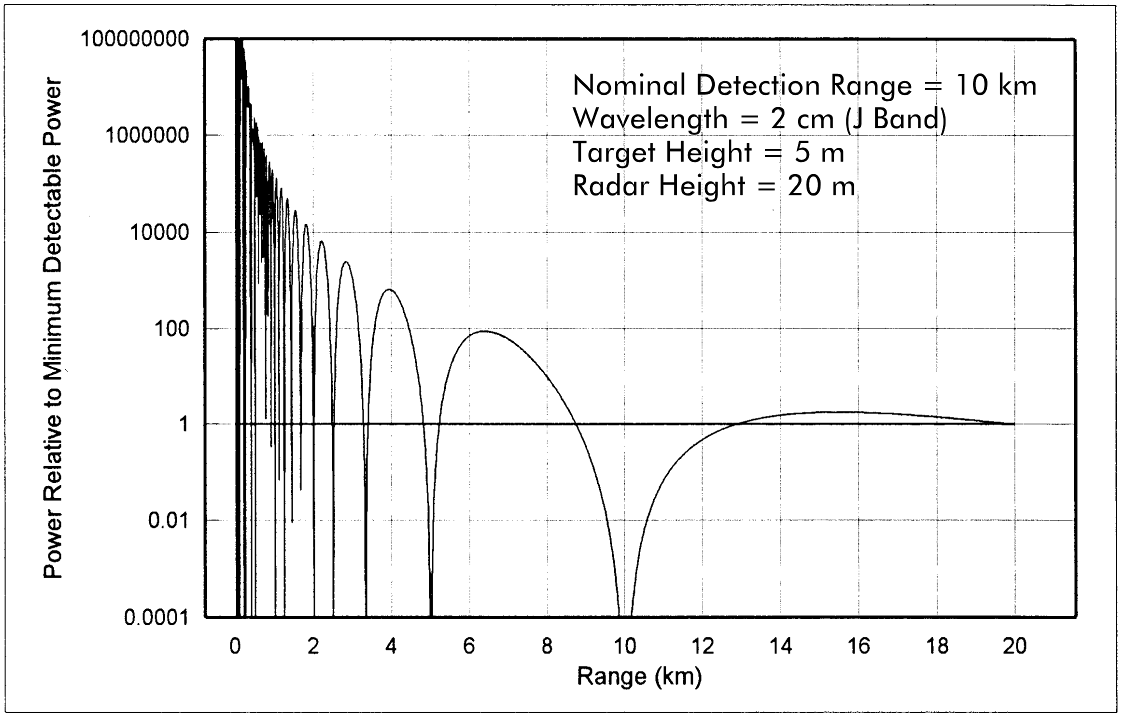


Figure 5-10 shows the effects of multipath on a generic tracking radar against the same cruise missile target. In this case, there is an enhancement of detection at very long ranges (13-20 km even though the nominal detection range is only 10 km). Unfortunately, this is not of much benefit to a radar that does not need greater than 10 km detection range. There is a broad null at the nominal detection range, which would cause complete loss of track even if it had been established at the longer ranges. At roughly 8.7 km, the missile becomes detectable again. There is another null at 5 km which is long enough to potentially cause loss of track again. As the missile comes closer and closer, the nulls become more frequent but also shorter in duration (a consequence of the *1/R*^{4} dependence on the received signal). Note that every null will go to zero relative power – the coarseness of the data used in plotting the figures causes some nulls to appear to stop at non-zero values. These nulls will clearly have some effect on tracking accuracy but the exact effects would have to be determined through simulation of the tracking loop. It is apparent that multipath can cause performance limitations in tracking as well as search radars.

Terrain Masking

Terrain masking refers to the difficulty in detecting objects in the open due to terrain, cultural (buildings, bridges, levees, etc.), or vegetation features interposing in the line of sight (see Fig. 5-11). It has significant impact on mission accomplishment in air-to-surface, surface-to-air, and surface-to-surface missions. Terrain masking will limit the ranges at which targets can be reasonably detected and attacked. It will force airborne platforms to fly in regimes that are less than optimal from survivability aspects. It will shorten the timelines associated with target acquisition, tracking and weapon delivery for both attacking and defending systems. Thus terrain masking should be a serious concern to system engineers.

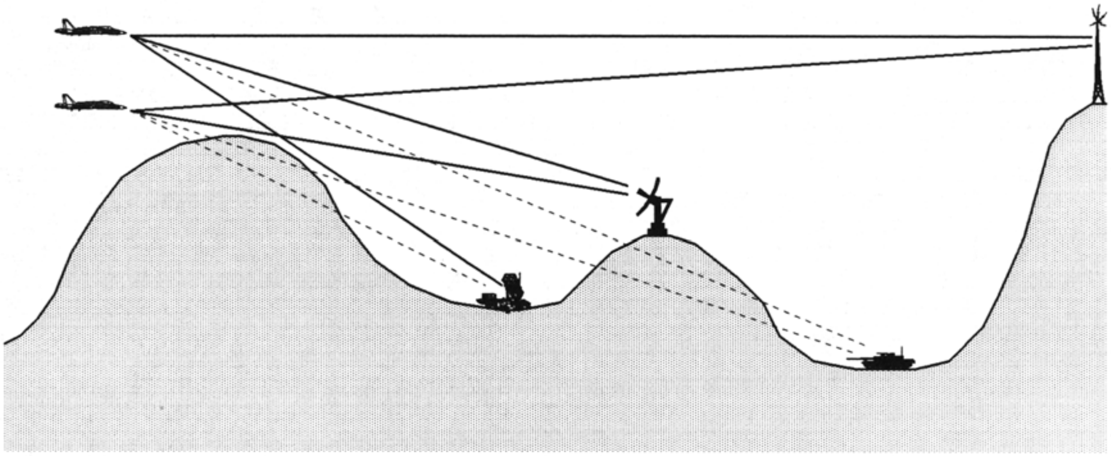
As is evident from Figure 5-11, altitude is a prime factor in terrain masking. At low altitudes, a clear line of sight exists to only a fraction of the terrain surface and thus to only a fraction of the important objects lying within a certain distance of the platform. If the platform is higher, then more (but not necessarily all) of the objects of interest become visible.

A number of earlier studies have looked at various geographic regions and performed analyses to determine the probability of obtaining a clear line of sight to points on the ground as functions of platform elevation and ground range (distance along the surface from the observation point to the target point). In some of these studies it was found to be convenient to fit the statistical probability of obtaining a clear line of sight to a point on the ground to a function of the form [11]

$$p_{CLOS} = e^{-aR^b H^c} \tag{5.46}$$

where R is the ground distance (in kilometers), H is the platform elevation (in kilometers), and a, b, and c are coefficients resulting from the curve fitting. An example from such a curve fit used the

Figure 5-11. Effect of platform elevation on line of sight.



values, *a* = -0.0003, *b* = 1.658, *c* = -1.315 to represent “gently rolling terrain”.[11] These values were used to generate the curves in Figure 5-12. Examination of the curves in this Figure indicates that the probability of having a line of sight out to 5 km is extremely limited even in benign terrain, unless platform altitudes in excess of 30-60 m are obtained. If high line of sight probabilities are required, then very large altitudes (1000 m or so) will be required.

Although line of sight studies have been performed several times in the past, their limited availability makes them difficult to use. They are also not guaranteed to have data relevant to every current scenario. Given the almost universal availability of digital terrain elevation data (DTED) maps and powerful laptop computers, it is almost as easy to generate one’s own line of sight probability data as it is to use the previous work. For example, Figure 5-13 shows a digitized topographic map of a portion of the National Training Center at Fort Irwin, CA.[12] Sometimes, examination of the terrain contours is enough to provide usable information. At the “randomly” selected point shown on the map (nominally on the 2800 ft contour line), simple examination of the directions with terrain contours higher than 2800 indicates that only within the roughly 120° arc indicated will lines of sight greater than 3000 meters be possible. This hints at an algorithm that one might use to obtain quantitative line of sight probability data for any geographical locale (for which data are available).

Figure 5-12. Clear line of sight probability for gently rolling terrain as a function of ground distance and platform altitude.

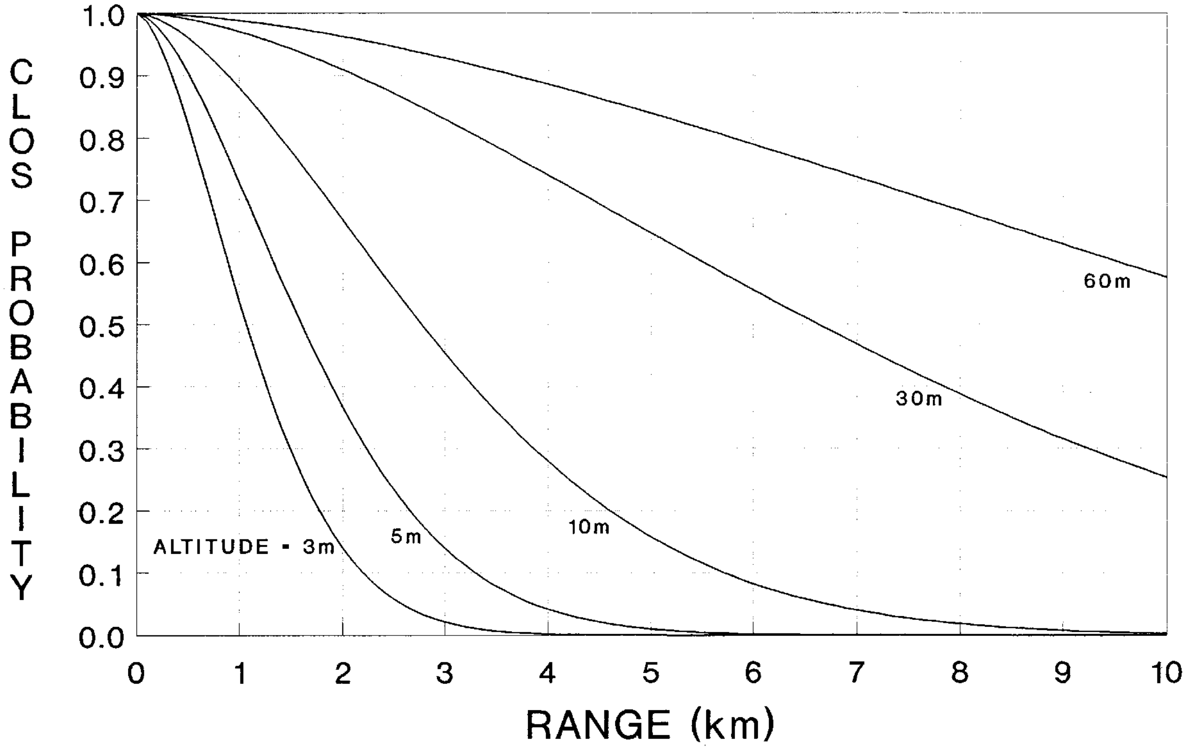
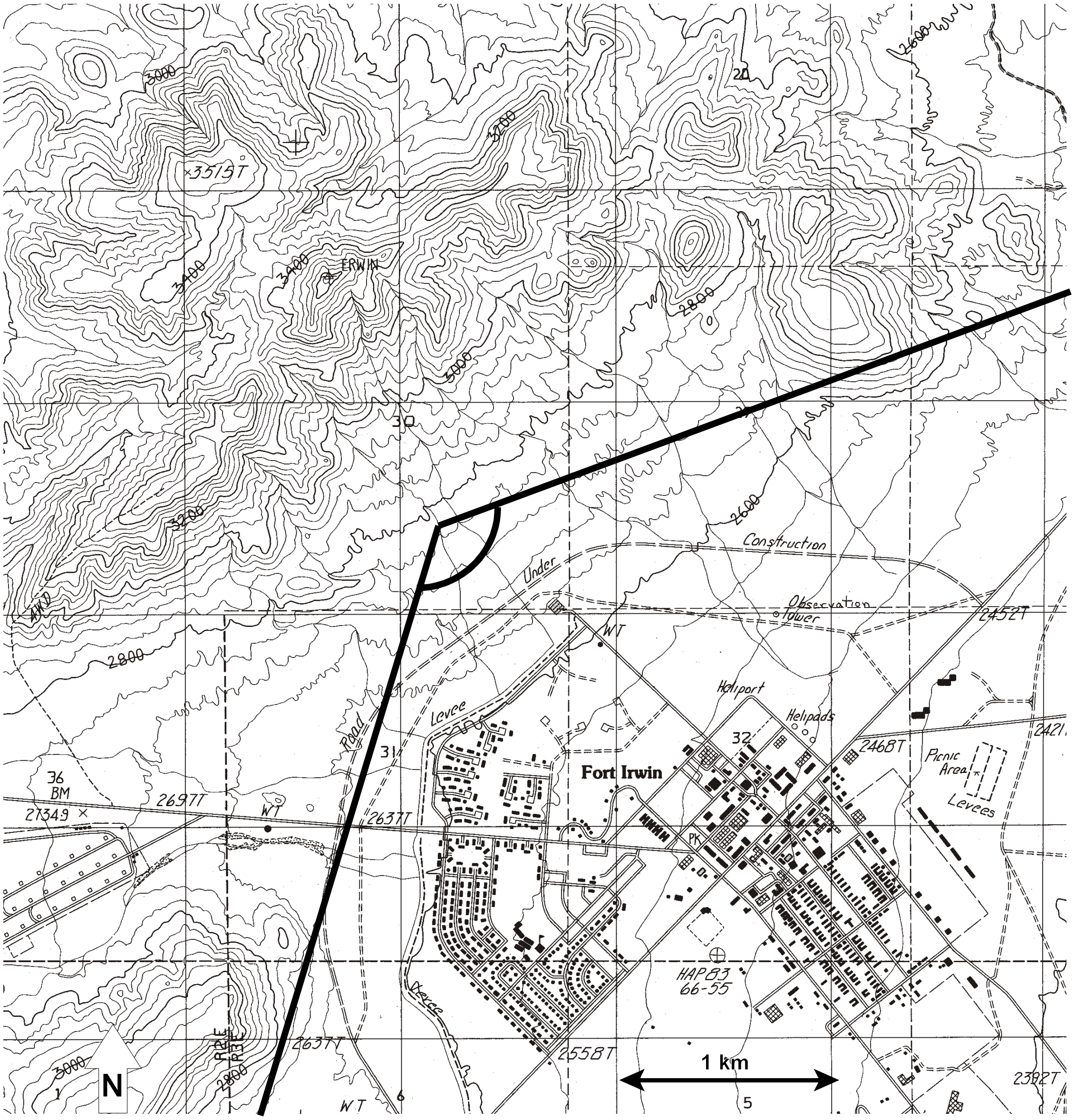


Figure 5-13. Topographic map of a small portion of Fort Irwin CA.[12]

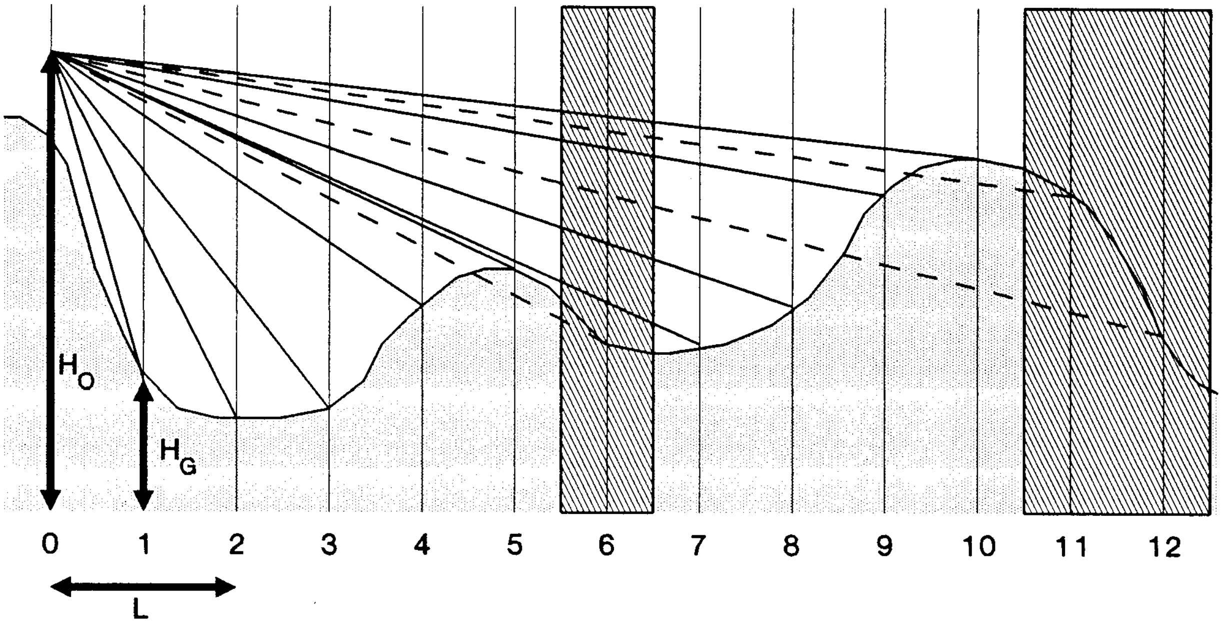


The algorithm proceeds as follows. First, randomly select a point within the terrain of interest and select an observation altitude. Second, for each of the major compass directions, extract the digital terrain values extending away from the selected point. Then, using the geometry indicated in Figure 5-14, calculate the tangent to the closest ground point using the expression

$$\tan\theta = L / \left(H_O - H_G\right) \qquad (5.)$$

where L is the ground distance, H_O is the observation platform altitude above sea level, and H_G is

Figure 5-14. Geometry used in line of sight probability analysis.

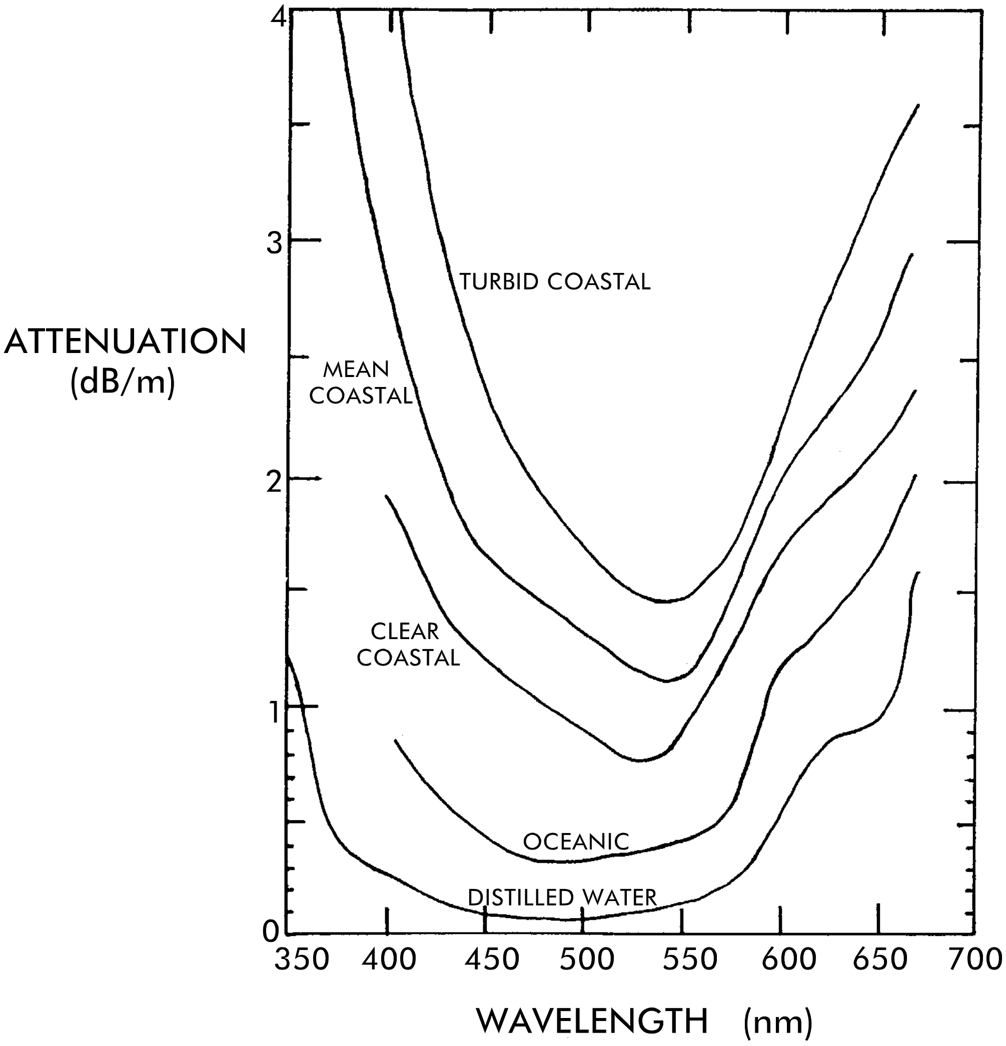


the ground elevation (above sea level). $\theta$is then the angle relative to nadir (the point directly below the observer) of the line of sight to the ground point. The first value of the tangent automatically becomes the running maximum value. The process is repeated for the next point. If the tangent of the next point exceeds the maximum prior tangent value, then a line of sight exists to that point. The current point then becomes the new prior maximum. This process is repeated for each succeeding ground point until a point is reached at which the tangent is less than the prior maximum (such as point 6 in Figure 5-14). If the tangent is smaller than the prior maximum, a clear line of sight does not exist to that ground point (some earlier ground point (5 in the figure above) masks it). Line of sight remains lost until the calculated tangent exceed the prior maximum again (establishing a new prior maximum). The process is repeated until all ground points out to a desired maximum range have been processed. The range values for this direction are then tagged (1 or 0) depending on whether a line of sight exists. The shaded regions in Figure 5-14 are those for which a clear line of sight was not obtained. The process is repeated for each compass direction and for a large number of other points randomly scattered throughout the geography of interest. After several hundred runs have been processed, the tagged values for each range are added and divided by the total number of runs. The result is the probability of clear line of sight versus range. The entire process is repeated for a variety of observation altitudes to obtain the height dependence. If desired, the line of sight probability values could be fit to an equation of the form of Eq. (5.). There are other algorithms that one can find in operations analysis texts, if the one described above does not seem attractive.[13]

Underwater Propagation

In Chapter 2 we calculated the attenuation coefficient in the visible and infrared for pure water (Figure 2-4) from its complex refractive index. At that time we noted that the only region that has acceptable transmission is the region around 500 nm (the blue-green portion of the spectrum. The attenuation of sea water is not greatly different from pure water but does have some minor differences. These are mainly due to the presence of particulates in sea water. Silt, plankton, and/or microbubbles (caused by wave breaking) can be present to appreciable extent. The particulates both absorb and scatter electromagnetic radiation with scattering being a significant contributor. The attenuation coefficient for a beam of radiation will depend on the degree of "turbidity" (or "muddiness"), that is, the amount of suspended particulates. Figure 5-15 compares the attenuation

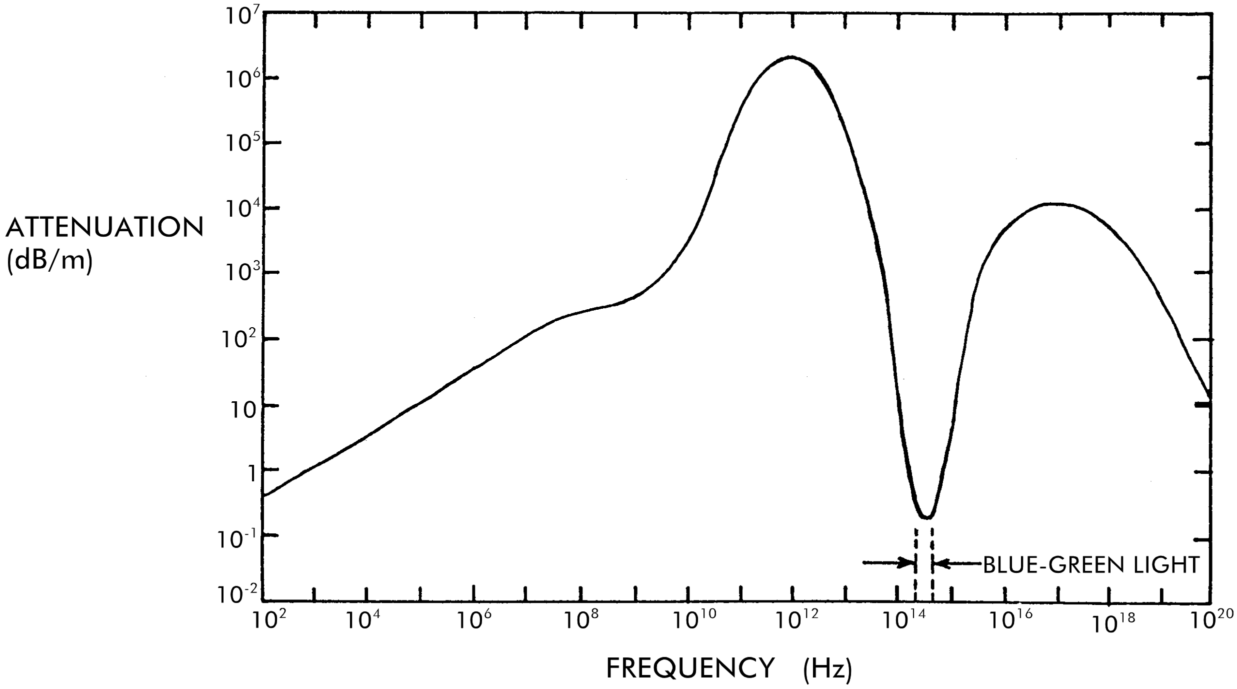
Figure 5-15. Attenuation coefficient of sea water at different locations.[14],[15]



coefficient of sea water at various locations to that of pure water. [14],[15] In general, mid-oceanic waters are relatively free of silt particulates and have reduced plankton levels. In the littoral waters both the plankton content (due to wider availability of nutrients) and the silt content (due to inflow of silt-laden river waters) are expected to increase. Attenuation is highest in estuaries and in the outflow of silt-bearing rivers, such as the Mississippi River.

Another aspect of underwater propagation is shown in Figure 5-16.[16] This figure shows the effective attenuation of electromagnetic radiation from gamma rays to radio frequencies. The window at blue-green wavelengths is clearly obvious. However, at frequencies below 1 kHz in the ELF (extremely low frequency) region, the attenuation again drops below 1 dB/m. Radio waves at ELF frequencies can penetrate deeply into sea water. A reasonable heuristic is that a radio communication system can operate to that depth at which an attenuation of 100 dB occurs. Note that even if this heuristic were increased to 200 dB, the difference in depth would be only a factor of two. Using this heuristic, we estimate that VLF frequencies (around 10 kHz) can penetrate to depths of the order of 30 meters, HF frequencies can only propagate to a depth of 1 meter or so, and higher frequencies will propagate even shorter distances. On the other hand ELF frequencies could penetrate to at least 100 meters depth or deeper if frequencies of 10-100 Hz were used. The only real problem with communication at ELF frequencies is the extremely small bandwidth that can be employed. At a carrier frequency of 100 Hz, it may take a minute or more to transmit a 100 byte message.

Figure 5-16. Attenuation of electromagnetic radiation in sea water. [16]



Surprisingly, the attenuation also falls below 1 dB/m in the gamma ray region of the spectrum. As we see in the Chapter 7, the attenuation coefficient of water falls below 0.01 cm^{-1} at photon energies above 100 MHz. However, only large particle accelerators are currently capable of producing such gamma rays, so it may be some time before technology permits the exploitation of this gamma ray window.

References

[1] Hoock, Donald W. Jr. and Sutherland, Robert A., "Obscuration Countermeasures" in David H. Pollock (Ed.) Countermeasure Systems Vol.7 of The Infrared & Electro-Optical Systems Handbook (SPIE Optical Engineering Press, Bellingham WA, 1993).

[2] Middleton, W. E. K., Vision Through The Atmosphere (Univ. Of Toronto Press, Toronto CA, 1952).

[3] Huschke, R. E., "Atmospheric Visual and Infrared Transmission Deduced from Surface Weather Observations: Weather and Warplanes IV", Rand Corporation, Santa Monica CA (1976).

[4] Department of the Army, "Chemical Smoke Generator Units and Smoke Operations", Field Manual FM 3-50 (April 1967).

[5] Milham, Merrill, "A Catalog of Optical Extinction Data for Various Aerosols/Smokes", Edgewood Arsenal Special Publication ED-SP-77002 (June 1976).

[6] DOD-HDBK-178, "Quantitative Description of Obscuration Factors for Electro-Optical and Millimeter Wave Systems", (25 July 1986).

[7] Pasquill, F., "The Estimation of the Dispersion of Windborne Material", *Meteorological Magazine*, 90, #1063, 33-49 (February 1961).

[8] Pasquill, F., Atmospheric Diffusion 2nd Ed. (John Wiley & Sons, New York NY, 1974).

[9] Turner, D. Bruce, "A Diffusion Model for an Urban Area", *J. Applied Meteorology*, 3, 83-91 (February 1964).

[10] Skolnik, Merrill I., Introduction to Radar Systems 2nd Ed. (McGraw-Hill Book Co., New York NY, 1980), pp.442-447.

[11] Data extracted from anonymous briefing to author from Verac Corp. (circa 1983).

[12] United States Geological Survey, "Fort Irwin Quadrangle, California - San Bernardino Co., 7.5 Minute Series (Topographic)" (1986).

[13] Olson, Warren K., (Ed.), Military Operations Research Analyst's Handbook, Vol. I., (Military Operations Research Society, Alexandria VA, 1994).

[14] Watson, R.D., et al, "Prediction of the Fraunhofer Line Detectivity of Luminescent Materials," Proceedings of the Ninth Symposium on Remote Sensing of Environment, Environmental Research Institute of Michigan, Ann Arbor MI (April (1974), p. 1969.

[15] Suits, Gwynn H., “Natural Sources”, Chapter 3 in Wolfe, William L. and Zissis, George J., (Eds.), The Infrared Handbook, Revised Edition, (Environmental Research Institute of Michigan, Ann Arbor MI, 1985), p. 3-111.

[16] Tanimoto, Douglas H., “Promising Applications of High Energy Lasers: DARPA Perspective”, Electro-Optical Systems and Technology Conference, Boston MA (17 November 1980).

Problems

5-1. A manufacturer claims that his new sensor achieves a contrast of 0.5 for a standard target against the background. The background intensity was measured to be 2 units. What can be said about the intensity of the signal from the standard target?

5-2. A second manufacturer claims that his new sensor is indistinguishable from that of the first manufacturer. However, this second manufacturer measures a contrast of 0.25 under exactly the same conditions. Can he be correct in his claims? If both manufacturers are completely correct in their claims, what must the target intensity be?

5-3. A smoke has a visible mass extinction coefficient of 1.0 m^2/g. The same smoke has an infrared mass extinction coefficient of 0.25 m^2/g. If this smoke is used in a smoke screen with a *CL* value twice the minimum needed for an effective visible smoke screen, what is the two- way extinction (out and back through the screen) that this screen provides in the infrared.

5-4. The sun is midway between the horizon and the zenith. Cloud cover is 4/10. Wind speed is 9 knots. What is the Pasquill stability class? Is this condition favorable for employing smoke?

5-5 There is no rain but the relative humidity is approaching 90%. Smoke from a burning vehicle rises roughly 50 meters before it abruptly stops rising and slowly drifts laterally. Are the conditions good for employing smoke?

5-6. A new countermeasure smoke consists of gold-plated microballoons (hollow spheres) with radii randomly distributed between 0.5mm and 1.5mm. If the scattering coefficient of a cloud of this smoke is 100 km^{-1} at 300 GHz, what scattering coefficient do you estimate the smoke to have

- at 1 μm?
- at 10 μm?
- at 30 GHz?
- at 3 GHz?

5-7. Describe the characteristics of Mie scattering and quantitatively discuss the impact of scattering of scattering on attenuation due to fog, rain, dust/smoke, and chaff.

5-8. A laser radar has a beamwidth of 100 μrad and is located 10 meters above a flat plain. It is attempting to detect and track incoming cruise missiles also at 10 meters altitude and nominal 10 km range. Will multipath fading significantly affect the performance of this system? Why or why not?

5-9. Over a spread of ranges of interest a particular radar has a CNR which falls 13-25 dB below that needed to reliably detect a class of inbound threats. Might multipath have some significant effect on increasing the detectability of those threats? Would your answer change if the CNR shortfall was only 3-15 dB?

5-10. Plot the probability of clear line of sight to the ground as a function of platform altitude for a ground range of 10 km. Use the values a = -0.0003, b = 1.658, and c = -1.315 to define the terrain.

5-11. Identify the two wavelength regions of the electromagnetic spectrum which propagate well in sea water.

5-12. A blue-green laser communication system can tolerate a total extinction of 100 dB in the direct beam. For the clearest oceanic water, how deep can this system transmit information?

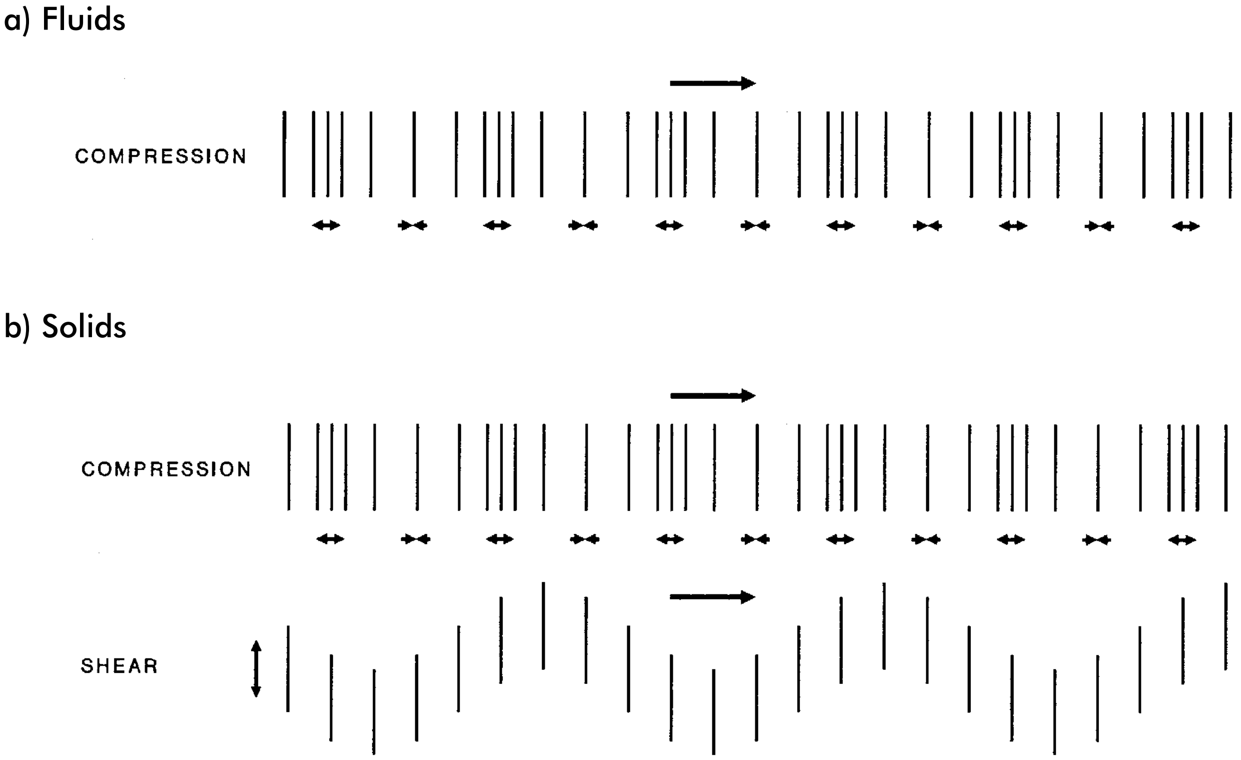
CHAPTER 6

PROPAGATION OF ACOUSTIC RADIATION

What is Acoustic Radiation?

Acoustic radiation is composed of stress waves in a fluid or a solid. In a fluid compression and rarefaction are the only mechanical stresses allowed. Acoustic radiation in a fluid will be composed of pressure waves, as shown in Figure 6-1a. However, in solid materials, shear stresses are also possible. Thus, acoustic radiation in a solid can have shear wave components as well as pressure wave components, as shown in Figure 6-1b. Seismic waves, which are really nothing but earth motion-produced acoustic waves propagating through the ground have readily apparent shear and pressure components, that manifest themselves as the "s" and "p" wave components, respectively, of the seismic waves seen in seismometer recordings of earthquakes. In the following we will consider only pressure waves, as most applications have air or water as the propagation medium.

Figure 6-1. Types of stress waves comprising acoustic radiation.



The existence and properties of acoustic pressure waves comes from consideration of three fundamental equations of materials: the equation of motion, the continuity equation, and the equation of state. The equation of motion is essentially a statement of Newton's second law ($F = ma$) of motion. The continuity equation is an expression of conservation of mass in the system. The equation of state is the relation between density and pressure in the material.

The lossless (no viscous damping) equation of motion of a fluid is [1]

$$-\nabla p = \rho_0 \frac{\partial \vec{v}}{\partial t} \tag{6.1}$$

where p is the total pressure including ambient and wave pressures, ρ_0 is the ambient density, and $\vec{v}$ is the velocity of the fluid. The pressure gradient ($-\nabla p$) is the net force acting on a volume of fluid of unit area and unit thickness. This force causes the mass (ρ_0 times the unit volume) to accelerate with acceleration ($\partial v/\partial t$). The continuity equation for the fluid is given by

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \left(\rho_0 \vec{v}\right) = 0\,. \tag{6.2}$$

The continuity equation states that any change in density must be accompanied by a flow of material into or out of the volume under consideration.

The equation of state describes the pressure as a function of density

$$p = p(\rho)\,. \tag{6.3}$$

Using the chain rule for derivatives, it is easy to show that in the adiabatic limit (no heat transfer into or out of the material), we obtain

$$\frac{\partial p}{\partial t} = \left.\frac{\partial p}{\partial \rho}\right|_{\rho_0} \cdot \frac{\partial \rho}{\partial t} \tag{6.4}$$

Defining a quantity c by the relation

$$c^2 = \left.\frac{\partial p}{\partial \rho}\right|_{\rho_0} \tag{6.5}$$

we may substitute Eq.(6.4) into Eq.(6.2) to yield

$$\frac{1}{c^2}\frac{\partial p}{\partial t} + \rho_0 \nabla \cdot \vec{v} = 0 . \tag{6.6}$$

Taking the time derivative of both sides of Eq.(6.6) yields

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} + \rho_0 \nabla \cdot \frac{\partial \vec{v}}{\partial t} = 0 . \tag{6.7}$$

Substitution of Eq.(6.1) into this result yields

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} - \nabla \cdot \nabla p = 0 \tag{6.8}$$

or

$$\nabla^2 p - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0 . \tag{6.9}$$

Equation (6.9) has the classical form of a wave equation with *c* being interpreted as the wave propagation velocity.

Equation (6.9) differs from the electromagnetic wave equation only in that the pressure is a scalar quantity while the electric and magnetic fields are vector quantities. As a consequence, pressure waves do not exhibit polarization. Otherwise, the same kinds of waves that satisfy the electromagnetic wave equation will satisfy the acoustic wave equation. Plane wave solutions of the form

$$p(\vec{r},t) = p_0 \, e^{i\omega t \mp i\left(\vec{k}_0 \cdot \vec{r}\right)} \tag{6.10}$$

where *p* is the (acoustic) pressure deviation from the ambient hydrostatic pressure p_A (the total pressure is the sum of the acoustic pressure and the hydrostatic pressure) and

$$\omega = c\left|\vec{k}_0\right| = ck_0 = 2\pi c / \lambda \tag{6.11}$$

are permitted. These solutions oscillate sinusoidally in both space and time with temporal frequency $\omega/2\pi$ and free-space wavelength λ. Spherical waves of the form

$$p\left(r,t\right) = p_0 \frac{e^{i\omega t \mp ikr}}{r} \tag{6.12}$$

are also viable solutions to the acoustic wave equation.

Associated with any acoustic pressure level *p* is an acoustic intensity *I* defined by the equation

$$I = \frac{p_0^2}{2\rho_0 c} . \tag{6.13}$$

Both acoustic pressure and acoustic intensity are characterized in units of decibels. That is,

$$I\left(\text{in dB}\right) \equiv 10\log_{10}\left[I\left(\text{in W / m}^2\right) / I_{REF}\right] \tag{6.14a}$$

or

$$p_0\left(\text{in dB}\right) \equiv 20\log_{10}\left[p_0\left(\text{in }\mu\text{Pa}\right) / p_{REF}\right] \tag{6.14b}$$

where I_{REF} is a reference intensity that is produced by a reference sound pressure p_{REF}. In atmospheric acoustics, the reference pressure is p_{REF} = 20 μPa; in underwater acoustics, the reference pressure is p_{REF} = 1 μPa. Because *I* is proportional to the square of p_0, the definition of the decibel by Equations (6.14) means that the value of acoustic intensity **in decibels** is equal to the value of the acoustic pressure **in decibels** for any value of pressure or intensity. That is,

$$I\left(\text{in dB}\right) \equiv p_0\left(\text{in dB}\right) . \tag{6.15}$$

This duality between pressure and intensity makes it extremely convenient to use the logarithmic decibel units rather than the conventional units of Pascals and Watts/square meter.

The Speed of Sound

For an ideal gas we can evaluate the expression for the propagation velocity explicitly. The equation of state for an ideal gas may be written as

$$p = \frac{\rho R T}{M} \tag{6.16}$$

in its isothermal form, where R is the ideal gas constant (= 8.31441 J/mole-K), T is the gas temperature, and M is the molar weight of the gas (= 0.028964 kg for air). In its adiabatic form the equation of state is

$$\frac{p}{p_0} = \left(\frac{\rho}{\rho_0} \right)^{\gamma} \tag{6.17}$$

where

$$\gamma = c_p / c_v \tag{6.18}$$

is the ratio of specific heats (= 1.4 for air and =1.667 for monatomic gases). Taking the density derivative of Eq.(6.17) yields an expression into which subsequent substitution of Eq.(6.16) yields

$$\left. \frac{\partial p}{\partial \rho} \right|_{\rho_0} = \left. p_0 \gamma \frac{(\rho)^{\gamma-1}}{(\rho_0)^{\gamma}} \right|_{\rho_0} = \frac{p_0 \gamma}{\rho_0} = c^2 = \frac{\gamma R T}{M} \tag{6.19}$$

For air, we can easily evaluate Eq.(6.19) to yield

$$c\,(\text{in m / s}) = 20.05\, T^{1/2} \tag{6.20}$$

where T is in K. For $T = 300$ K, we calculate c = 347.3 m/s.

In Eq. (6.20) we have ignored any temperature dependence or gas mixture dependence of the ratio of specific heats. To first order, γ for air may be considered to be a constant and equal to 1.4. To a higher order of fidelity, γ has a weak dependence on temperature, carbon dioxide fraction, and relative humidity. This results because any absorption caused by molecular species must be included in the specific heats (and thus in the ratio of specific heats). If we denote the result from Eq.(6.20) by the symbol c_0, then taking the molecular absorptions into account the speed of sound may be written as [2]

$$\frac{c}{c_0} \cong 1 + \frac{1}{35}\left[X_{N_2} \left(\frac{\theta_{N_2}}{T} \right)^2 e^{-\theta_{N_2}/kT} \frac{\left(\omega / \omega_{N_2} \right)^2}{1 + \left(\omega / \omega_{N_2} \right)^2} \right] + \frac{1}{35}\left[X_{O_2} \left(\frac{\theta_{O_2}}{T} \right)^2 e^{-\theta_{O_2}/kT} \frac{\left(\omega / \omega_{O_2} \right)^2}{1 + \left(\omega / \omega_{O_2} \right)^2} \right] \tag{6.21}$$

where the terms X_x, θ_x, and ω_x are defined in the following section on absorption of sound waves. In most cases we may neglect the more complicated behavior of the speed of sound in air and use the simple result of Eq.(6.20).

The equation of state for sea water is considerably more complex than that of an ideal gas. Direct evaluation of Eq.(6.5) is therefore impractical. Empirical measurements, however, have yielded the following results for the speed of sound in sea water [3]

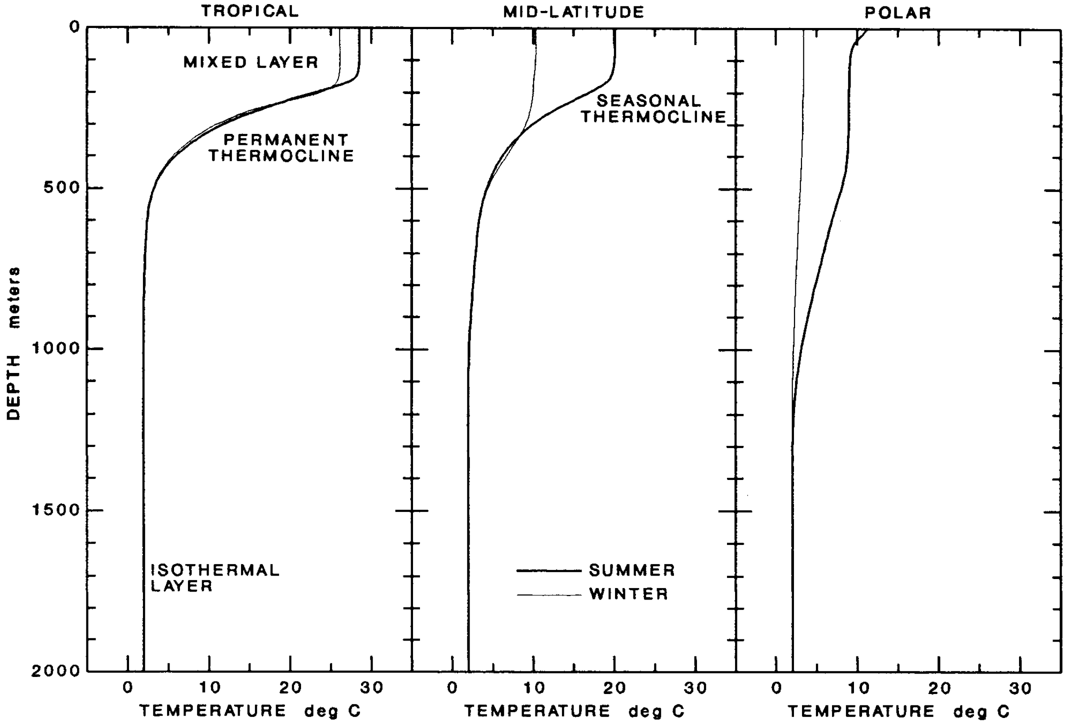
$$\begin{aligned} c = {} & 1402.06 + 1.340S - 0.01025TS \\ & + 4.950T - 0.05304T^2 + 2.374 \times 10^{-4} T^3 \\ & + 0.0163z + 1.675 \times 10^{-7} z^2 - 7.139 \times 10^{-13} Tz^3 \end{aligned} \tag{6.22}$$

where c is the speed of sound in m/s, T is the temperature in °C, S is the salinity in parts per thousand (which are equivalent to the so-called **practical salinity units** (psu)), and z is the depth in m.

Both temperature and salinity are highly variable with respect to location as well as depth in the oceans of the world. Figure 6-2 illustrates typical seasonal variations in temperature vs. depth at three different latitudes. Sea surface temperatures range from as high as 30-35 °C at tropical latitudes to -1 or -2 °C in Arctic and Antarctic waters. At very great depths (below 1000 m) the temperature tends to be more uniform and decreases very slowly from roughly 3-4 °C to as low as 0.5-2.5 °C depending on location. The temperatures tend to be more or less constant in a mixed layer from the surface to depths of the order of 100-200 m and then decrease with increasing depth until the nominal isothermal layer is reached. At depths below 100-200 m there is generally a decrease in temperature, known as a thermocline. The portion of the thermocline that varies seasonally is called the seasonal thermocline. At greater depths (typically below 400-600 m) the seasonal variations tend to disappear and the temperature decrease is called the permanent thermocline.

The inflow of water from large rivers will create plumes of usually warmer water that may extend hundreds of miles into the open ocean. The interaction of ocean currents with the oceanic topography and with each other complicates the normal behavior of temperature versus depth. Cold water can be forced up from great depths to mix with warmer surface waters. Warm currents such

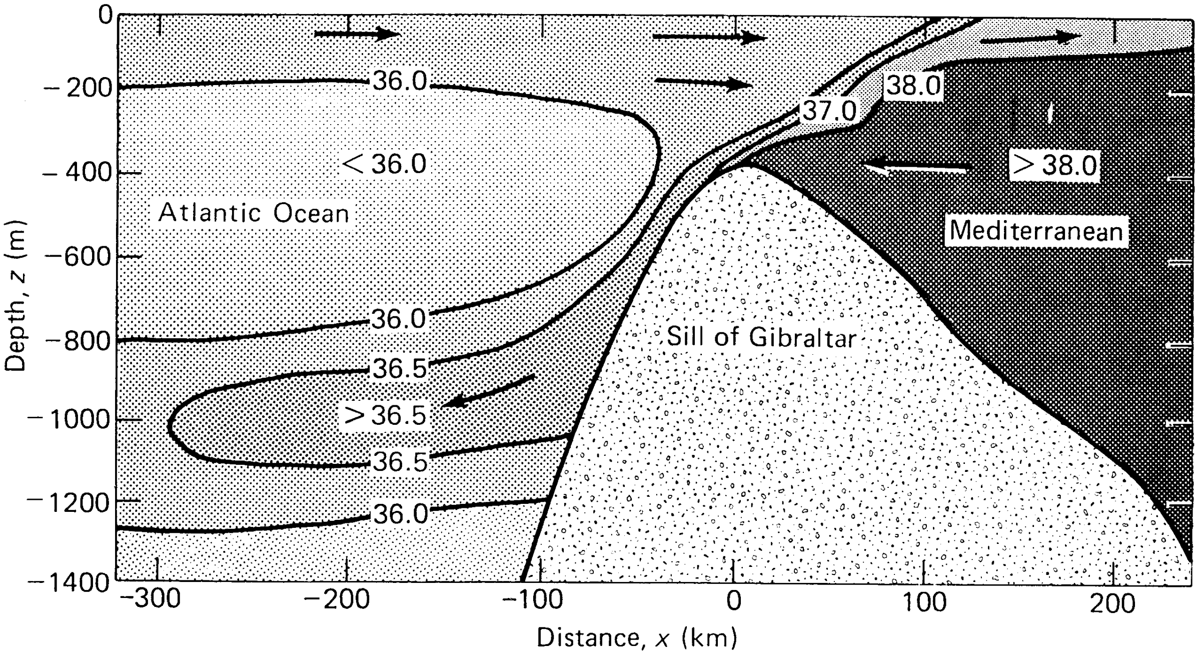
Figure 6-2. Temperature variation with depth at different latitudes.



as the Gulf Stream will form warm and cold eddies as they move past colder waters. Thus, there are many ways in which it is possible for layers of warmer water to exist between layers of colder water.

The salinity of the ocean is as variable as the temperature. The average salinity at the surface of the North Atlantic is 36.5 psu, while it is only 35.0 in the Pacific and Indian Oceans, and as low as 33 psu in the Arctic Ocean. In the Eastern Mediterranean Sea, which has limited freshwater inflow and severe evaporation, the salinity can reach as high as 39.1 psu. Evaporation significantly increases the salinity at the surface of the ocean. If the Straight of Gibraltar were to become blocked, it is estimated that the Mediterranean would evaporate completely within 1000 years. Long before it completely dried up, the salinity would exceed 300 psu, the salinity of the Dead Sea and other salt lakes. Precipitation can significantly reduce salinity. For example, 1 cm of precipitation mixed with the upper 10 m of seawater (the amount that might reasonably be mixed on time scales of a few hours in moderate Sea States) will reduce the salinity by 1 psu. Freshwater has extremely low salinity (by definition less than 3 psu). High flow rivers such as the Mississippi River and the Amazon River can produce plumes of freshwater that extend for hundreds of kilometers before they are thoroughly mixed. Surface and subsurface currents carrying water from high salinity regions to low salinity regions, and vice versa, cause a fairly complex variation of salinity with depth. For example, Figure 6-3 shows the undersea saline “waterfall” pouring over the sill of Gibraltar from the highly saline Mediterranean Sea. As shown in Figure 6-4, this outflow extends almost to the East

Figure 6-3. The outflow of supersaline Mediterranean Sea waters through the Straights of Gibraltar.[4],[5]

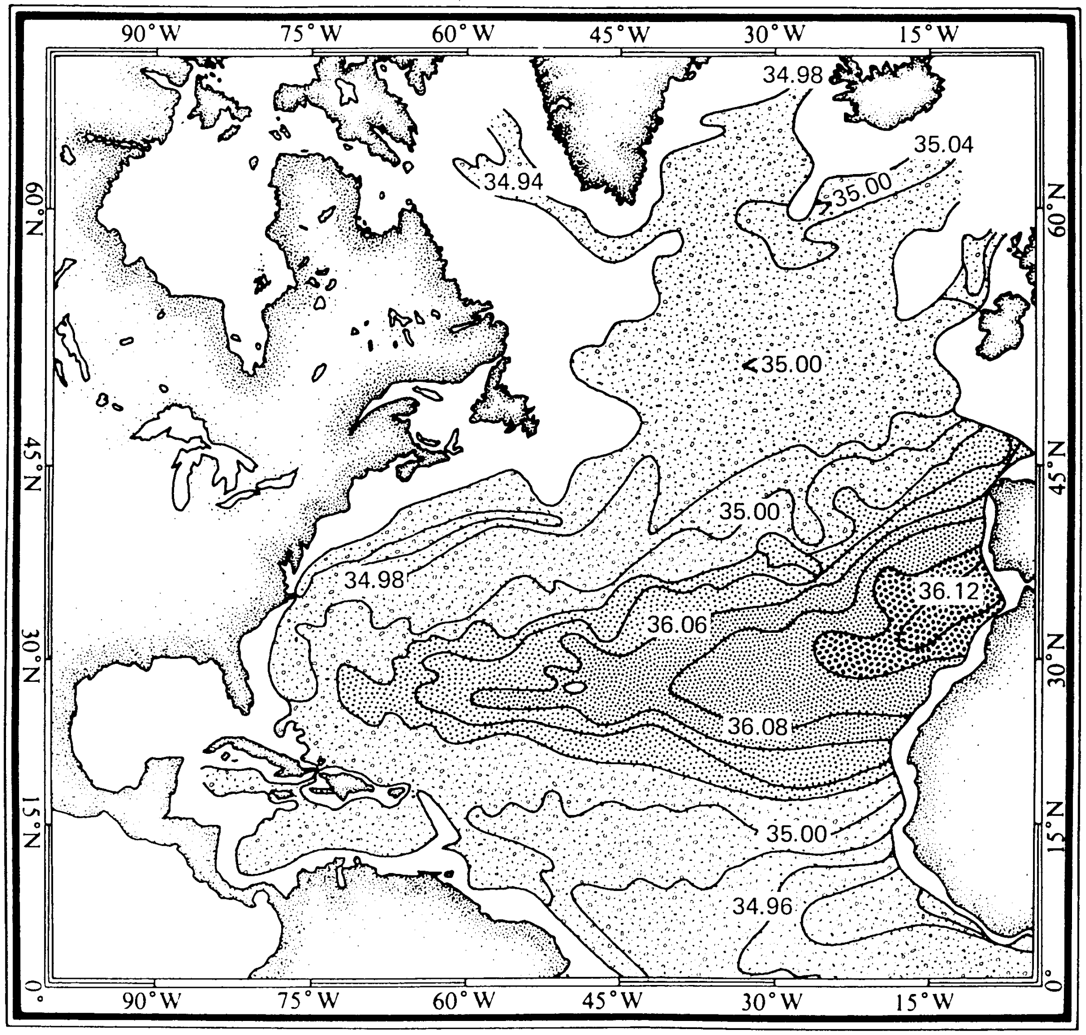


coast of Florida (the Gulf Stream blocks its continued westerly flow).

The various temperature gradients, thermal inversion layers, salinity gradients, supersaline and freshwater plumes produce inhomogeneities in the speed of sound that complicate the propagation of acoustic radiation. These speed of sound variations define the nature of acoustic propagation in the underwater environment. The peculiarities of underwater sound propagation will be discussed in a subsequent section.

Similar stratified variations in humidity and temperature affect the speed the sound in air. These variations will cause atmospheric acoustic effects almost identical to those that are observed in underwater acoustic propagation.

Figure 6-4. Spreading of Mediterranean Sea salinity across the North Atlantic ocean waters at intermediate depths.[4],[6]



Attenuation of Acoustic Radiation

As pressure waves propagate through the air or water medium, a number of factors work to attenuate the pressure levels (or intensities) in those waves. We will define the attenuation coefficient α (in nepers/m) or a (in dB/m) by the intensity relation

$$I(R) = I(R_0)\left(\frac{R_0}{R}\right)^2 e^{-\alpha(R-R_0)} = I(R_0)\left(\frac{R_0}{R}\right)^2 10^{-0.1a(R-R_0)} \qquad (6.23)$$

which also incorporates the spherical wave spreading factor ($1/R^2$). This is a common but not universally accepted definition. Using this definition, the pressure attenuation expression is

$$p(R) = p(R_0)\left(\frac{R_0}{R}\right)e^{-0.5\alpha(R-R_0)} = p(R_0)\left(\frac{R_0}{R}\right)10^{-0.05a(R-R_0)} . \qquad (6.24)$$

The exponential and logarithmic forms of the attenuation coefficients are related by

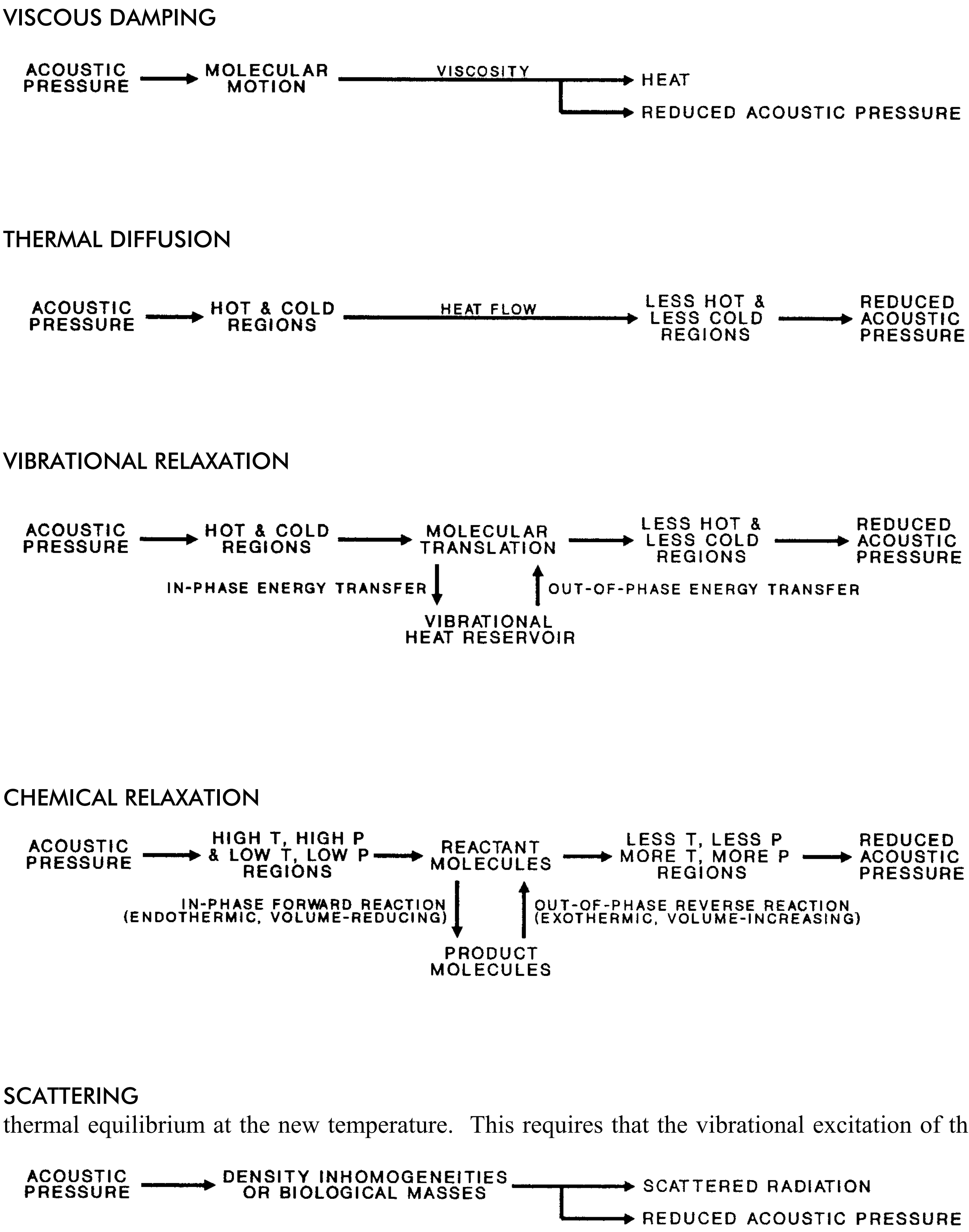
$$a = 4.3430\alpha \qquad \text{and} \qquad \alpha = 0.2303a . \qquad (6.25)$$

Figure 6-5 describes the various factors that contribute to acoustic attenuation. Every fluid has a viscosity. Viscosity attenuates acoustic radiation in the following way. Pressure waves consist of alternating high pressure and low pressure regions. As the wave propagates, the fluid molecules must move out of the average pressure regions into the high pressure regions and out of the low pressure regions into the average pressure regions. Once the wave motion has established the high and low pressure regions, the molecules will move under the pressure gradients back towards their original positions to reestablish pressure equilibrium. Viscosity resists molecular motion. Thus, the wave energy associated with the molecular motion is reduced (converted to heat) by viscosity. This reduces the pressure differences that can be established – the fundamental definition of acoustic attenuation.

Thermal diffusion acts to attenuate acoustic radiation in the following fashion. When the pressure wave produces high and low pressure regions, the high pressure regions will become warmer and the low pressure regions will become colder. Thermal diffusion acts to move energy from warm regions to cold regions. As energy diffuses out of the warmer regions, they cool and the pressure will be reduced. As energy diffuses into the cooler regions, they warm up and the pressure will be increased. Thermal diffusion acts to reduce the pressure in the high pressure regions and increase the pressure in the low pressure regions – once again this reduction of pressure differences is the definition of acoustic attenuation.

Vibrations of the fluid molecules can attenuate acoustic radiation in the following fashion. As the pressure in a region increases, it heats up. As it heats up, the molecules will try to establish

Figure 6-5. Mechanisms of attenuation of acoustic radiation.


VISCOUS DAMPING
ACOUSTIC PRESSURE
MOLECULAR MOTION
VISCOSITY
HEAT
REDUCED ACOUSTIC PRESSURE
THERMAL DIFFUSION
ACOUSTIC PRESSURE
HOT & COLD REGIONS
HEAT FLOW
LESS HOT & LESS COLD REGIONS
REDUCED ACOUSTIC PRESSURE
VIBRATIONAL RELAXATION
ACOUSTIC PRESSURE
HOT & COLD REGIONS
MOLECULAR TRANSLATION
LESS HOT & LESS COLD REGIONS
REDUCED ACOUSTIC PRESSURE
IN-PHASE ENERGY TRANSFER
OUT-OF-PHASE ENERGY TRANSFER
VIBRATIONAL HEAT RESERVOIR
CHEMICAL RELAXATION
ACOUSTIC PRESSURE
HIGH T, HIGH P & LOW T, LOW P REGIONS
REACTANT MOLECULES
LESS T, LESS P MORE T, MORE P REGIONS
REDUCED ACOUSTIC PRESSURE
IN-PHASE FORWARD REACTION (ENDOTHERMIC, VOLUME-REDUCING)
OUT-OF-PHASE REVERSE REACTION (EXOTHERMIC, VOLUME-INCREASING)
PRODUCT MOLECULES
SCATTERING
ACOUSTIC PRESSURE
DENSITY INHOMOGENEITIES OR BIOLOGICAL MASSES
SCATTERED RADIATION
REDUCED ACOUSTIC PRESSURE

thermal equilibrium at the new temperature. This requires that the vibrational excitation of the molecules increase in concert with the new Boltzmann factors ($e^{-E/kT}$). However, because the vibrational energies often exceed the thermal energy (kT), vibrational excitation does not occur instantly. It takes a number of collisions before the geometry and kinetic energies of the molecules are just right to cause a vibrational excitation to occur. In the same way, not every collision allows stored vibrational energy to be released. Thus, molecular vibrations slowly absorb energy when the high pressure regions heat up. The energy absorption tends to reduce the peak temperature and thus the peak pressure obtained. When the pressure falls as the wave moves on, the excited molecules do not immediately release their stored energy. They are most likely to release their energy when the temperature is lowest. The release of energy tends to increase the temperature and pressure of the low pressure regions. Thus, the in-phase absorption and out-of-phase release of energy by the vibrations lowers the maximum pressure in the high pressure regions and increases the minimum pressure in the low pressure regions – the fundamental definition of acoustic attenuation.

In a highly analogous fashion, chemical reactions can attenuate acoustic radiation. Consider an ionic reaction in a solution

$$AB \Leftrightarrow A^{+} + B^{-} + \Delta E \tag{6.26}$$

Under normal conditions, the chemical salt AB dissolves into A^{+} and B^{-} ions with the release of a small amount of energy ΔE. At high pressures, the dissolved species are driven to recombine (reactions often act to counteract changes in the environment – fewer species leads to reduced pressure). Formation of the salt causes a pressure reduction by reduction of the number of fluid components as well as by absorbing energy. In regions of lower pressure, the salt will tend to dissolve, but this process is not instantaneous. It is most likely to occur when the pressure is lowest. The dissolution process releases energy, increasing the temperature, and thus increasing the pressure. The process also directly increases the pressure by increasing the number of species. Thus, the in-phase recombination and out-of-phase dissolution of chemical salts lowers the maximum pressure in the high pressure regions and increases the minimum pressure in the low pressure regions – once again, the fundamental definition of acoustic attenuation.

Many entities in the air and ocean environments can scatter acoustic radiation. Density fluctuations, particulates (silt, dust, aerosols, raindrops, etc.), and biological species (insects, plankton, fish, etc.) all act as scatterers of acoustic radiation. However, contrary to the first four processes, scattering does not act by reducing the high pressures and increasing the low pressures. It works by redirecting acoustic energy away the original directions. This direct reduction in energy flow produces the attenuation.

Let us consider first the attenuation of acoustic radiation in air. The ANSI standard ANSI S1.26-1978 gives detailed instructions for calculating the attenuation of acoustic radiation in air.[7] This publication, however, uses a different definition of the attenuation coefficient. Specifically, it defines the coefficient α' based on pressure attenuation

$$p(x) = p(0)e^{-\alpha' x} \quad . \tag{6.27}$$

The ANSI attenuation coefficient is related to our coefficients α and a by the simple relations

$$\alpha = 0.5\alpha' \qquad \text{and} \qquad a = 2.172\,\alpha' . \tag{6.28}$$

Empirically, it has been determined that only viscosity, thermal diffusion, and vibrational relaxation in oxygen and nitrogen are major contributors to absorption in air. The total attenuation coefficient is the sum of the individual contributions

$$\alpha' = \alpha_{VD/TD} + \alpha_{N_2} + \alpha_{O_2} . \tag{6.29}$$

Thermal diffusion and viscosity are usually treated together. Their contribution is given by

$$\alpha_{VD/TD} = 1.84\times 10^{-11}\,\nu^2\,\left(T/T_0\right)^{1/2} / \left(p_A / p_{A0}\right) \tag{6.30}$$

where $\nu = \omega/2\pi$ is the acoustic frequency, $T_0 = 293.15$ K, and $p_{A0} = 101325$ Pa.

The contribution from vibrational relaxation in nitrogen is given by

$$\alpha_{N_2} = \left(\omega/c_0\right)\left(X_{N_2}/35\right)\left(\theta_{N_2}/T\right)^2 e^{-\theta_{N_2}/T}\left[\frac{2\left(\omega/\omega_{N_2}\right)}{1+\left(\omega/\omega_{N_2}\right)^2}\right] \tag{6.31}$$

where the speed of sound is

$$c_0 = 343.23\left(T/T_0\right)^{1/2} , \tag{6.32}$$

the partial fraction of nitrogen is

$$X_{N_2} = 0.781 , \tag{6.33}$$

and the characteristic vibrational temperature for nitrogen is

$$\theta_{N_2} = 3352.2 \text{ K} . \tag{6.34}$$

The characteristic frequency for nitrogen is given by

$$\omega_{N_2} = 2\pi\left(p_A/p_{A0}\right)\left(T/T_0\right)^{-1/2}\left(9 + 350h\, e^{\left\{-6.142\left[\left(T/T_0\right)^{-1/3}-1\right]\right\}}\right) . \tag{6.35}$$

The attenuation produced by molecular oxygen is treated in an analogous fashion, with

$$\alpha_{O_2} = (\omega / c)\left(X_{O_2} / 35\right)\left(\theta_{O_2} / T\right)^2 e^{-\theta_{O_2}/T} \left[\frac{2(\omega / \omega_{O_2})}{1 + \left(\omega / \omega_{O_2}\right)^2} \right] \tag{6.36}$$

where *c* is the same for oxygen as for nitrogen, the partial fraction of oxygen is

$$X_{O_2} = 0.209 \ , \tag{6.37}$$

and the characteristic vibrational temperature for oxygen is

$$\theta_{O_2} = 2239.1 \text{ K} . \tag{6.38}$$

The characteristic frequency for oxygen is given by

$$\omega_{O_2} = 2\pi(p_A / p_{A0})\left\{24 + 4.41 \times 10^4 h \left[\frac{0.05 + h}{0.391 + h}\right]\right\} \tag{6.39}$$

where *h* is the molar concentration of water vapor in percent. That is,

$$\begin{aligned} h &= \left(\frac{\rho_{H_2O}(\text{in g} / \text{m}^3)}{18.01528(\text{g} / \text{mole})} \right) \left(\frac{100(\text{percent})}{44.6148(\text{mole} / \text{m}^3)} \right) \left(\frac{p_A}{p_{A0}} \right)^{-1} \left(\frac{T}{T_0} \right) \\ &= 0.1244 \, \rho_{H_2O} \left(\frac{p_A}{p_{A0}} \right)^{-1} \left(\frac{T}{T_0} \right) \end{aligned} \tag{6.40}$$

where is ρ_{H20} is the absolute humidity (in g/m^3) at atmospheric pressure p_A and temperature *T*.

Combining all of the contributions yields

$$\begin{aligned} \alpha' = (\omega / 2\pi)^2 \Big\{ &\left[1.84 \times 10^{-11} (p_A / p_{A0})^{-1} (T / T_0)^{1/2} \right] \\ &+ (T / T_0)^{-5/2} \left\{ 0.08030 \left[e^{-2239.1/T} \right] / \left[\omega_{O_2} + \left(\omega^2 / \omega_{O_2} \right) \right] \right\} \\ &+ (T / T_0)^{-5/2} \left\{ 0.6710 \left[e^{-3352.2/T} \right] / \left[\omega_{N_2} + \left(\omega^2 / \omega_{N_2} \right) \right] \right\} \Big\} \end{aligned} \tag{6.41}$$

For typical normal values of temperature, pressure, and humidity, the attenuation as a function of frequency is shown in Figure 6-6. Figures 6-7 and 6-8 show the effects of humidity on acoustic attenuation in the atmosphere. These curves were calculated using the equations above with the attenuation coefficient being transformed to the standard version used throughout the remainder of this text using Eq. (6.28).

Figure 6-6. Attenuation of acoustic radiation in air.

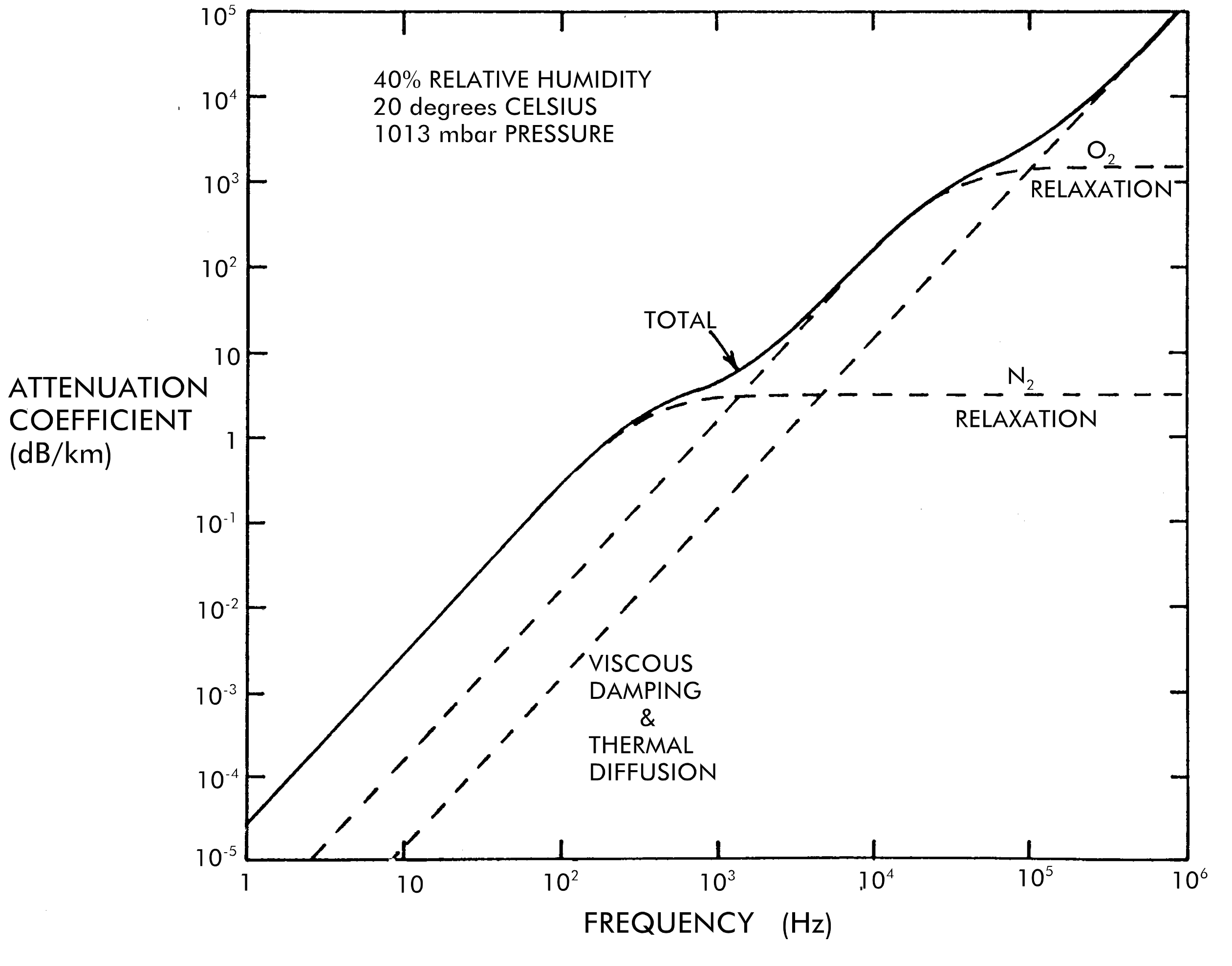


Figure 6-7. Effects of atmospheric humidity on acoustic attenuation in air at 0°C.

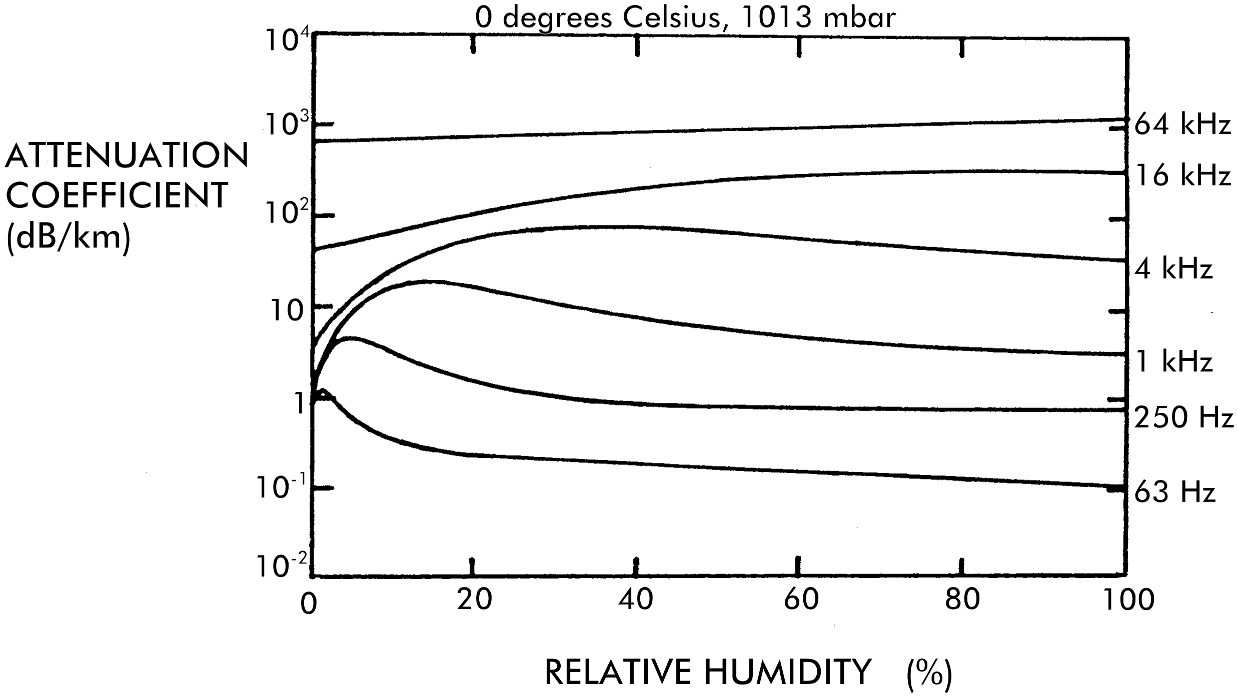
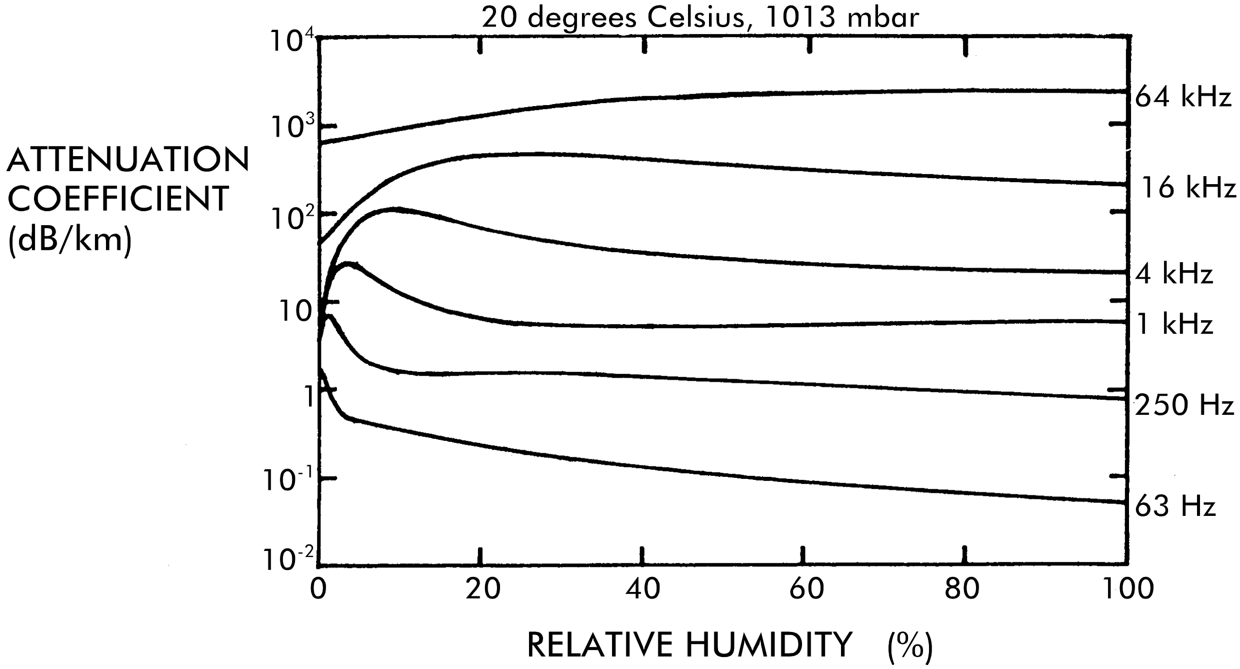


Figure 6-8. Effects of atmospheric humidity on acoustic attenuation in air at 20°C.



Several things should be apparent from an examination of Figure 6-6. First, high frequencies have higher attenuations than low frequencies. This means that if acoustic waves are to be propagated very long distances through the atmosphere, then those waves must have relatively low frequencies. The low frequency "rumble" of thunder can be heard for many miles; the high frequency "crack" can only be heard at short distances. The nominal break point in frequency occurs at 1 kHz. At this frequency, the attenuation coefficient is roughly 5 dB/km. At higher frequencies, propagation for a few kilometers will produce attenuations of tens of dB. At lower frequencies, propagation for a few kilometers produces relatively small attenuations. Thus, long range detection can only be done at frequencies below 1 kHz. If it is desired to make long range detection difficult, then the system should strive to produce only very high frequency sound and any low frequencies should be damped to the maximum extent possible.

The acoustic attenuation coefficient in sea water has major contributions from viscosity, chemical relaxation, and scattering.[4] The total attenuation coefficient is given

$$\alpha = \alpha_{\text{Scattering}} + \alpha_{\text{Boric Acid}} + \alpha_{\text{Magnesium Sulfate}} + \alpha_{\text{Viscosity}} \tag{6.42}$$

The scattering component is essentially independent of acoustic frequency. It has a value of

$$\alpha_{\text{Scattering}} = 6.9 \times 10^{-7} \quad m^{-1} \tag{6.43}$$

The boric acid relaxation term involves the chemical reaction

$$\mathrm{H_3BO_3} \Leftrightarrow \mathrm{H^+} + \mathrm{H_2BO_3^-} . \tag{6.44}$$

The relaxation contribution to the attenuation has the form

$$\alpha_{\text{Boric Acid}} = \frac{A_2}{t_2} \frac{\left(\omega t_2\right)^2}{1+\left(\omega t_2\right)^2} \quad m^{-1} \tag{6.45}$$

where the coefficient A_2 is approximately

$$A_2 = 3.269 \times 10^{-9} \quad s - m^{-1} \tag{6.46}$$

and the relaxation time constant is

$$\frac{1}{t_2} = 3.8\left(\frac{S}{35}\right)^{1/2} 10^{\left(7-\frac{1051}{T+273.15}\right)} \quad s^{-1} \tag{6.47}$$

The magnesium sulfate relaxation involves the chemical reaction

$$\mathrm{MgSO_4} \Leftrightarrow \mathrm{Mg^{2+}} + \mathrm{SO_4^{2-}}. \tag{6.48}$$

The relaxation contribution to the attenuation has the form

$$\alpha_{\text{Magnesium Sulfate}} = \frac{A_3\, S(s)\, P(p)}{t_3} \frac{(\omega t_3)^2}{1+(\omega t_3)^2} \quad m^{-1} \tag{6.49}$$

where the coefficient *A_3 S(s),* which contains an explicit salinity dependence *S(s)* is given by

$$A_3 S(s) = s \cdot 3.73 \times 10^{-7}, \tag{6.50}$$

the relaxation time constant is given by

$$\frac{1}{t_3} = 1.38 \times 10^{\left(11 - \frac{1520}{T+273.15}\right)} \quad s^{-1}, \tag{6.51}$$

and the pressure dependence is given by

$$P(p) = 1 - 6.46 \times 10^{-9}\, p_A \tag{6.52}$$

where p_A is the ambient hydrostatic pressure in Pascals.

The viscosity contribution to the total acoustic attenuation takes the form

$$\alpha_{\text{Viscosity}} = \frac{8}{3} \frac{\mu \omega^2}{\rho c^3} P(p) \omega_3 t_3 \quad m^{-1} \tag{6.53}$$

where ρ is the density of the seawater, *c* is the speed of sound, μ is the effective viscosity and is given by

$$\mu = 2 \times 10^{-3} \quad kg - m^{-1} - s^{-1} \tag{6.54}$$

and

$$\omega_3 = 9 \times 10^5 \quad s^{-1}. \tag{6.55}$$

Figure 6-9 shows the total acoustic attenuation in sea water and the contributions to the attenuation from all of the major sources. As in propagation in the atmosphere, acoustic propagation in sea water favors low frequencies. The nominal cutoff frequency in sea water is 20 kHz where the attenuation coefficient again is of the order of 5 dB/km. Frequencies below 20 kHz can be detected at ranges in excess of a few kilometers (if other propagation effects permit), while frequencies above 20 kHz will be severely attenuated after a few kilometers of propagation. The lowest frequencies (1-100 Hz) can propagate halfway around the world with substantial but tolerable attenuations.

Figure 6-9. Acoustic attenuation of sea water.

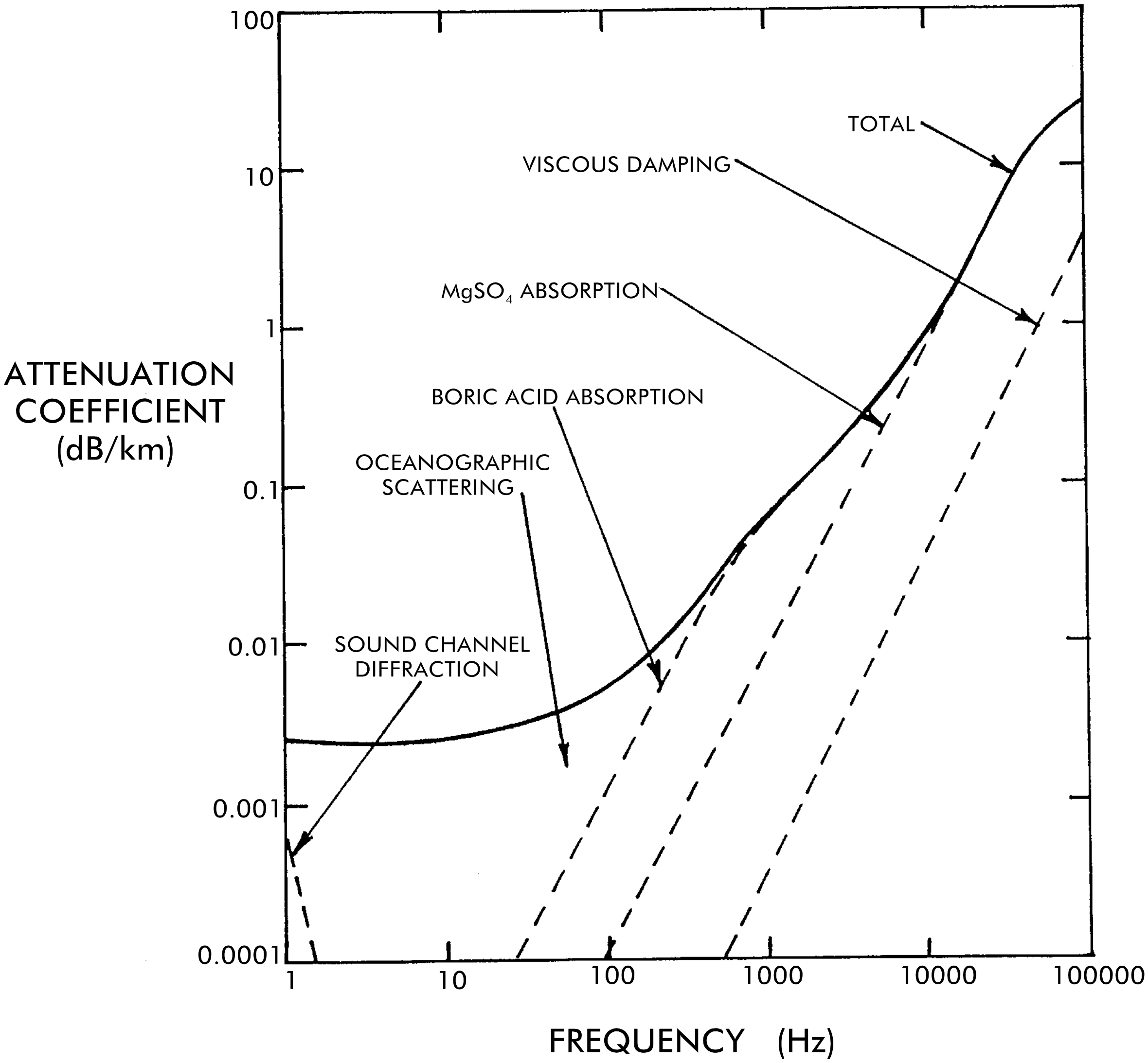
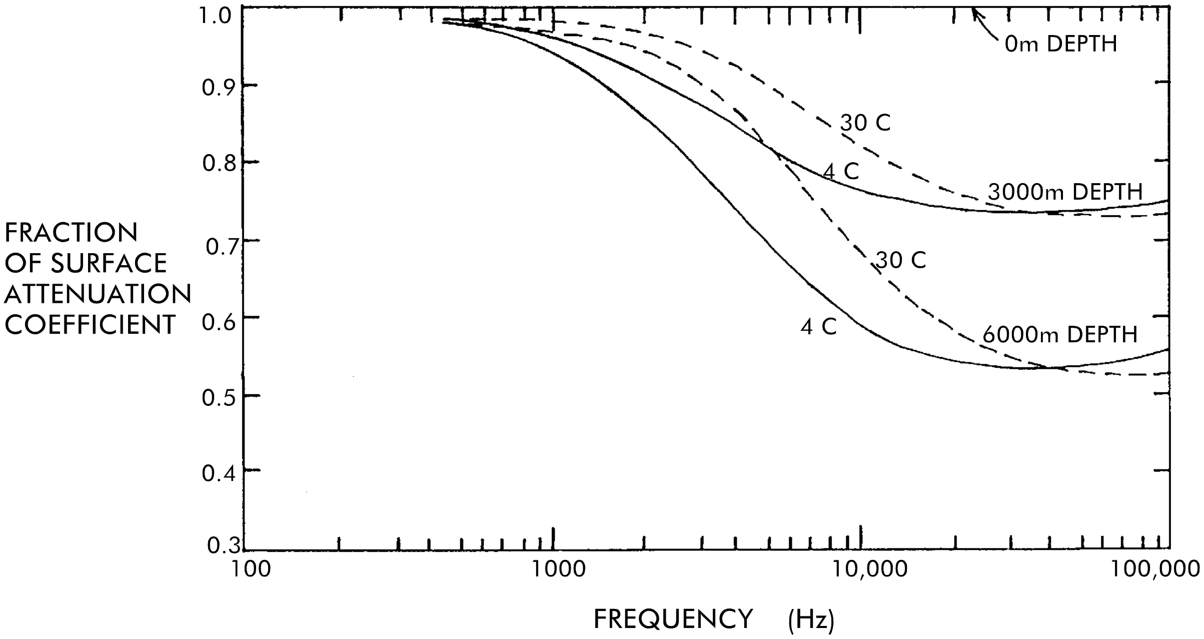


Figure 6-10. Effects of temperature and depth on acoustic attenuation in sea water.



Reflection and Refraction of Acoustic Radiation

As evidenced in Eqs. (6.20) and (6.22), the speeds of sound in air and the ocean are both functions of temperature. The speed of sound in the ocean is also a function of salinity and depth. Temperature is a highly variable function of altitude in the atmosphere, and both temperature and salinity vary considerably with depth in the ocean. Thus, non-zero vertical gradients in the speed of sound are expected. By analogy with the speed of light in the atmosphere, which also exhibits strong vertical gradients (Chapter 4), we expect anomalous refractive propagation of acoustic radiation in both the atmosphere and in the ocean.

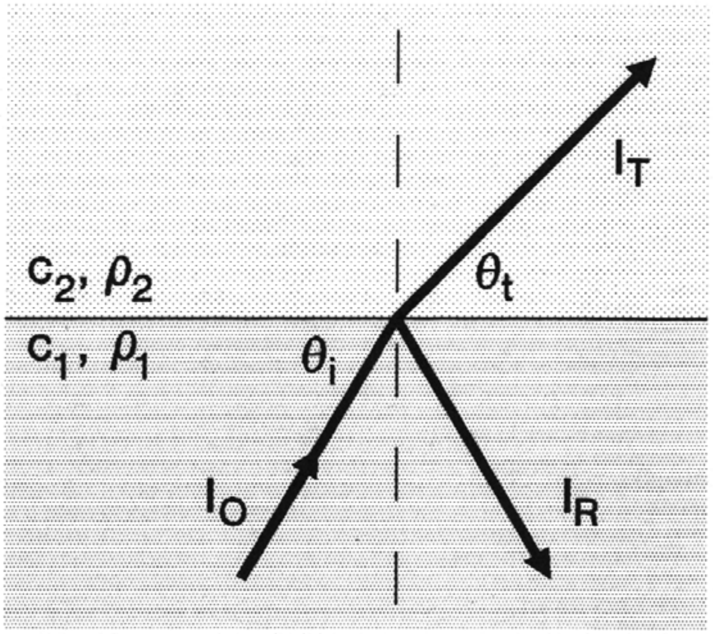
Before discussing anomalous behavior, we should review the effects of normal reflection and refraction at an interface between two acoustic media. The first medium is characterized by speed of sound c_1 and density ρ_1 while the second medium is characterized by speed c_2 and density ρ_2. For convenience, we may define an acoustic "refractive index" η by

$$\eta = c_{REF} / c \qquad (6.56)$$

where c_{REF} is an arbitrarily chosen velocity. Thus we may define a relative refractive index n at the interface by

$$n = c_1 / c_2 = \eta_2 / \eta_1 . \qquad (6.57)$$

Figure 6-11. Reflection and refraction of acoustic radiation at an interface.



We need also define a density mismatch factor *m* by

$$m = \rho_2 / \rho_1 . \tag{6.58}$$

A ray of acoustic energy incident on the interface at an angle θ_i will be bent exiting the interface at an angle θ_t, with the angles related by the refractive indices

$$\frac{\cos\theta_i}{c_1} = \frac{\cos\theta_t}{c_2} \quad \Leftrightarrow \quad \eta_1 \cos\theta_i = \eta_2 \cos\theta_t \quad . \tag{6.59}$$

Eq. (6.59) is the acoustic equivalent of Snell's Law. Note the definition of angles used in Urick's presentation (on which our own is modeled) is just the opposite of that used in electromagnetism. If the electromagnetic definition (angle relative to the surface normal) were used, then the cosines in Eq.(6.59) would be replaced by sines and the expression would become identical to Snell's Law (Eq.(4.6)).

The incident acoustic radiation is not just refracted, a portion of the radiation is reflected. The fraction of radiation that is reflected is given by the expression

$$\frac{I_R}{I_0} = \left[\frac{m\sin\theta_i - \left[n^2 - \cos^2\theta_i\right]^{1/2}}{m\sin\theta_i + \left[n^2 - \cos^2\theta_i\right]^{1/2}} \right]^2 . \tag{6.60}$$

This expression exhibits different behavior depending on the relative values of *m* and *n.* These behaviors are shown in Figure 6-12. There are two special angles that become apparent. The first of these is "Brewster's Angle"

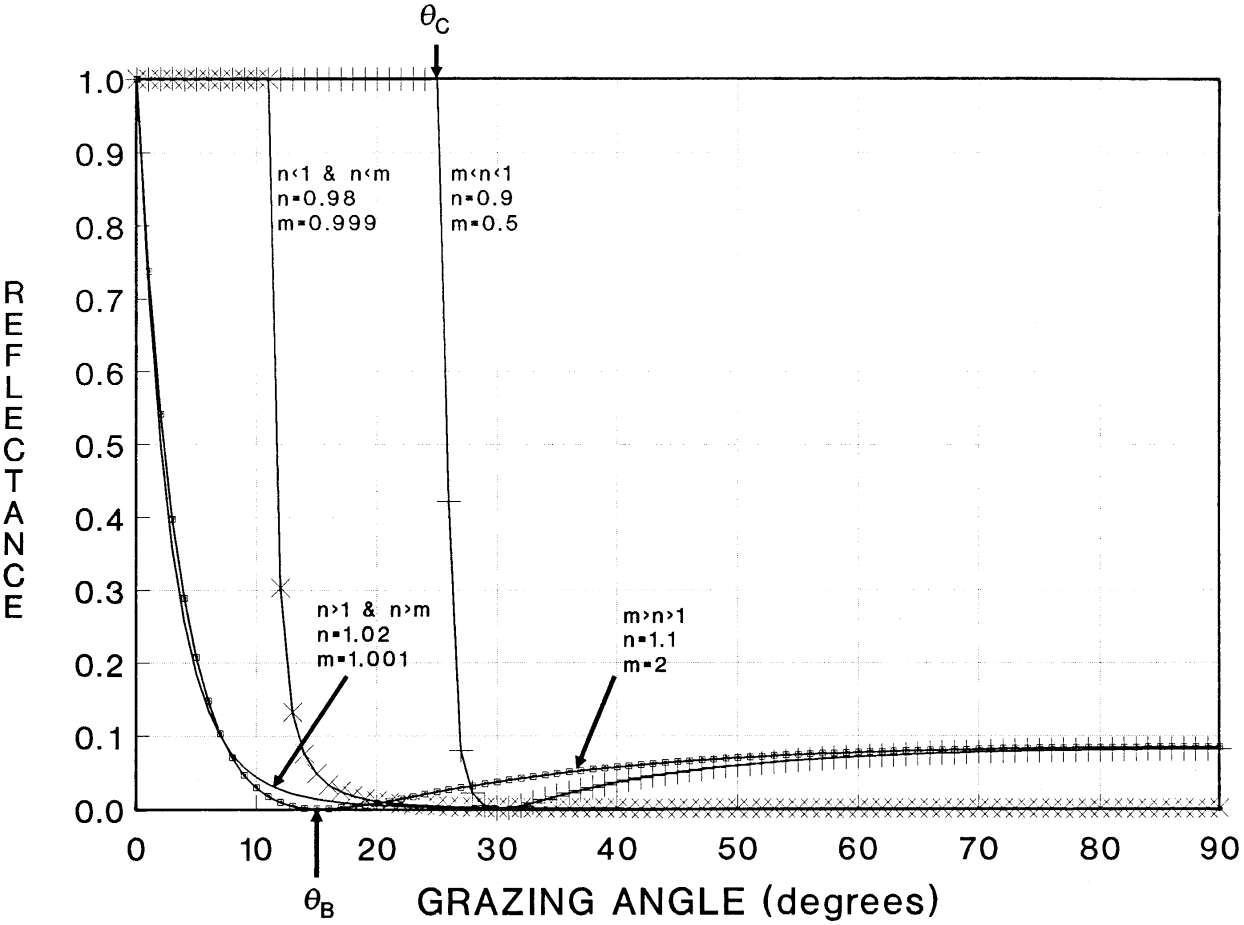
$$\theta_B \equiv \cos^{-1}\left[\frac{m^2 - n^2}{m^2 - 1} \right]^{1/2} \quad , \tag{6.61}$$

the angle at which the acoustic energy is completely transmitted through the interface with no reflected component. The other angle is the "critical angle"

$$\theta_C \equiv \cos^{-1} n , \tag{6.62}$$

the angle at which the acoustic energy is completely reflected from the interface with no transmitted component. Both angles are indicated in the figure (for the values of *m* and *n* used in plotting the figure). In either air or water, if acoustic radiation encounters a layer with strongly different refractive index, the ray will be strongly reflected. Such layers may be used to hide from sonar

Figure 6-12. Reflectance versus incidence angle for several regimes of relative refractive index and density mismatch.



systems.

Given the above expressions, let us consider atmospheric propagation first. The vertical gradient of refractive index is given by

$$\frac{d\eta}{dh} = \frac{d\eta}{dT} \cdot \frac{dT}{dh} . \tag{6.63}$$

Substituting Eq.(6.20) into Eq.(6.56) and differentiating yields

$$\frac{d\eta}{dT} = c_{REF} \frac{d\left(20.05T^{1/2}\right)^{-1}}{dT} = \frac{c_{REF}}{20.05}\left(-\frac{1}{2}\right)T^{-3/2} = -\frac{\eta}{2T} . \tag{6.64}$$

From Chapter 3, we recall that the normal lapse rate of the atmosphere is approximately 0.0064 K/m. Thus, the normal atmospheric gradient of sound speed is

$$\frac{d\eta}{dh} = \left(-\frac{\eta}{2T}\right)\left(-0.0064\right) = \frac{+0.0032\eta}{T} \quad . \tag{6.65}$$

Such a refractive gradient will cause some sort of acoustic refraction. By analogy with the results in Chapter 4, we have the radius of curvature defined by

$$r = -\eta\left(\frac{1}{d\eta / dh}\right)\cos\theta , \tag{6.66}$$

the absolute curvature defined by

$$\rho = \frac{1}{\eta\cos\theta}\cdot\frac{d\eta}{dh}, \tag{6.67}$$

and the curvature relative to the earth's surface defined by

$$\rho' = \frac{1}{R_e} + \frac{1}{\eta\cos\theta}\cdot\frac{d\eta}{dh} . \tag{6.68}$$

Upon substitution of Eq.(6.65) into Eq.(6.68) we find that *ρ'* is positive and the radiation curves away from the surface. On the other hand if a temperature inversion exists at some height, then the refractive index gradient can be negative. If the gradient is sufficiently negative then the ray curvature may be negative and the acoustic radiation would be trapped. The requirement for this is

$$\rho' \; = \; \frac{1}{R_e} + \frac{1}{\eta\cos\theta}\cdot\left(\frac{-\eta}{2T}\right)\cdot\frac{dT}{dh} \; = \; \frac{1}{R_e} - \frac{1}{2T\cos\theta}\cdot\frac{dT}{dh} \; < \; 0 \tag{6.69}$$

or

$$\frac{dT}{dh} > \frac{2T\cos\theta}{R_e} \tag{6.70}$$

For $T = 300$ K and $R_e = 6.37\text{x}10^6$ m, we find the requirement $dT/dh > 9.4\text{x}10^{-5}$ K/m.. This is rather easily satisfied by almost any temperature inversion. The effects of anomalous propagation are commonly heard near airports. On days when a low-lying inversion is present, the jet noise may disturb individuals miles away from the airport. Yet, as soon as the inversion layer disappears, the jet noise abates.

Now let us consider propagation in the ocean medium. By analogy with electromagnetic waves, we expect that a medium exhibiting a strong minimum in propagation velocity (maximum of refractive index) will exhibit ducting. A typical sound speed versus depth profile in the deep ocean is shown in Figure 6-13. We are not disappointed in that there is at least one such minimum (at 900 m depth in this profile). Figure 6-14 shows ray tracings for this profile at two acoustic emission depths (one near the surface and one near the point of minimum sound speed).

Figure 6-13. Typical sound speed versus depth profile for the deep ocean.[8]

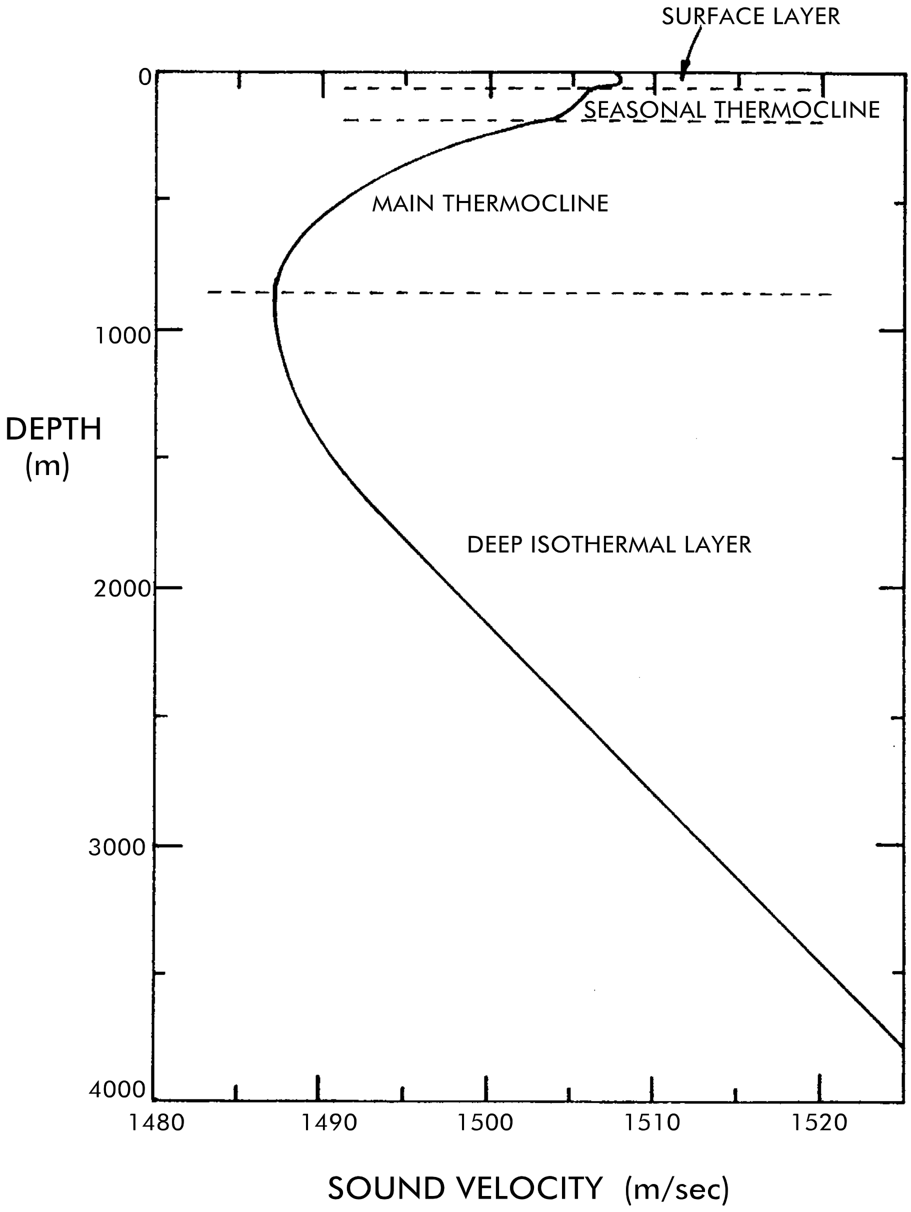
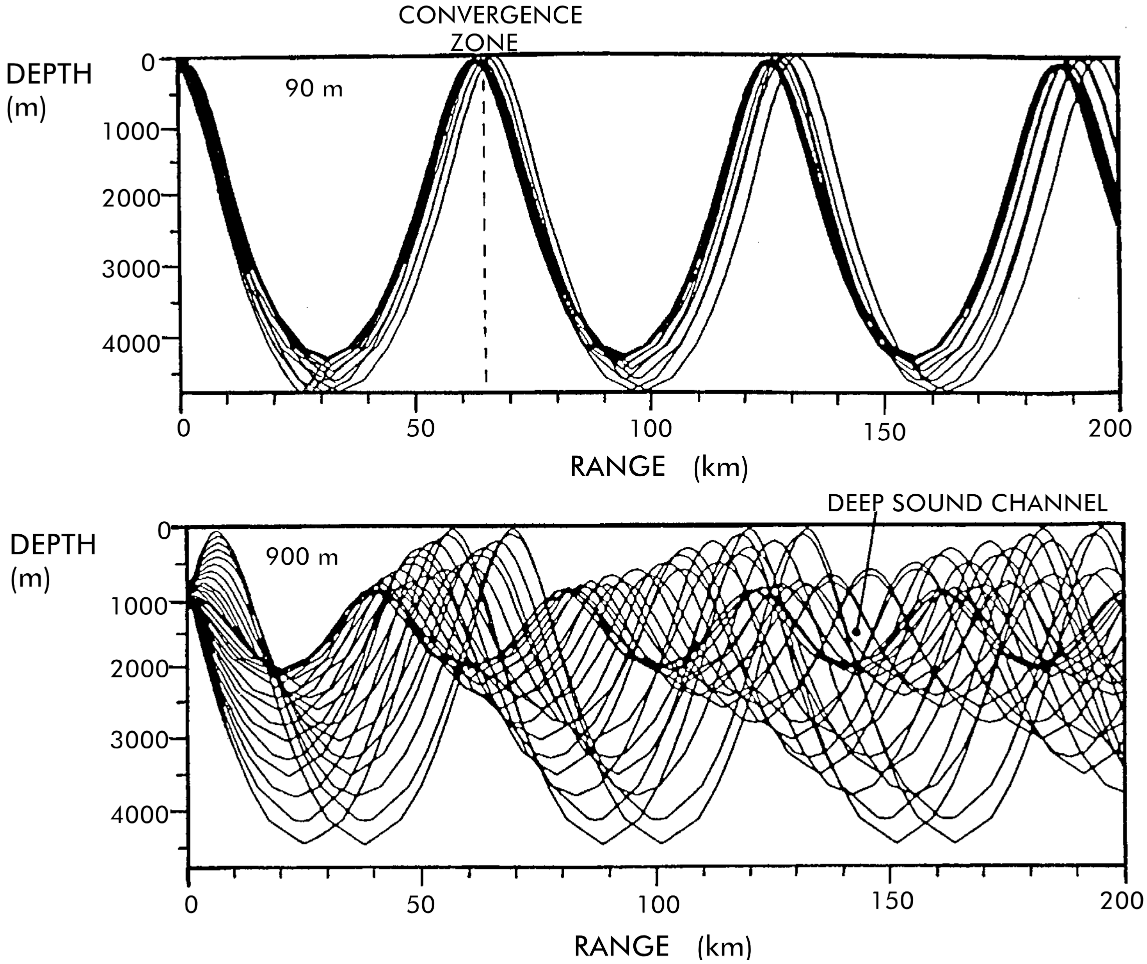


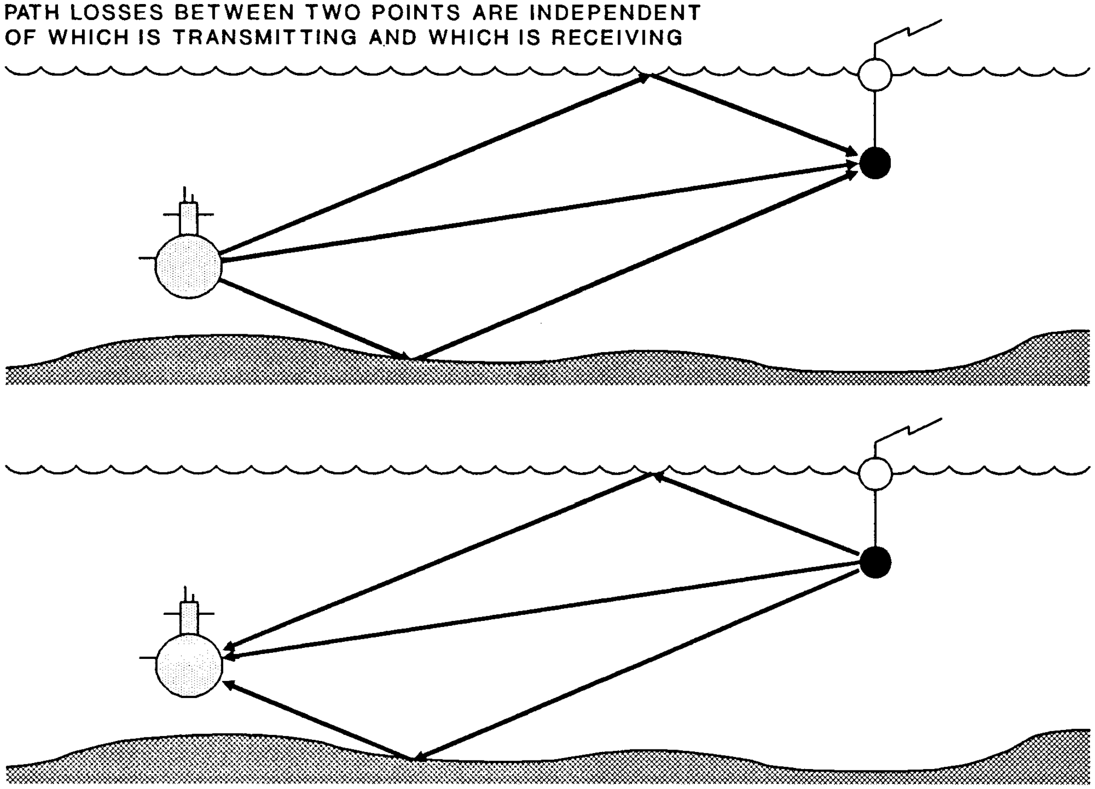
Figure 6-14. Ray tracing for the sound speed profile of Figure 6-13. [8]



Several phenomena are apparent in these ray tracings. First, rays emitted near the velocity minimum oscillate regularly about that minimum. Those making small angles with the axis of this channel oscillate very nearly sinusoidally. Those with larger angles exhibit a significant vertical asymmetry characteristic of the different gradients above and below this channel. This propagation channel is called the "**deep sound channel**" or the "**SOFAR**" channel. [9] Second, rays emitted nearer the surface are also ducted by the channel. However, they tend to stay bunched together. Those rays initially headed toward the surface are immediately either reflected from the surface or refracted back towards the downward direction and then travel in the company of those rays originally emitted downwards. After crossing the axis of the deep sound channel, the entire ray bundle is then refracted back to the surface. The rays reach the surface at a relatively well-defined range (65 km in this example). The rays are then either refracted or reflected back towards the depths and the process repeats. Each time the rays come to the surface, they have spread much less than $1/4\pi R^2$ would predict. The spreading is still proportional to $1/R^2$ at long ranges, but the constant of proportionality is much closer to 3 than to $1/4\pi$. The regions where the ray bundles reach the surface are called **convergence zones** and are regions of considerably enhanced signal detectability. Between the convergence zones are shadow zones where rays from the indicated source simply do not propagate. Thus, not only are the convergence zones regions of enhanced detectability, they are separated by regions of virtual undetectability of the signals. The deep sound channel also has shadow zones near the surface and exhibits weak convergence zone effects. However, in the deep sound channel at sufficient ranges, the shadow zones essentially disappear.

The problem to detection posed by shadow zones deserves some additional comment. Acoustic media possess the property of reciprocity (ray paths are reciprocal). That is, the direction of the energy flow does not matter. Thus, if the noise emitted by a submarine experiences a specific transmission loss on a path (whether that path is a direct refracted path, a surface reflection path or a bottom reflection path as indicated in Figure 6-15) to a sonobuoy, then the reciprocal path from the sonobuoy to the submarine must experience the same transmission loss. If a sonobuoy is in a shadow zone relative to the noise emitted by a submarine, then the submarine is in the equivalent of the shadow zone of any active emissions from the sonobuoy. Transmission loss on a path is the same regardless of whether the path is an active or passive path or the direction of the path. That is the essence of reciprocity.

Figure 6-15. Reciprocity in acoustic transmission and reception.

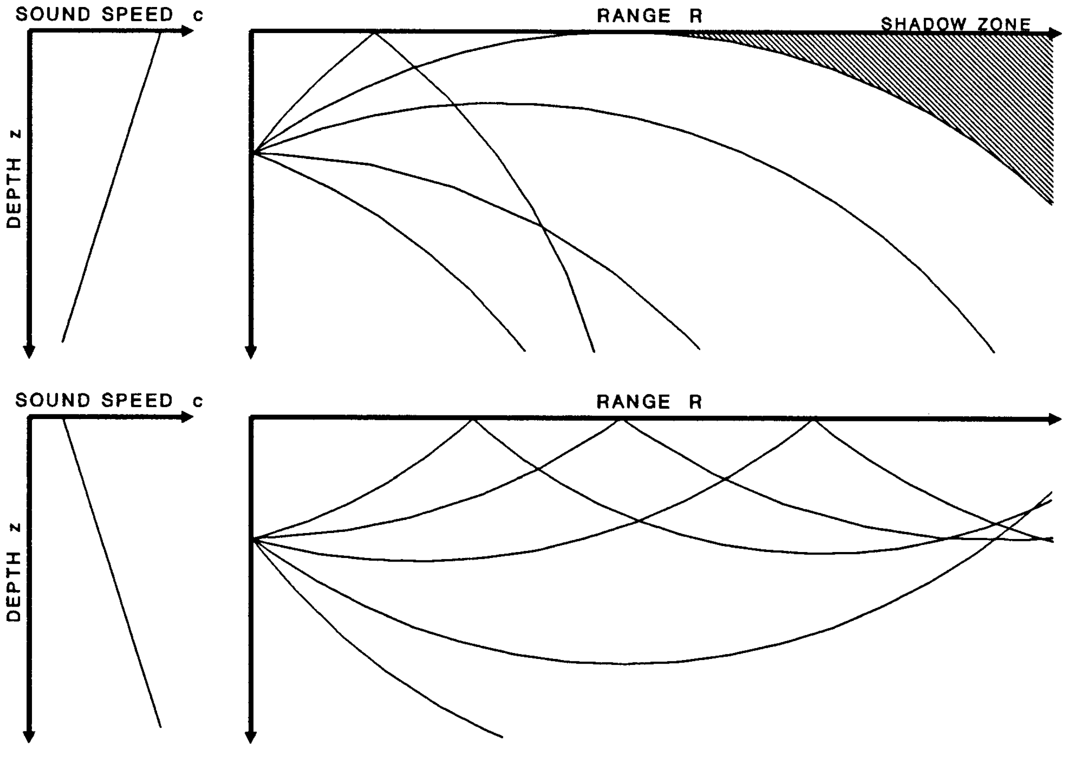

PATH LOSSES BETWEEN TWO POINTS ARE INDEPENDENT
OF WHICH IS TRANSMITTING AND WHICH IS RECEIVING

It should be noted in every instance that as soon as the acoustic radiation has propagated more than a few kilometers, the ubiquity of refractive bending renders meaningless any angle of arrival information in the vertical direction. That is, elevation angle cannot by itself be used to determine the depth of the target. However, in an active sonar system, the addition of range data coupled with knowledge of the refractivity of the medium, would permit accurate estimation of target depth. In most cases azimuthal angle of arrival information is of use in determining target location. However, if the propagation distance is long enough, it may traverse a major current or river outflow. Since these are notorious for having significant temperature and/or salinity gradients, there presence can cause significant refraction in the azimuthal direction. This would cause a degradation in the location utility of the angle of arrival information.

To this point our arguments have been qualitative. It would be useful to have some relations useful for predicting underwater refractive behavior. Let us begin with the surface layer. Potential variability in salinity and temperature with depth can produce gradients of almost any type. The radius of curvature expression, Eq.(6.66), can be rewritten in terms of the sound speed and depth,

$$r = -\eta \left(\frac{1}{d\eta / dh} \right) \cos\theta = c \left(\frac{1}{dc/dz} \right) \cos\theta , \qquad (6.71)$$

Figure 6-16. Effects of surface layer gradients on shallow source propagation.



where now a positive radius of curvature corresponds to rays bending towards the ocean surface. In the upper diagram of Figure 6-16 the sound speed gradient dc/dz is negative and linear. The rays emitted roughly parallel to the surface (cos θ ~ 1) will have a constant radius of curvature and bend away from the surface. If the acoustic emitter is at a nonzero depth there will be one angle of emission at which the ray just grazes the surface. Rays emitted more vertically will be reflected back in the downward direction. There is a region in which no rays can propagate – a shadow zone. If the emitter is at depth z_S, the range R_{SHADOW} to the beginning of this shadow zone is

$$R_{SHADOW} = \left(2\, r\, z_S - z_S^2\right)^{1/2} = \left(\frac{2\, c(z_S)\, z_S}{dc/dz} - z_S^2\right)^{1/2} \tag{6.72}$$

If the sound speed gradient is positive and linear, as in the lower diagram in Figure 6-16, then the rays bend toward the surface. In this case there is no shadow zone. Each ray will bend toward the surface until it strikes the surface and reflects. The angle of reflection is equal to the angle of incidence. The reflected ray will curve back towards the surface until the ray either leaves the linear gradient or the ray strikes the surface again.

Because the sea surface is not flat, but is modulated by a multiplicity of wave patterns, some of the radiation reaching the surface will be scattered into all directions. This scattering causes reverberation (multiple overlapping echoes that act as a noise source) in active sonar systems. It also attenuates the acoustic signal. It has been empirically determined that each reflection from the sea surface imposes a loss (in dB) [10]

$$b = 1.04 \cdot (\text{Sea State}) \cdot \sqrt{\nu} \tag{6.73}$$

where ν is the acoustic frequency in kHz and (Sea State) is the sea state number. Overall transmission loss is increased by a term proportional to the number of surface bounces in the path.

If the acoustic radiation encounters the sea bottom before it is refracted back towards the surface, then an analogous reflection process will occur. The reflection will be accompanied by significant scattering and transmission loss. Figure 6-17 shows the reflection of common ocean sediment as a function of frequency and angle of incidence.[11] The situation is complicated by considerable variability from location to location even at a single frequency. Figure 6-18 shows a series of typical loss vs. angle profiles for the 1-4 kHz frequency region.[8] Each number represents a different bottom type (characterized by the reflection behavior not any geological or oceanographic character). Depth sounding charts sometimes indicate the bottom type as represented by these or a similar set of curves. As with surface bounces, the total path attenuation will contain a term proportional to the number of bottom bounces encountered.

In many instances the water in the surface layer is fairly well mixed. This mixed state should be accepted as the norm. Temperature and salinity gradients are the exception. If there is no

Figure 6-17. Bottom reflection loss as a function of frequency and angle of incidence.[11]

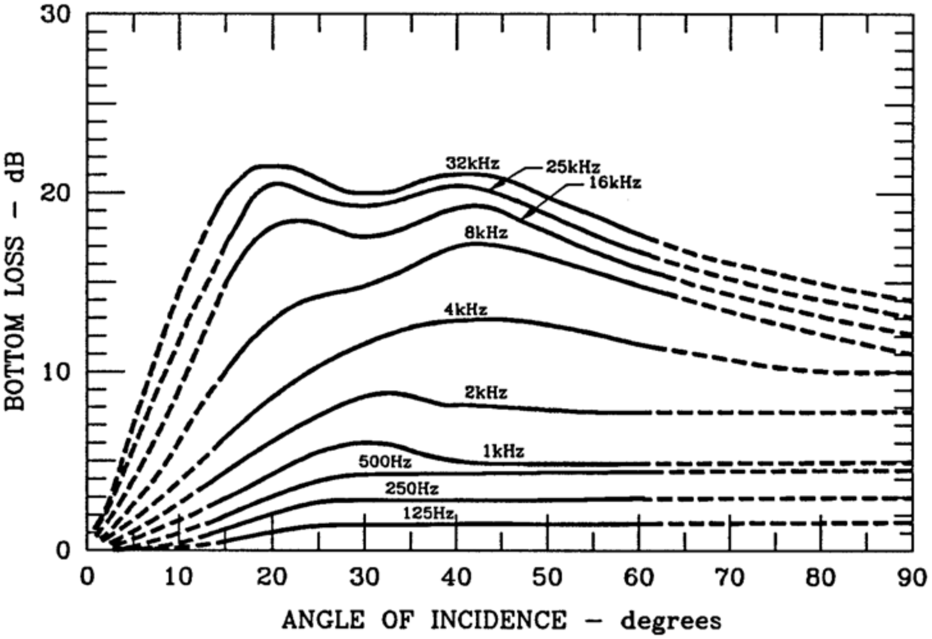
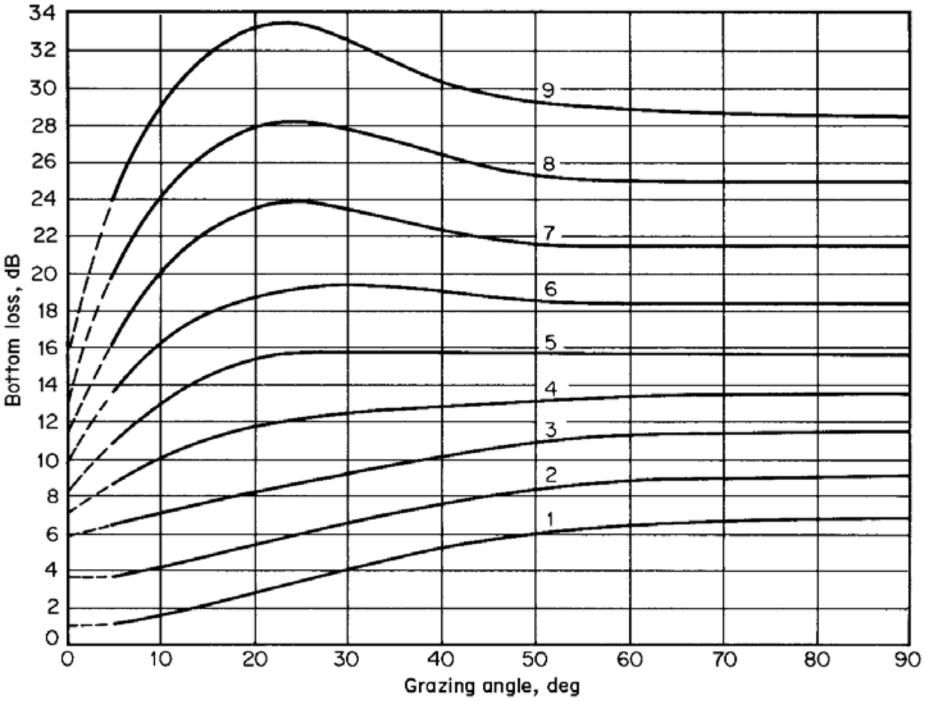


Figure 6-18. Bottom reflection loss over 1-4 kHz frequency range for different bottom types.[8]



temperature or salinity gradient, then ignoring higher order corrections, from Eq.(6.22) we determine the speed of sound gradient to be

$$\frac{dc}{dz} \approx 0.0163 \quad \mathrm{m^{-1}} . \tag{6.74}$$

If we assume normal salinity (S=35 psu) and ignore high order terms in Eq.(6.22), we can obtain an approximate relation for the sound speed

$$c \approx 1449 + 5.3T - 0.053T^2 + 0.0163z . \tag{6.75}$$

For rough estimates we may simplify this further by picking T=15 °C, yielding

$$c \approx 1516 \quad \mathrm{m/s} \tag{6.76}$$

which is accurate to ± 5% over the typical 0 to 30 °C range of temperatures.

If the bottom of the mixed layer is at depth D_1 then the skip distance, the distance between reflections off the surface of a ray that barely grazes the bottom of the mixed layer (as identified in Figure 6-19), can be determined from

$$R_S = \left(4D_1\left(2r - D_1\right)\right)^{1/2} \tag{6.77}$$

the formula for the length of a chord of a circle of radius r and chord center to circle distance D_1, where the radius of curvature may be approximated by

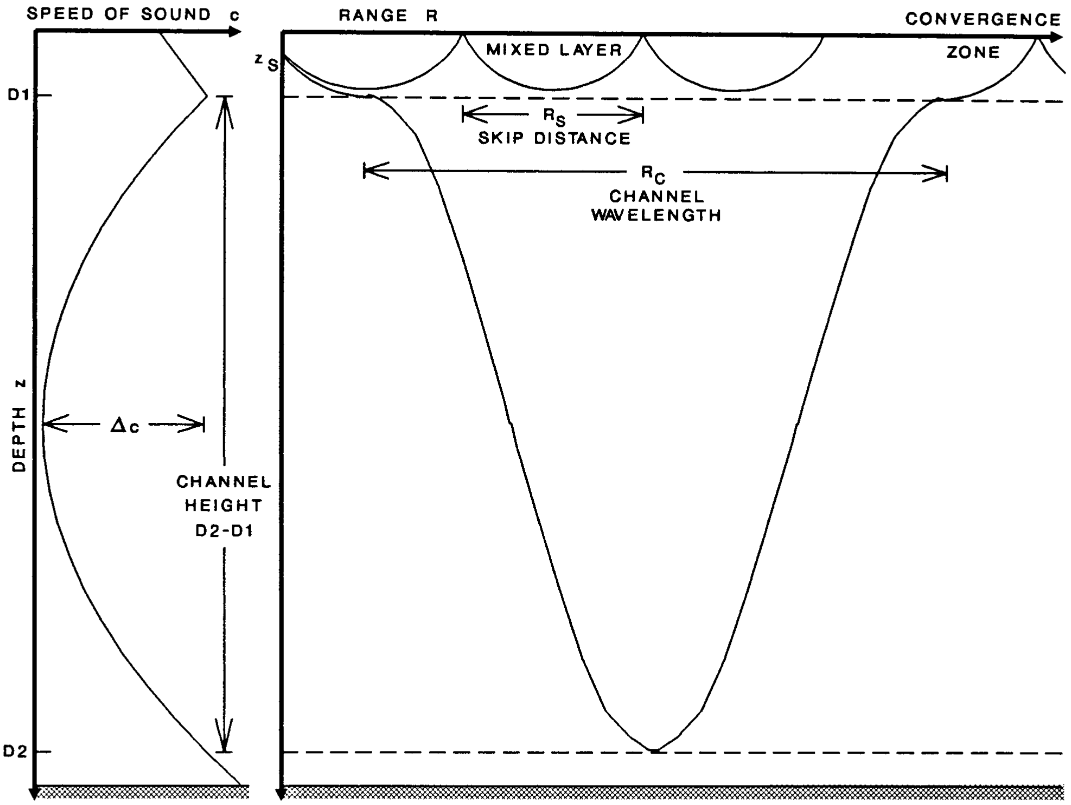
$$r = \frac{c}{dc/dz} = 93{,}000 \quad \mathrm{m} . \tag{6.78}$$

These expressions reduce to [9]

$$R_S \approx \sqrt{8D_1 r} \approx 863\sqrt{D_1} . \tag{6.79}$$

The speed of sound profile in the deep sound channel is not parabolic (for one thing it is strongly asymmetric about the minimum) but neither is it truly bi-linear – the combination to two linear gradient segments – (although the lower half quickly assumes the linear gradient associated with the deep isothermal layer. Empirically it is found that a bi-linear approximation is much better than a parabolic approximation to this channel. In a bi-linear approximation, the total repeat distance or "wavelength" will be the sum of the contributions from the upper linear segment and the lower linear segment. Thus, the upper and lower gradients may be averaged without error. One

Figure 6-19. Geometry for computation convergence zone range and skip distance.



estimator of the gradient is to take the difference between the maximum velocity c_{MAX} at the top of the channel and the minimum velocity c_{MIN} in the middle of the channel and divide by half of the total height of the channel (D_2 -D_1). Using Eq.(6.79) we may then write the expression for the "wavelength" as

$$\begin{aligned} R_C &= R_{UPPER} + R_{LOWER} \\ &= 2\left(8\left(\frac{D_2 - D_1}{2} \right)\left(\frac{c_{MAX}\left(\left(D_2 - D_1 \right)/2 \right)}{c_{MAX} - c_{MIN}} \right) \right)^{1/2} \\ &= 2\sqrt{2}\left(D_2 - D_1 \right)\left(\frac{c_{MAX}}{c_{MAX} - c_{MIN}} \right)^{1/2} \end{aligned} \qquad (6.80)$$

The range to the convergence zone is the sum of the channel "wavelength" plus any effect due to the mixed layer

$$R_{CZ} = R_C + \begin{cases} R_S & \text{– if both source and receiver are near} \\ & \quad \text{the top of the mixed layer} \\ 0.5R_S & \text{– if one is near the top of the mixed} \\ & \quad \text{layer and the other is near the bottom} \\ 0 & \text{– if both are near the bottom of the mixed} \\ & \quad \text{layer or if there is no mixed layer} \end{cases} \tag{6.81}$$

The equations above are merely rough estimates. Much more accurate predictions could be obtained by actually performing the ray tracing for measured velocity profiles (if available) using the technique described in Chapter 4.

Seismic waves

So far we have said nothing about the third major acoustic propagation medium – solid ground. Seismic waves do play a significant role in some military operations. For example, there have been numerous attempts to gain a tactical advantage by tunneling under enemy fortifications. The "Crater" of the Civil War siege of Petersburg and the North Korean tunnels under the Demilitarized Zone are two significant examples. In the latter case seismic waves have been exploited to monitor such activity. Seismic waves have also been exploited to remotely detect vehicular and personnel traffic.[12]

In a linear elastic homogeneous solid, the acoustic equations are obtained from Cauchy's equation of motion [13],[14]

$$\rho \frac{\partial^2 \xi_i}{\partial t^2} = \sum_{j=1}^{3} \frac{\partial \sigma_{ij}}{\partial x_j} \tag{6.82}$$

where ρ is the normal material density, $\xi_i(x,t)$ is the ith Cartesian component of the displacement from their normal position of particles normally at position x, the quantities σ_{ij} are the components of the stress tensor (applied forces per unit area), and ε_{ij} are the components of the strain tensor (macroscopic changes in length per unit length). The stress is related to the strain by

$$\sigma_{ij} = 2\mu\varepsilon_{ij} + \lambda\delta_{ij}\sum_{k=1}^{3}\varepsilon_{kk} \tag{6.83}$$

with the strains defined by

$$\varepsilon_{ij} = \frac{1}{2}\left[\frac{\partial \xi_i}{\partial x_j} + \frac{\partial \xi_j}{\partial x_i}\right] \tag{6.84}$$

and where δ_{ij} is the Kronecker delta (δ_{ij} =1 if i=j, and δ_{ij} = 0 if i≠j), and λ and μ are the Lamé constants. These constants are related to common material parameters by the relations

$$\lambda = \frac{\nu E}{(1+\nu)(1-2\nu)} \tag{6.85}$$

and

$$\mu = G = \frac{E}{2(1+\nu)} \tag{6.86}$$

where E is the elastic modulus (or Young's modulus is the ratio of extensional stress to the resulting extensional strain in a cylinder pulled on from both ends), ν is Poisson's ratio (the ratio of lateral contraction of a cylinder being pulled on from both ends to its longitudinal extension), and G is the shear modulus (the ratio of shear stress to the resulting shear strain – or transverse displacement per unit length). E and ν are seen to be related to λ and μ by

$$E = \frac{(3\lambda + 2\mu)\mu}{\lambda + \mu} \tag{6.87}$$

and

$$\nu = \frac{\lambda}{2(\lambda + \mu)} . \tag{6.88}$$

The bulk modulus B (the ratio of hydrostatic pressure to the resulting volume change – a measure of incompressibility) is related to λ and μ by

$$B = \frac{3\lambda + 2\mu}{3} \tag{6.89}$$

and to E and ν by

$$B = \frac{E}{3(1 - 2\mu)} . \tag{6.90}$$

Usually at least two of the quantities E, B, G, and/or ν can be found for a material. The other quantities can be derived by suitable manipulation of the above parameters.

Substitution of the stress-strain relation into Cauchy's equation yields a "wave equation"

$$\rho \frac{\partial^2 \xi}{\partial t^2} = (\lambda + 2\mu - \mu)\nabla(\nabla \cdot \xi) + \mu \nabla^2 \xi \tag{6.91}$$

If the displacements are irrotational ($\nabla \times \xi = 0$) or purely **compressional** in nature, then one can set

$$\xi = \nabla \Phi \tag{6.92}$$

and rewrite Eq. (6.91) as

$$\nabla^2 \Phi - \frac{1}{c_P^2} \frac{\partial^2 \Phi}{\partial t^2} = 0 \tag{6.93}$$

where

$$c_P^2 = \frac{\lambda + 2\mu}{\rho} = \frac{E(1-\nu)}{\rho(1+\nu)(1-2\nu)} \tag{6.94}$$

is the compressional (P) wave velocity.

On the other hand, if the displacements are solenoidal ($\nabla \cdot \xi = 0$) or purely **shear** in nature, then one can set

$$\xi = \nabla \times \Psi \tag{6.95}$$

and rewrite Eq. (6.91) as

$$\nabla^2 \Psi - \frac{1}{c_S^2}\frac{\partial^2 \Psi}{\partial t^2} = 0 \tag{6.96}$$

where

$$c_S^2 = \frac{\mu}{\rho} = \frac{E}{2\rho(1+\nu)} \tag{6.97}$$

is the shear (S) wave velocity. Both Eqs. (6.93) and (6.96) have the form of classical wave equations.

Earthquakes are the source of the strongest seismic waves. Both P and S waves are produced by earthquakes. As we shall see, the P waves travel faster than the S waves. The sharp jolt that often precedes the strong shaking of an earthquake is due to the passage of the P waves. The slower S waves are also longer lasting and more destructive. They cause much of the side-to-side and up-and-down motions that are observed. The difference in S and P wave velocity is used to determine the range to the epicenter of the earthquake by measuring the time between the arrival of the P and S waves. The distance *d* the earthquake waves have traveled is

$$d = c_P T_P = c_S T_S = \frac{c_P c_S}{c_P - c_S}\left(T_S - T_P\right) \tag{6.98}$$

where T_P is the time between the earthquake event and the detection of the P waves and T_S is the time between the earthquake event and the detection of the S waves.

The S and P waves are bulk (or body) waves and travel through the mass of the earth. Earthquakes can also generate surface waves. These waves occur only at the earth's surface. Motions are restricted to roughly one wavelength or less in depth. There are two common forms of surface waves: Rayleigh waves and Love waves. Love waves are horizontal shear waves that only occur in areas with strong horizontal stratification near the surface. They travel more slowly than S waves (velocity approx. 2/3 c_S). Rayleigh waves are waves that combine vertical and axial motions. A point on the surface moves up and down, forward and backward in a circular motion. Rayleigh waves have the slowest propagation velocity (approx. ½ c_S), last the longest, and often have the largest amplitude (even larger than S waves). Because they combine vertical and back-and-forth motion, they often cause most of the destruction produced by an earthquake. Besides their importance in earthquake damage, Love and Rayleigh waves are important because they are generally not produced by nuclear explosions. This coupled with the tendency for explosions to produce a relative strong P wave makes it difficult to disguise a nuclear explosion as an earthquake.

Attenuation mechanisms in solids (especially rocks and soil) are not as well known as in fluids and gases. Nevertheless it has been found that attenuation can be described by an exponential attenuation coefficient a. The peak-to-peak amplitude A of a seismic wave decays with propagation distance as

$$A(L) = A(0)e^{-aL} . \tag{6.99}$$

The attenuation coefficient appears to increase linearly with increasing frequency, just like the viscous damping contributions to the acoustic attenuation coefficients in air and water. For convenience, the attenuation coefficients are often listed as

$$a(f) = a_0 f \tag{6.100}$$

where a_0 has units of sec/meter and f is the acoustic frequency in Hz.

In Table 6-1 we have listed the speeds of sound in a number of rock and soil types.[15] For comparison the table also includes the speeds of sound for a number of solid materials used in the fabrication of military equipment.[16] Where available, both shear and compression wave velocities are included. In Table 6-2 the acoustic parameters (sound speeds and attenuation coefficients) of several rock samples are described.[17] Examination of this last table shows that at frequencies of 1 kHz most samples have attenuation coefficients in excess of 1 km^{-1}. Such high frequencies will not propagate for distances more than a few kilometers. However, at frequencies below a few Hz, the attenuation coefficients are of the order of 0.01 km^{-1} or less. Very low frequency seismic waves can easily propagate around the world without excessive attenuation.

Table 6-1. Speeds of sound in various solid materials.[15],[16]

		Speed of Sound (m/s)		
Material	Density	Range of values	Compression	Shear
Loose, Dry Soils	–	180-1000	–	–
Clay	–	760-1900	–	–
Wet Soils	–	760-1900	–	–
Compact Soils	–	910-2600	–	–
Sandstone	–	910-4300	–	–
Shale	–	1800-5300	–	–
Limestone/Chalk	–	2100-6400	–	–
Metamorphic Rock	–	3000-6400	–	–
Volcanic Rock	–	3000-6700	–	–
Solid Plutons	–	4000-7600	–	–
Jointed Granite	–	2400-4600	–	–
Weathered Rock	–	600-3100	–	–
Aluminum	2.7	–	6420	3040
Iron	7.9	–	5950	3240
Stainless Steel	7.9	–	5790	3100
Titanium	4.5	–	6070	3125
Fused Silica	2.2	–	5968	3764
Plexiglass	1.18	–	2680	1100
Polyethylene	0.90	–	1950	540

Table 6-2. Characteristics of seismic waves in selected rocks.[17]
In the loss parameter expressions *f* is the acoustic frequency in Hz

	Speed of Sound (M/s)		Attenuation Coefficient (sec/m)	
Material	Compression	Shear	Compression	Shear
Granite:				
Kamyk	6100	3500	$3.9 \times 10^{-6} f$	—
Sample 3	5500	3300	$2.7 \times 10^{-5} f$	$3.3 \times 10^{-5} f$
Sample 4	4880	2990	$1.4 \times 10^{-5} f$	$1.7 \times 10^{-5} f$
Limestone				
Solenhofen	5560	2900	$5.2 \times 10^{-6} f$	$5.8 \times 10^{-6} f$
Sample I-1	6100	3200	$3.1 \times 10^{-6} f$	$2.5 \times 10^{-6} f$
Sandstone				
Sample 116	5000	—	$3.5 \times 10^{-6} f$	—
Chalk				
Chislehurst	2300	1240	$1.0 \times 10^{-6} f$	$1.2 \times 10^{-6} f$
Shale				
Pierre	2200	810	$4.5 \times 10^{-5} f$	$4 \times 10^{-6} f$

In some seismic references, attenuation is described in terms of a Q factor. The Q of any oscillator is the fractional energy loss per cycle of oscillation. It can be related to an attenuation coefficient by

$$a = \omega / 2cQ \tag{6.101}$$

where ω is the angular frequency of the radiation and c is the appropriate sound velocity.

The primary vertical changes in seismic velocity are due to changes in rock composition with changes in depth. However, it has been determined that except for discontinuities caused by composition changes, the sound velocity slowly increases with increasing depth. The positive sound speed gradient means that seismic waves will bend towards the surface of the earth. Table 6-3 shows the sound speed characteristics of typical oceanic crust. At the surface there is a roughly 1.5 m/sec per meter of depth gradient in the P wave velocity. Recalling Eq. (6.71) for the radius of curvature of an acoustic wave in a linear velocity gradient, we calculate a radius of curvature of approximately 1.5 km. Surface ducting is almost certain to occur in this medium. Note: however, that the crust under the ocean is almost certainly saturated with water and under considerable hydrostatic pressure. Consequently, this may not be a good model for estimating near-surface seismic properties on dry land. Gradients at deeper depths are typically much less strong (barring interfaces between different materials) and radii of curvature can exceed 100 km or even 1000 km along deep subsurface paths. The curvature is of clear significance in determining the distance to distant (100's to 1000's of km) seismic events (earthquakes or underground nuclear tests).

Table 6-3. MARMOD, a generic marine seismic model of the oceanic crust.[14]

DEPTH (km)	c_p (km/sec)	c_S (km/sec)	ρ (g/cm^3)
0.0	4.50	2.40	2.0
1.5	6.80	3.75	2.8
6.0	7.00	3.85	2.9
6.5	8.00	4.60	3.1
10.0	8.10	4.70	3.1

Linear gradients are assumed between points.

References

[1] Beranek, Leo L., Acoustics (McGraw-Hill Book Co., New York NY, 1954).

[2] Morse, Philip M. and Ingard, K. Uno, Theoretical Acoustics (Princeton University Press, Princeton NJ, 1968).

[3] Mackenzie, K. V., “Nine-term Equation for Sound Speed in the Oceans”, *J. Acoust. Soc. Am.*, 70 807 (1981).

[4] Apel, J. R., Principles of Ocean Physics (Academic Press, London UK, 1987).

[5] Schott, G., Geographie des Atlantischen Ozeans, 3rd Ed. (C. Boysen, Hamburg, Germany, 1942).

[6] Worthington, L. W., and Wright, W. R., North Atlantic Ocean Atlas of Potential Temperature and Salinity in the Deep Water, Vol. 2, Woods Hole Oceanographic Institution Atlas Series (1970).

[7] Acoustical Society of America, “American National Standard Method for the Calculation of the Absorption of Sound by the Atmosphere”, ANSI S1.26-1978 (American Institute of Physics, New York NY, 23 June 1978).

[8] Urick, Robert J., Principles of Underwater Sound 3rd Ed. (McGraw-Hill Book Co., New York NY, 1983).

[9] Coppens, Alan B., Sanders, James V., and Dahl, Harvey A., Introduction to the SONAR Equations (Naval Postgraduate School, Monterey CA, 1981).

[10] Schulkin, M., *J. Acoust. Soc. Am.*, 44, 1152-1154 (1968).

[11] Fassnacht, Frank, Naval Combat Systems unpublished lecture notes Naval Postgraduate School (1995).

[12] Dickson, Paul, The Electronic Battlefield (Indiana University Press, Bloomington IN, 1976).

[13] Pierce, Allan D. and Berthelot, Yves H., “Acoustics” in Eshbach’s Handbook of Engineering Fundamentals 4th Ed., Byron D. Tapley and Thurman R. Poston, Eds. (John Wiley & Sons, New York NY, 1990).

[14] Shearer, Peter M., Introduction to Seismology (Cambridge University Press, Cambridge UK, 1999).

[15] Brode, H. L., "Height of Burst Effects at High Overpressures", RM-6301 Rand Corporation (1970). Reproduced in Kosta Tsipis, Arsenal (Simon & Schuster, New York NY, 1983).

[16] Weast, Robert C. (Ed.), Handbook of Chemistry and Physics 49th Ed. (Chemical Rubber Co., Cleveland OH, 1968), p. E-38.

[17] White, James E., Seismic Waves: Radiation, Transmission, and Attenuation (McGraw-Hill Book Co., New York NY, 1965) p. 89.

Problems

6-1. What are the speeds of sound in air (at sea level) and in water? Give a numerical answer accurate to 20%. Note: these are numbers that should be committed to memory.

6-2. What is the speed of sound in helium at a temperature of 300 K?

6-3. Plot the speed of sound vs. depth for the following data:

Depth(m)	Temperature(C)	Salinity(psu)
0	25	34
100	25	34
200	24	34
300	20	34
400	15	34
500	10	34
600	5	34
700	2	34
800	2	34
⋮	⋮	⋮
2000	2	34
2100	2	38
⋮	⋮	⋮

What peculiar propagation characteristics might this sound speed profile produce?

6-4. Compare and contrast the frequency behavior of acoustic attenuation in air and in water.

6-5. A source has a sound pressure level of 80 dB at a frequency of 1 kHz at a distance of 1 m from the source. Assuming the threshold for audibility of sound at 1 kHz is 20 dB and spherical spreading of the sound waves, estimate the distance in air at which this source can barely be heard. Estimate the distance in water at which this same source can barely be heard.

6-6. How does acoustic reflection from a material interface differ from electromagnetic reflection from a material interface?

6-7. A source and a receiver are at the surface of a body of water. The water has a mixed layer 100 m deep and a deep sound channel with maximum depth of 2100 m. The speed of sound is 1510 m/s at the top of the deep sound channel and 1490 m/s at the middle of the channel. At what range should the first convergence zone be expected?

6-8. A helicopter drops passive sonobuoys at ranges of 40, 80, 120, 160, and 200 km from the source in problem 6-7. Which if any of these sonobuoys are likely to detect the source?

6-9. An underground river flows into the ocean from a submerged cliff face at a depth of 300 m. The flow rate is so high and the local currents are such that an unmixed layer of fresh water 50 m thick extends from the river mouth out to a distance of 200 km. The river water temperature is 10 °C whereas the ocean water temperature is 5°C. Salinity of the ocean water is 35 psu. Will this river create an acoustic ducting layer or a reflecting layer? Justify your answer by calculations.

6-10. Compare and contrast the propagation of electromagnetic waves in the atmosphere with the propagation of acoustic waves in sea water. Qualitative comparisons will suffice.

6-11. Aluminum has a density of 2.6 x 10^3 kg/m^3, an elastic modulus of 70 GPa, and a shear modulus of 28 GPa. Estimate the compressional and shear speeds of sound in aluminum.

6-12. A seismic sensor placed in a limestone formation detects the sound of digging with pickaxes. The shear wave component from each blow of the pick arrives 50 ms after the compressional wave component. How far away from the sensor is the source of the digging?

CHAPTER 7

NUCLEAR RADIATION: ITS ORIGIN AND PROPAGATION

The Nuclear Radiation Particle "Zoo"

Particle physics has proven the existence of many dozens of subatomic or nuclear particles. Of these less than a dozen will be encountered in common military situations. We describe these critical few below. The particles are often categorized by the nature of their physical characteristics, the ways in which they are produced, and the way in which they interact with other particles. These categorizations are also discussed below.

Nuclear particles are subject to two or more of four fundamental forces. **Gravity** is the force of attraction between masses (or more properly between volumes of mass-energy). All fundamental particles are subject to gravity. However, the gravitational force is so weak that outside of collapsed stellar objects (white dwarfs, neutron stars, and black holes) it seldom has an impact on nuclear processes. The **electromagnetic** force is the force of attraction or repulsion between charged particles. Charged nuclear particles are subject to the electromagnetic force. The **weak** force is a short-range force that manifests itself only in radioactive decay processes. Leptons and hadrons are subject to the weak force. The **strong** (or nuclear) force is a short-range force that causes baryons to bind together and form nuclei. It also governs the major interactions between hadrons. Only hadrons are subject to the strong force.

A **lepton** is any one of six fundamental particles with spin ½ which are capable of interacting via the weak force but not the strong (or nuclear) force. The six leptons are electrons, muons, and tau particles, electron neutrinos, muon neutrinos, and tau neutrinos. Leptons have a lepton with anti-leptons (the anti-particles of leptons) having the negative lepton number of the corresponding normal leptons. Lepton number is conserved in nuclear reactions.

A **baryon** is any half-integer spin and usually very massive fundamental particle capable of interacting via both the weak and strong forces. Baryons have a baryon number with anti-baryons (the anti-particles of baryons) having the negative baryon number of the corresponding normal baryons. Baryon number is conserved in nuclear reactions.

A **hadron** is any fundamental particle that is capable of interacting via the strong force. All baryons are hadrons. **Mesons** are integer-spin hadrons. Mesons are not conserved in nuclear reactions.

Electrons are light (0.511 MeV) stable leptons with a -1 electrical charge. Electrons combine with nuclei to form atoms. They may be ejected from atoms by nuclear scattering processes. They are produced in a common form of radioactive decay called beta-minus decay leading to the common name of beta (or more properly beta-minus) particles. They are also produced together with positrons are produced in pair production absorption of gamma rays.

Positrons are the anti-particles of electrons having +1 electrical charge, but also having the same mass and spin as electrons. Positrons are emitted in beta-plus radioactive decay. They are also produced along with electrons (one electron and one positron for each absorbed gamma ray) in pair production absorption of gamma rays. When a positron collides with an electron, the two particles may annihilate each other producing two gamma rays. Their complete masses of the betas are converted into energy in the form of two 0.511 MeV gamma rays.

Gamma rays are photons (or quanta) of electromagnetic radiation. Gamma rays are often produced when excited nuclei make transitions to lower energy states. When electrons in excited atoms make similar transitions we call the resulting photons **x-rays**. Gamma rays typically have energies of 0.1-10 MeV while x-rays typically have energies of 1-100 keV.

Protons are stable baryons of moderate mass (938 MeV) having +1 electrical charge and spin ½. Protons are one of the primary constituents of nuclei.

Neutrons are electrically neutral baryons of moderate mass (940 MeV). Neutrons are the second major constituent of nuclei. Because of this, neutrons and protons are collectively referred to as **nucleons**. Free neutrons decay to protons and electrons with a half-life of 930 seconds. Neutrons bound in nuclei are stable. Free neutrons are produced in many particle interactions.

Neutrinos are massless, electrically neutral leptons produced in weak interactions such as beta decay. They are produced to balance the lepton number in the beta decay processes. If an electron (beta-minus) is produced in a decay process, an anti-electron neutrino must be produced to balance the lepton number. If a muon decays into an electron, a muon neutrino and an anti-electron neutrino must be produced to balance the lepton number. Practically speaking, the interaction probability of neutrinos with normal matter is so low that they effectively do not interact. An average neutrino can pass straight the center of the earth without interacting.

Alpha particles are bare nuclei of Helium-4. They consist of two protons & two neutrons bound together by the strong force. The heavy (3.727 GeV), charge +2, spin 0 particles are extraordinarily stable. They are emitted in the alpha decay of many radioactive nuclei and in a variety of particle-nucleus interactions.

Muons are heavy (106 MeV) leptons produced in the interaction of cosmic rays with nuclei in the upper atmosphere. They are also produced in a variety of high-energy particle interactions. Muons decay to electrons (and two neutrinos) with a half-life of 2.2 μsec.

Deuterons are bare nuclei of deuterium (Hydrogen-2). They contain 1 proton & 1 neutron. **Tritons** are bare nuclei of tritium (Hydrogen-3) and contain 1 proton & 2 neutrons. **Helium-3** nuclei contain 2 protons & 1 neutrons. All three are occasionally emitted in radio-active decay. They are also produced in many particle-nucleus interactions. The parent gases (deuterium, tritium, and Helium-3) are available in sufficient quantities that the ions can be accelerated in particle accelerators to produce different artificial isotopes for industrial and medical purposes.

The particles described above and their properties are summarized in Table 7-1 below.

Table 7-1. Commonly encountered nuclear particles.

PARTICLE	INTERACTIONS?			ELEC	MASS		LEPTON	BARYON	LIFETIME
NAME	EM	WK	ST	CHG	(MeV)	SPIN	NUMBER	NUMBER	(sec)
GAMMA (γ)	Y	N	N	0	0	1	0	0	Stable
ELECTRON (e^-,β^-)	Y	Y	N	-1	0.51099906	½	1e	0	Stable
POSITRON (e^+,β^+)	Y	Y	N	+1	0.51099906	½	-1e	0	"Stable"
PROTON (p)	Y	Y	Y	+1	938.27231	½	0	1	Stable
ANTIPROTON (p^-)	Y	Y	Y	-1	938.27231	½	0	-1	"Stable"
NEUTRON (n)	N	Y	Y	0	939.56563	½	0	1	930
e NEUTRINO	N	Y	N	0	0	½	±1e	0	Stable
μ NEUTRINO	N	Y	N	0	0	½	±1μ	0	Stable
MUON (μ)	Y	Y	N	±1	105.658389	½	±1μ	0	2.198×10^{-6}
ALPHA (α)	Y	Y	Y	+2	3727.3799	0	0	4	Stable
DEUTERON (d)	Y	Y	Y	+1	1875.61339	1	0	2	Stable
TRITON (t)	Y	Y	Y	+1	2808.5561	½	0	3	3.875×10^{8}
HELIUM-3 (^{3}He)	Y	Y	Y	+1	2808.5375	½	0	3	Stable
PION (π)	N	N	Y	0	134.975	1	0	0	8.9×10^{-17}
	Y	N	Y	±1	139.579	1	0	0	2.604×10^{-8}

A **nucleus** (the plural is **nuclei**) is a collection of protons and neutrons bound together by the strong (nuclear) force in a more or less stable arrangement. Unstable nuclei transform themselves into more stable nuclei via processes collectively known as radioactive decay. Positively-charged nuclei are massive (many GeV) and may possess either integer or half-integer spin depending on the numbers of protons and neutrons they contain.

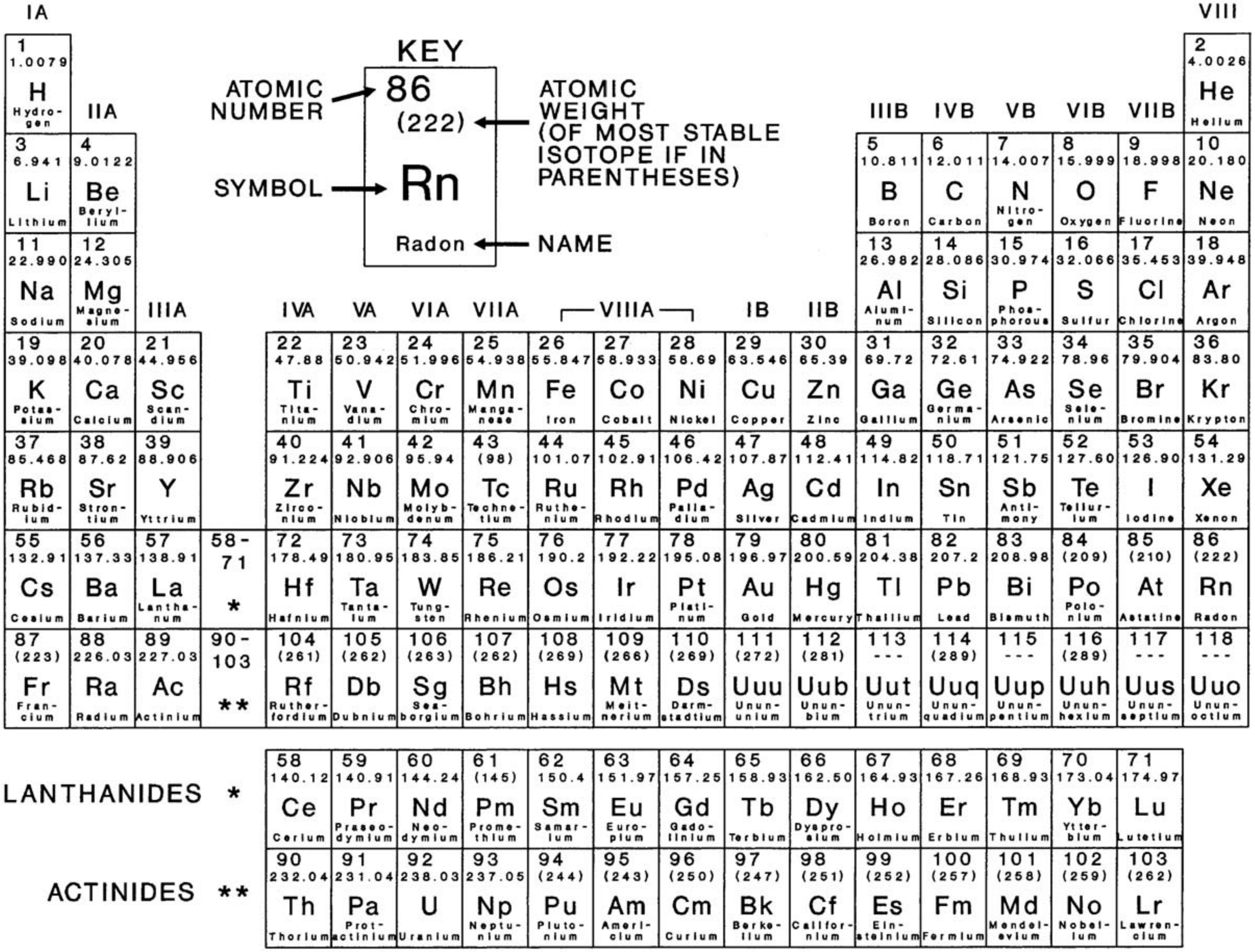
Nuclei at rest tend to attract electrons because of their positive charge. Nuclei which have attracted a number of electrons equal to the number of protons, and thus have become electrically neutral, are called atoms. The chemical properties of an atom is determined by the number of electrons it has (and thus the number of protons in its nucleus). Over the years the chemical community was able to identify a moderate number of distinct chemical **elements**, atoms which react in unique and predictable ways. With the development of the “nuclear” theory of atoms we are able to organize the chemical elements into a **periodic table** that is based on **atomic number** (Z equal to the number of protons in each nucleus). Figure 7-1 shows the author’s version of the periodic table. This version has been updated to include the latest data as of roughly 1 January 2000. Elements beyond Z=109 have not yet been officially named. Many individuals use an interim nomenclature based on the value of Z. Thus Z=110 is ununnilium (un=1, un=1, nil=0, +ium) with corresponding symbol Uun. Z=111 is unununium (Uuu), Z=112 is ununbium (Uub), Z=113 is ununtrium (Uut), etc. The naming of newly discovered elements evokes strong passions in the participants. The interim nomenclature is sufficiently unbiased, no individual name will ever be a candidate for permanent adoption, and can automatically accommodate all new discoveries, so its continued use for new elements is virtually assured.

Each element has a different unique atomic number. The number of neutrons (N) in a nucleus of any given element is not fixed. Different neutron numbers lead to different total masses of a nucleus. The **mass number** (A) of a nucleus is the sum of its atomic number (number of protons) and its neutron number. Components of the same element with different masses (different neutron numbers and mass numbers) are called **isotopes**. **Radioisotopes** are isotopes that are unstable and subject to radioactive decay. The atomic weight of an isotope is roughly equal to its mass number. The atomic weight of an element is an average of the atomic weights of each isotope weighted by the fractional natural abundance of those isotopes. The term **nuclide** is used to distinguish unique species with different atomic numbers and neutron numbers. Any possible nucleus can be represented as a distinct nuclide. **Radionuclides** are nuclides that are unstable and subject to radioactive decay.

The difference between the mass number and the atomic weight is due to three factors. One is the fact that atomic weights for atoms contain the weights associated with the bound electrons (roughly 0.05% of the total mass). Second, when nucleons are brought into close contact to form a nucleus, their net energy is less than the energy of the isolated particles. This energy difference is called the **binding energy**. As we shall see in a later section, the binding energy is a complicated function of the number of neutrons and protons. Since mass and energy are equivalent, the binding energy reduces the mass of the nucleus below the mass that would result from summing the masses of the separated neutrons and protons. Third, atomic weights are normalized to the atomic weight of Carbon-12 (A=12, Z=6) with that weight being defined as 12.000 $\cdots$. Thus, a proton has an atomic weight of 1.00727648, a neutron has an atomic weight of 1.00866491, an electron has a atomic weight of 0.00054858, and a hydrogen atom (proton + electron) has an atomic weight of 1.00782504.

Nuclear binding energies are useful in a variety of ways. They can be used in calculations of energy production/absorption in nuclear reactions. They can be used to assist in predicting the kinds of reactions that a nucleus may undergo. A common form of expression is the **binding energy**

Figure 7–1. Periodic Table of the Elements.


KEY
ATOMIC NUMBER
86
(222)
SYMBOL
Rn
Radon
ATOMIC WEIGHT (OF MOST STABLE ISOTOPE IF IN PARENTHESES)
NAME

IA	IIA	IIIA		IVA	VA	VIA	VIIA	VIIIA			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIII
1 1.0079 H Hydrogen																		2 4.0026 He Helium
3 6.941 Li Lithium	4 9.0122 Be Beryllium												5 10.811 B Boron	6 12.011 C Carbon	7 14.007 N Nitrogen	8 15.999 O Oxygen	9 18.998 F Fluorine	10 20.180 Ne Neon
11 22.990 Na Sodium	12 24.305 Mg Magnesium												13 26.982 Al Aluminum	14 28.086 Si Silicon	15 30.974 P Phosphorous	16 32.066 S Sulfur	17 35.453 Cl Chlorine	18 39.948 Ar Argon
19 39.098 K Potassium	20 40.078 Ca Calcium	21 44.956 Sc Scandium		22 47.88 Ti Titanium	23 50.942 V Vanadium	24 51.996 Cr Chromium	25 54.938 Mn Manganese	26 55.847 Fe Iron	27 58.933 Co Cobalt	28 58.69 Ni Nickel	29 63.546 Cu Copper	30 65.39 Zn Zinc	31 69.72 Ga Gallium	32 72.61 Ge Germanium	33 74.922 As Arsenic	34 78.96 Se Selenium	35 79.904 Br Bromine	36 83.80 Kr Krypton
37 85.468 Rb Rubidium	38 87.62 Sr Strontium	39 88.906 Y Yttrium		40 91.224 Zr Zirconium	41 92.906 Nb Niobium	42 95.94 Mo Molybdenum	43 (98) Tc Technetium	44 101.07 Ru Ruthenium	45 102.91 Rh Rhodium	46 106.42 Pd Palladium	47 107.87 Ag Silver	48 112.41 Cd Cadmium	49 114.82 In Indium	50 118.71 Sn Tin	51 121.75 Sb Antimony	52 127.60 Te Tellurium	53 126.90 I Iodine	54 131.29 Xe Xenon
55 132.91 Cs Cesium	56 137.33 Ba Barium	57 138.91 La Lanthanum	58-71 *	72 178.49 Hf Hafnium	73 180.95 Ta Tantalum	74 183.85 W Tungsten	75 186.21 Re Rhenium	76 190.2 Os Osmium	77 192.22 Ir Iridium	78 195.08 Pt Platinum	79 196.97 Au Gold	80 200.59 Hg Mercury	81 204.38 Tl Thallium	82 207.2 Pb Lead	83 208.98 Bi Bismuth	84 (209) Po Polonium	85 (210) At Astatine	86 (222) Rn Radon
87 (223) Fr Francium	88 226.03 Ra Radium	89 227.03 Ac Actinium	90-103 **	104 (261) Rf Rutherfordium	105 (262) Db Dubnium	106 (263) Sg Seaborgium	107 (262) Bh Bohrium	108 (269) Hs Hassium	109 (266) Mt Meitnerium	110 (269) Ds Darmstadtium	111 (272) Uuu Ununurnium	112 (281) Uub Ununbium	113 --- Uut Ununtrium	114 (289) Uuq Ununquadium	115 --- Uup Ununpentium	116 (289) Uuh Ununhexium	117 --- Uus Ununseptium	118 --- Uuo Ununoctium

LANTHANIDES *	58 140.12 Ce Cerium	59 140.91 Pr Praseodymium	60 144.24 Nd Neodymium	61 (145) Pm Promethium	62 150.4 Sm Samarium	63 151.97 Eu Europium	64 157.25 Gd Gadolinium	65 158.93 Tb Terbium	66 162.50 Dy Dysprosium	67 164.93 Ho Holmium	68 167.26 Er Erbium	69 168.93 Tm Thulium	70 173.04 Yb Ytterbium	71 174.97 Lu Lutetium
ACTINIDES **	90 232.04 Th Thorium	91 231.04 Pa Protactinium	92 238.03 U Uranium	93 237.05 Np Neptunium	94 (244) Pu Plutonium	95 (243) Am Americium	96 (250) Cm Curium	97 (247) Bk Berkelium	98 (251) Cf Californium	99 (252) Es Einsteinium	100 (257) Fm Fermium	101 (258) Md Mendelevium	102 (259) No Nobelium	103 (262) Lr Lawrencium

per nucleon which is the approximate energy that must be supplied to remove a single nucleon from a nucleus. In practice there are differences between removing a proton or a neutron and there are differences related to the stability of the resulting nucleus. However, binding energy per nucleon is useful for thinking purposes.

The **mass defect** is the difference in mass between the isolated particles that might make up a specific nucleus and the mass of that nucleus [1]:

$$\begin{aligned}\Delta M &= Z\left(m_e + m_p\right) + Nm_n - M_{nucleus} \\ &= Zm_H + Nm_n - M_{nucleus}\end{aligned} \tag{7.1}$$

where the masses are usually expressed in atomic mass units (amu) based on the mass of Carbon-12. The total energy difference is the product of Δ*M* and the energy of 1 amu (= 931.4943 MeV)

$$\begin{aligned}\Delta E &= 931.4943\left(Zm_H + Nm_n - M_{nucleus}\right) \\ &= 931.4943\left(1.00782504Z + 1.00866491N - M_{nucleus}\right)\end{aligned} \tag{7.2}$$

This energy is equal to the total binding energy, thus the binding energy per nucleon is

$$\frac{\Delta E}{A} = \left(\frac{931.4943}{A}\right)\left(1.00782504Z + 1.00866491N - M_{nucleus}\right). \tag{7.3}$$

For example, consider Helium-4, with $M_{nucleus}=4.00260323$, $Z=2$, $N=2$, and $A=4$, then

$$\begin{aligned}\frac{\Delta E}{A} &= \left(\frac{931.4943}{4}\right)\left(1.00782504(2) + 1.00866491(2) - 4.00260323\right) \\ &= 7.07392 \text{ MeV / nucleon}\end{aligned}$$

Several other noteworthy examples are deuterium (Hydrogen-2) with $M_{nucleus}=2.01410178$, $Z=1$, $N=1$, and $A=2$,

$$\begin{aligned}\frac{\Delta E}{A} &= \left(\frac{931.4943}{2}\right)\left(1.00782504(1) + 1.00866491(1) - 2.01410178\right) \\ &= 1.11228 \text{ MeV / nucleon}\end{aligned}$$

Uranium-235 with $M_{nucleus}=235.043924$, $Z=92$, $N=143$, and $A=235$,

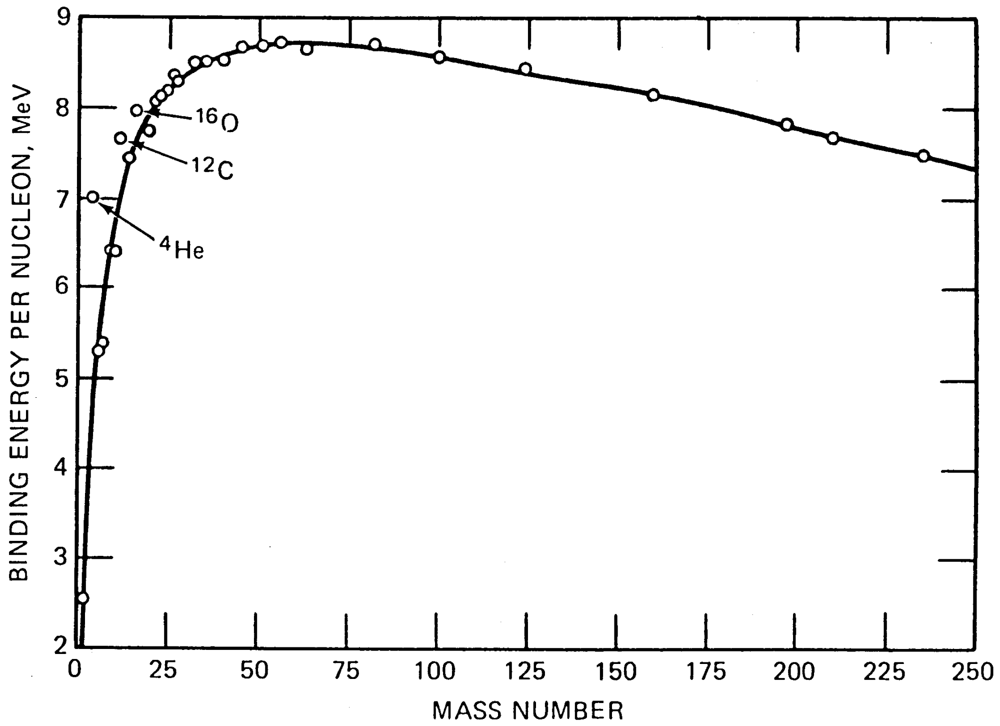
$$\frac{\Delta E}{A} = \left(\frac{931.4943}{235}\right)\left(1.00782504(92) + 1.00866491(143) - 235.043924\right)$$
$$= 7.59093 \text{ MeV / nucleon}$$

and Iron-56 with $M_{nucleus}$*=55.934940, Z=26, N=30,* and *A=56,*

$$\frac{\Delta E}{A} = \left(\frac{931.4943}{56}\right)\left(1.00782504(26) + 1.00866491(30) - 55.934940\right)$$
$$= 8.79028 \text{ MeV / nucleon}$$

A complete presentation of the results of Eq. (7.3) is shown in Figure 7-2 which shows the binding energy per nucleon for the longest lived nuclides at each mass number. The curve is relatively smooth with a few prominent bumps. The most noticeable bump is at Helium-4. The very high binding energy means that alpha particles are among the most stable of all light particles. It is for this reason that they show up as radioactive decay products. It is more energetically favorable for many nuclei to emit an alpha particle than it is for them to emit a proton, neutron, deuteron, triton, or other similar particle. The peak of the binding energy curve occurs near the value of A=56. Iron-56 is among the most tightly bound of all nuclei.

Figure 7-2. The Curve of Binding Energy.[1]



The curve of binding energy is most enlightening when studied in detail, predicting the energetic favorability of nuclear fission and the iron/nickel/cobalt endpoint of stellar fusion processes. We will return to this curve in later discussions.

Nuclear Particle Reactions

A nuclear particle reaction is any process in which an incident nuclear particle (electron, gamma ray, neutron, etc.) interacts with either a target nuclear particle or, more likely, with a target atom or nucleus to cause a change in one or all of the participants. Reactions may produce entirely different particles, they may cause particles to disappear, or they may cause an internal change in energy of one or more of the participants. A number of the more commonly experienced nuclear reactions are identified and defined below.

The strength of many of these reactions are described by reaction cross sections. The **cross section** (σ) for a nuclear reaction is the effective area (cm^2) that an atom, nucleus, or particle presents to an incident particle with respect to "colliding and undergoing a specific nuclear reaction". Cross sections for nuclear reactions are usually very small. A **barn** is a unit of cross section equal to 1 x 10^{-24} cm^2, that has been found to be particular relevant to nuclear reactions. A **differential cross section** is a distribution of cross section versus some parameter, e.g., scattering angle. The units of an angular differential cross section are cm^2/sr vs. θ.

Elastic scattering is a scattering process in which the incident particle preserves its identity throughout the scattering event, and in which significant momentum but limited energy is transferred from the incident particle to the target nucleus. The target nucleus is not excited to any higher internal energy in an elastic scattering process. The actual energy lost is a function of the angle at which the particle is scattered. The ratio of scattered particle energy to incident particle energy is given by the relation [1]

$$\frac{E_{scattered}}{E_{incident}} = \frac{A^2 + 2AM\cos\theta + M^2}{\left(A+M\right)^2} \tag{7.4}$$

where M is the mass of the incident particle, A is the mass of the target nucleus, and θ is the angle of scattering relative to the incident direction. The maximum energy loss occurs for θ = 180° and is given by

$$\left.\frac{E_{scattered}}{E_{incident}}\right|_{MIN} = \frac{A^2 + 2AM(-1) + M^2}{\left(A+M\right)^2} = \frac{\left(A-M\right)^2}{\left(A+M\right)^2} \tag{7.5}$$

Although we have stated that the energy loss is limited, this is really only true for heavy target nuclei. If *A=100M,* then the minimum ratio of scattered energy to incident energy is 0.96. However, if we are interested in elastic neutron scattering from hydrogen atoms, then *A~M* and the energy lost can range from zero for small-angle scattering to virtually all of the initial energy for backwards scattering. The large elastic scattering energy losses in hydrogen make it of interest in neutron moderation and neutron shielding applications.

Elastic scattering of gamma rays is called **Rayleigh scattering** or **coherent scattering**. The energy of the scattered gamma is essentially unchanged. Coherent scattering is only important at gamma energies below roughly 10 keV. Even in this low energy regime coherent scattering is usually very small compared to photoelectric absorption. Electron binding energies are usually so small that inelastic scattering processes are energetically favorable and overshadow coherent scattering.

Inelastic scattering is a scattering process in which the incident particle preserves its identity throughout the scattering event, and in which significant energy and momentum is transferred from the incident particle to the target nucleus. The target nucleus is left in an excited energy state, although the excitation energy may be rapidly radiated away as gamma rays.

Compton scattering is the inelastic scattering of gamma rays from electrons bound in atoms. The gamma ray scatters with significant energy loss; the lost energy is transferred to an electron which gains sufficient energy to be ejected from the target atom. The energy of the scattered gamma $h\nu$' is related to the energy of the incident gamma $h\nu$ by [2]

$$h\nu' = \frac{h\nu}{1+\left(h\nu / mc^2\right)\left(1-\cos\theta\right)} \tag{7.6}$$

where m is the mass of an electron, θ is the angle through which the gamma ray is scattered, and ν is the frequency of the incident photon. The recoil electron has energy given by

$$E_{RECOIL} = h\left(\nu - \nu'\right) - B.E. \tag{7.7}$$

where $B.E.$ is the binding energy of the electron (usually negligible). The differential cross section for scattering through an angle θ is given by the Klein-Nishina formula

$$\frac{d\sigma}{d\Omega}\left(\text{in cm}^2 / \text{sr}\right) = \frac{e^4}{32\pi^2 \varepsilon_0^2 m^2 c^4}\left(\frac{\nu'}{\nu}\right)^2\left(\frac{\nu}{\nu'}+\frac{\nu'}{\nu}-\sin^2\theta\right). \tag{7.8}$$

Integrating this function over all angles θ gives the total Compton scattering cross section. The recoil electron deposits its energy in the general vicinity of its production. The scattered gamma ray usually leaves the vicinity before reacting further. The fraction of the total energy that is deposited in the vicinity of a Compton event is sometimes called **Compton absorption**. The Compton absorption coefficient is the product of the total Compton scattering attenuation coefficient and the fraction of the energy absorbed. The Compton scattering coefficient is the fraction of the energy carried away by the scattered gamma times the total Compton scattering attenuation coefficient. This will become clearer after reading the next section on cross sections and attenuation coefficients.

The **photoelectric effect** is a process in which x-rays or gamma rays are absorbed by electrons bound in atoms. The absorption process transfers sufficient energy to an electron for it to be ejected from the target atom. The energy of the ejected electron is equal to the incident gamma energy $h\nu$ minus the binding energy *B.E.* of the electron in its atomic state.

$$E_{PHOTOELECTRON} = h\nu - B.E. \tag{7.9}$$

In metals and covalent solids, the electrons are not bound to individual atoms but are distributed throughout the solid. In this case the binding energy should be replaced by the work function of the solid. At low gamma ray photon energies the atomic binding energies of the electrons play a significant role in the shape of the photoelectric cross section curves. In general the probability of photoelectric scattering is proportional to [3]

$$\sigma = \frac{2^{11/2}\pi}{3}\frac{Z^5 r_e^2 \alpha^4 \left(mc^2\right)^{7/2}}{h\nu\left(h\nu - \varepsilon\right)^{5/2}} \tag{7.10}$$

where Z is the nuclear charge, r_e is the classical electron radius, α is the fine structure constant, and ε is the binding energy of the electron to be ejected.

Pair production is a process in which a gamma ray collides with any particle and disappears with the simultaneous creation of an electron and a positron. Only gamma rays with energies in excess of 1.022 MeV are capable of pair production. The electron and the positron each possess an initial kinetic energy that is equal to one-half the difference between the photon energy $h\nu$ and 1.022 MeV (the rest mass-energy of an electron and a positron)

$$E_{e^- \text{ or } e^+} = \tfrac{1}{2}\left(h\nu - 1.022 \text{ MeV}\right). \tag{7.11}$$

The cross section for pair production is not easily calculated. However, at energies greater than the 1.022 MeV threshold but below extreme relativistic energies, the cross section will take the form [4]

$$\sigma = \alpha Z^2 r_e^2 \left(\frac{h\nu - 2mc^2}{h\nu}\right)^3 F(\nu) \tag{7.12}$$

where $F(\nu)$ is a complicated function that varies from about 2 at low energies to around 10 at energies around 100 MeV.

All charged particles can interact with materials through the processes of **ionization** and **atomic excitation.** The electric field of the charged particle interacts with an electron in the material imparting a small energy to it. This energy may be sufficient to cause the electron to be ejected (ionization) or to be elevated to a higher internal energy level (atomic excitation). Both processes are most important when the charged particle energy is low, but obviously cannot occur if the charged particle energy is less than the excitation or ionization energy. The cross section for atomic excitation is given approximately by [5]

$$\sigma_{n \to m} = \frac{4\pi z^2 \alpha^2 \mu c^2}{E} \left| r_{nm} \right|^2 \ln\left(\frac{4E}{\Delta E} \right) \tag{7.13}$$

where E is the kinetic energy of the charged particle, z is the charge of the incident particle, μ is the reduced mass of the incident particle, $|r_{mn}|^2$ is the electric dipole matrix element connecting the initial atomic state n to final atomic state m, and ΔE is the $n \to m$ transition energy or the ionization energy, as appropriate. The reduced mass of a particle of mass M_1 interacting with a particle of mass M_2 is given by

$$\mu = \frac{M_1 M_2}{M_1 + M_2} . \tag{7.14}$$

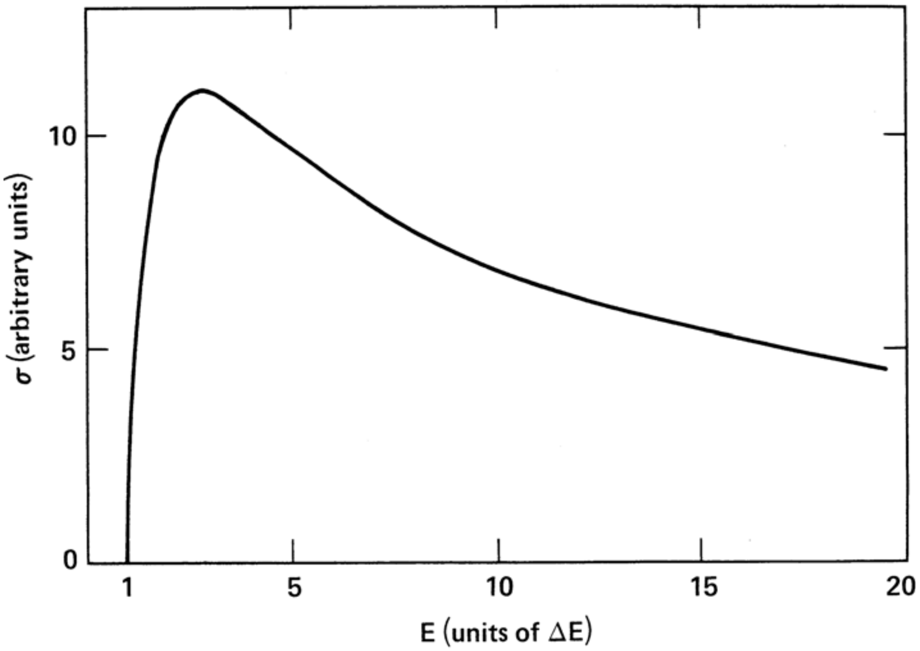
The cross section for ionization is essentially the same as Eq. (7.13) except that the bound state-bound state transition matrix element is replaced by a bound state-free electron matrix element. The energy dependence will be essentially the same.

The qualitative behavior of excitation/ionization is shown in Figure 7-3. There is a threshold at $E = \Delta E$, with a steep rise to a maximum, followed by a slow fall that is roughly proportional to $1/E$. The appearance of a threshold is required by energy conservation (an atom cannot be excited to a state ΔE higher than its original state unless the incident photon has an energy of at least ΔE). Eq. (7.13) is not quantitatively valid in the region below and just above the maximum. The cross section expression was developed under an assumption of moderate electron energies. Note that the term $ln(4E/\Delta E)$ implies a threshold at $E = \Delta E/4$, which is not physically allowable. Eq. (7.13) should give reasonable results for electron energies above a few hundred electron volts.

The magnitude of the total excitation and or ionization would be determined by integrating over all possible transitions. The total cross sections can be estimated by replacing the matrix elements in Eq.(7.13) with a quantity ηa_0^2 where a_0 is the Bohr radius and η is a number of the order of unity which is theoretically derivable from atomic transition sum rules for the target atom.[6],[7] The values of η for excitation and ionization will not necessarily be the same. We can estimate what values of the peak cross section may be expected for electrons. Since μc^2 = 511 keV, if we assume E = 100 eV, ΔE = 10 eV and η ~1, then the result is σ ~ 3.5 x 10^{-16} cm^2. For heavier charged particles the cross sections will scale as z^2 and approximately as m. We would expect alpha particle cross sections to be roughly 3 x 10^4 larger than the corresponding electron cross sections.

Induced fission is a process in which an incident particle (usually a neutron, but possibly another kind of particle) is absorbed by a nucleus, so disrupting the nucleus that it breaks into two smaller nuclei with the simultaneous emission of a number of other particles (usually neutrons).

Figure 7-3. Energy dependence of the atomic excitation or ionization cross section.



Transmutation is a collection of reaction processes in which an incident particle is absorbed by a target nucleus, with or without another nuclear particle being ejected, and producing a different nucleus as a product. The product nucleus may or may not be in an excited internal energy state. Transmutation reactions include: **photonuclear** processes in which a gamma ray is absorbed by a nucleus and one or more neutrons or protons or both are ejected, **capture** reactions in which the incident particle is captured with no particles other than gamma rays being emitted, and **particle-particle** exchange reactions in which one particle is absorbed by the target nucleus and one or more other particles are ejected.

Nuclei and many nuclear particles may be described by the shorthand notation

$$ {}_{Z}^{A}T^{N} $$

where Z is the atomic number (number of protons), N is the number of neutrons, $A = N+Z$ is the mass number, and T is the chemical symbol for the element with atomic number Z. For completeness in describing most reactions of interest we may assume the following definitions:

$$ {}_{0}^{1}\mathrm{n}^{1} \;=\; \text{neutron} $$

$$ {}_{1}^{1}\mathrm{p}^{0} \;=\; \text{proton} \;=\; {}_{1}^{1}\mathrm{H}^{0} $$

$$ {}_{0}^{0}\gamma^{0} \;=\; \text{gamma ray} $$

A typical nuclear transmutation reaction can be described by the equation

$$ {}_{Z1}^{A1}T1^{N1}\begin{pmatrix}\text{Target}\\ \text{Nucleus}\end{pmatrix}+{}_{Z3}^{A3}T3^{N3}\begin{pmatrix}\text{Incident}\\ \text{Particle}\end{pmatrix}\rightarrow {}_{Z4}^{A4}T4^{N4}\begin{pmatrix}\text{Exiting}\\ \text{Particle}\end{pmatrix}+{}_{Z2}^{A2}T2^{N2}\begin{pmatrix}\text{Product}\\ \text{Nucleus}\end{pmatrix} \quad (7.15)$$

with associated conservation laws for *N*, *A*, and *Z* being expressed as

$$N1+N3=N4+N2 \quad (7.16a)$$

$$A1+A3=A4+A2 \quad (7.16b)$$

$$Z1+Z3=Z4+Z2 \quad (7.16c)$$

Often reaction (7.15) will be written using the shorthand

$$ {}_{Z1}^{A1}T1^{N1}\left(X,Y\right){}_{Z2}^{A2}T2^{N2} \quad (7.17)$$

where *X* is the incident particle and *Y* is the exiting particle.

Figure 7-4 lists the product nuclei relative to a standard target nucleus for most of the common transmutation reactions. [8] The shorthand described above is used extensively in this figure. Figure 7-4 also describes the daughter nuclei for a number of radioactive decay modes. As an illustration of the use of this figure, an α capture reaction (uppermost righthand corner of chart) produces the product nucleus with

$$ {}_{Z+2}^{A+4}\,T+2^{\;N+2} \qquad \alpha \text{ capture}$$

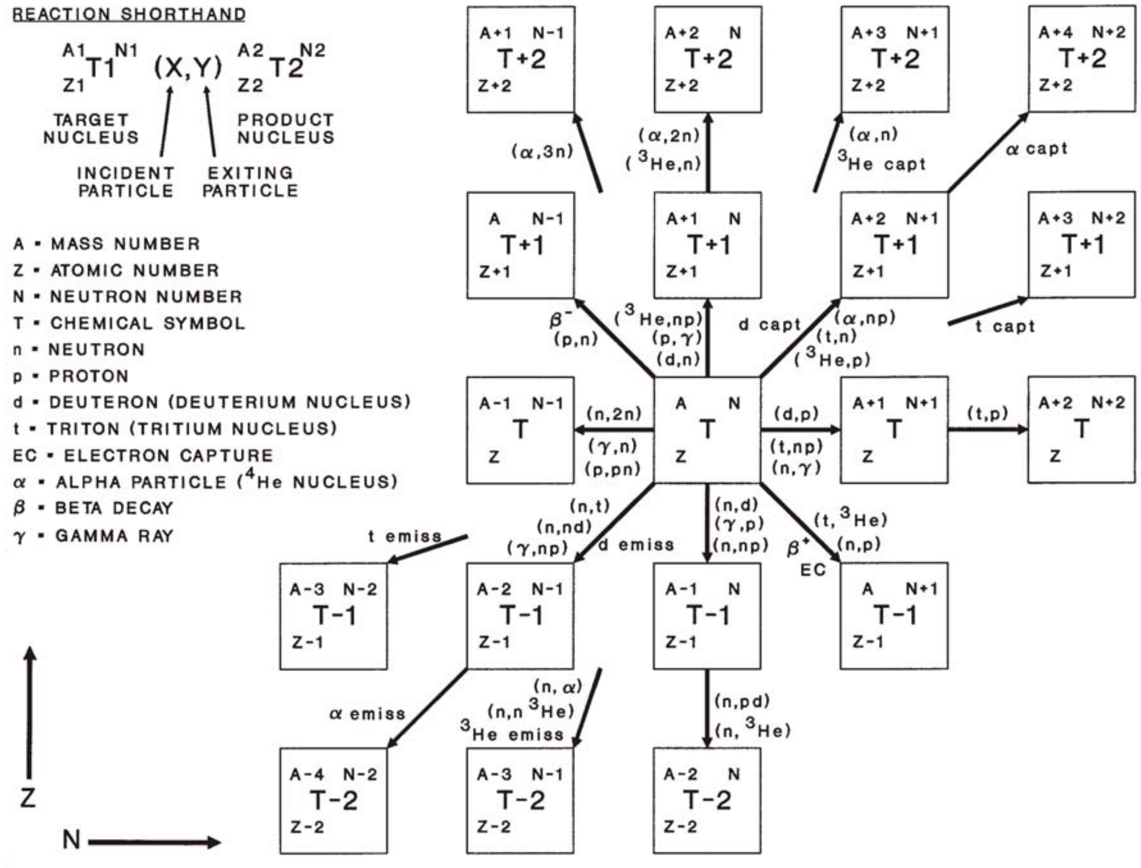
while either a (n,2n), a (γ,n), or a (p,pn) reaction (left-hand middle of chart) will result in a product nucleus with

$$ {}_{Z}^{A-1}T^{\;N-1} \qquad \text{(n,2n), } (\gamma\text{,n}), \text{ or (p,pn).}$$

Note that (p,pn) denotes that one proton is incident on the target while both a proton and a neutron are emitted. Alpha decay (lower left-hand corner of chart) will result in a product nucleus with

$$ {}_{Z-2}^{A-4}\,T-2^{\;N-2} \qquad \alpha \text{ decay.}$$

Figure 7-4. Common transmutation reaction products and radioactive decay products.[8]



Not all of the reactions described above occur for all particles and not all occur at all energies of those particles. In Table 7-2 we have summarized the dominant reactions of the more common nuclear particles and the energy ranges at which those reactions are usually important. The reactions may or may not be possible outside of those ranges. As discussed above, many reactions have thresholds below which they cannot occur.

Table 7-2. Dominant reactions of nuclear particles

PARTICLE	DOMINANT REACTIONS	ENERGY
PHOTON	ATOMIC EXCITATION	MeV AND BELOW
	RAYLEIGH SCATTERING	eV AND BELOW
	PHOTOELECTRIC ABSORPTION	100 keV AND BELOW
	COMPTON SCATTERING	keV AND ABOVE
	PAIR PRODUCTION	1 MeV AND ABOVE
	TRANSMUTATION – e.g. (γ,n), (γ,p)	MeV AND ABOVE
ELECTRON	ATOMIC EXCITATION	10 eV AND ABOVE
	ELECTRO-IONIZATION	keV AND ABOVE
	PAIR PRODUCTION	MeV AND ABOVE
	TRANSMUTATION – e.g. (e,n), (e,p)	MeV AND ABOVE
PROTON	ATOMIC EXCITATION	10 eV AND ABOVE
	IONIZATION	keV AND ABOVE
	ELASTIC SCATTERING	keV AND ABOVE
	INELASTIC SCATTERING	MeV AND ABOVE
	TRANSMUTATION – e.g. (p,n), (p,α)	MeV AND ABOVE
NEUTRON	ELASTIC SCATTERING	ALL ENERGIES
	INELASTIC SCATTERING	MeV AND ABOVE
	ABSORPTION (TRANSMUTATION)	ALL ENERGIES
	FISSION (HEAVY ELEMENTS ONLY)	MeV AND ABOVE (except fissile nuclei)
ALPHA	ATOMIC EXCITATION	10 eV AND ABOVE
	IONIZATION	keV AND ABOVE
	ELASTIC SCATTERING	keV AND ABOVE
	INELASTIC SCATTERING	MeV AND ABOVE
	TRANSMUTATION	MeV AND ABOVE

Nuclear Particle Reaction Rates

Before attempting to describe nuclear reaction rates a few definitions are in order. The number of particles passing through a unit area per unit time is called the **flux density**.

$$\textit{Flux density} = \phi = \text{ particles/sec/cm}^2 \qquad (7.18)$$

Fluence is the time integral of flux density. Therefore, fluence is the total number of particles which have passed through a unit area.

$$\textit{Fluence} = \text{particles/cm}^2 \qquad (7.19)$$

The reaction rate relative to a single target atom is the product of the flux density and the cross section

$$R_{atom} = \sigma\phi \, . \qquad (7.20)$$

The reaction rate per unit volume of target material is the product of the target atomic density (N) in atoms/cm^3 times the reaction rate per atom

$$R_{volume} = N\sigma\phi \, . \qquad (7.21)$$

The reaction rate per unit area occurring in a small distance element dx is the volume reaction rate times the distance element

$$R_{dx} = N\sigma\phi \, dx = -d\phi \, . \qquad (7.22)$$

This same rate is the rate at which particles are lost from an incident beam. That is, it is equal to the change in the flux density. Rearrangement of this equation gives the following differential equation for the flux density

$$\frac{d\phi}{dx} = -N\sigma\phi \qquad (7.23)$$

This kind of differential equation can be easily solved by rearranging to form the logarithmic derivative and integrating. That is

$$\frac{d\phi}{\phi} = d\ln\phi = -N\sigma \, dx \qquad (7.24)$$

and

$$\int_0^x d\ln\phi = \int_0^x -N\sigma \, dx \, . \qquad (7.25)$$

This has solutions

$$\ln \phi(x) - \ln \phi(0) = -N\sigma x \tag{7.26}$$

and

$$\phi(x) = \phi(0)e^{-N\sigma x} \; . \tag{7.27}$$

Thus the flux density is seen to decrease exponentially with propagation distance into the target. This type of behavior is characteristic of many kinds of attenuation processes.

The characteristic distance over which the flux density decreases is the 1/e length. That is, we set the exponential equal to 1/e and solve for the distance:

$$e^{-N\sigma x} = 1/e = e^{-1} \tag{7.28}$$

with logarithm

$$-N\sigma x = -1 \tag{7.29}$$

and subsequent solution

$$x = 1/N\sigma \; . \tag{7.30}$$

The 1/e characteristic length is commonly called the attenuation (or absorption) length. From Eq. (7.30) we see that the attenuation length is inversely proportional to the density of the target and inversely proportional to the reaction cross section.

Attenuation of Nuclear Radiation

The **macroscopic cross section** Σ is the product of the target atomic density (atoms/cm^3) and the cross section per atom. It is equivalent to the attenuation length.

$$\Sigma = N\sigma = \mu \tag{7.31}$$

The macroscopic cross section is also called the **linear attenuation coefficient** μ. Both quantities have units of cm^{-1}. The reciprocal of the linear attenuation coefficient is the average distance a particle travels in a material before reacting. The **mass attenuation coefficient** (μ/ρ) is the linear attenuation coefficient (μ in 1/cm)divided by the mass density (ρ in g/cm^3) of the material. It represents the total reaction cross section per unit mass of target material. It has units of cm^2/g.

$$\frac{\mu}{\rho} = N\sigma / \rho \tag{7.32}$$

Equations (7.8) and (7.12) have alternate forms which use the mass attenuation coefficient. The alternate forms are

$$\frac{d\phi}{dx} = -\left(\frac{\mu}{\rho}\right)\rho\phi \tag{7.33}$$

and

$$\phi(x) = \phi(0)e^{-(\mu/\rho)\rho x} . \tag{7.34}$$

Just as we saw in the case of electromagnetic radiation, attenuation (or extinction) of nuclear radiation has two parts: absorption and scattering. Scattering takes radiation moving in one (usually fixed) direction and converts it into radiation moving in another (usually random) direction. Scattering removes radiation from a beam of radiation but does not eliminate it from the total environment. Absorption converts radiation into absorbed energy. Radiation is removed not only from a beam but from the total environment. The removed energy is absorbed within the "absorbing" material and converted to internal atomic and nuclear excitations and ultimately heat. The total attenuation coefficient can be written as the sum of the scattering and absorption components:

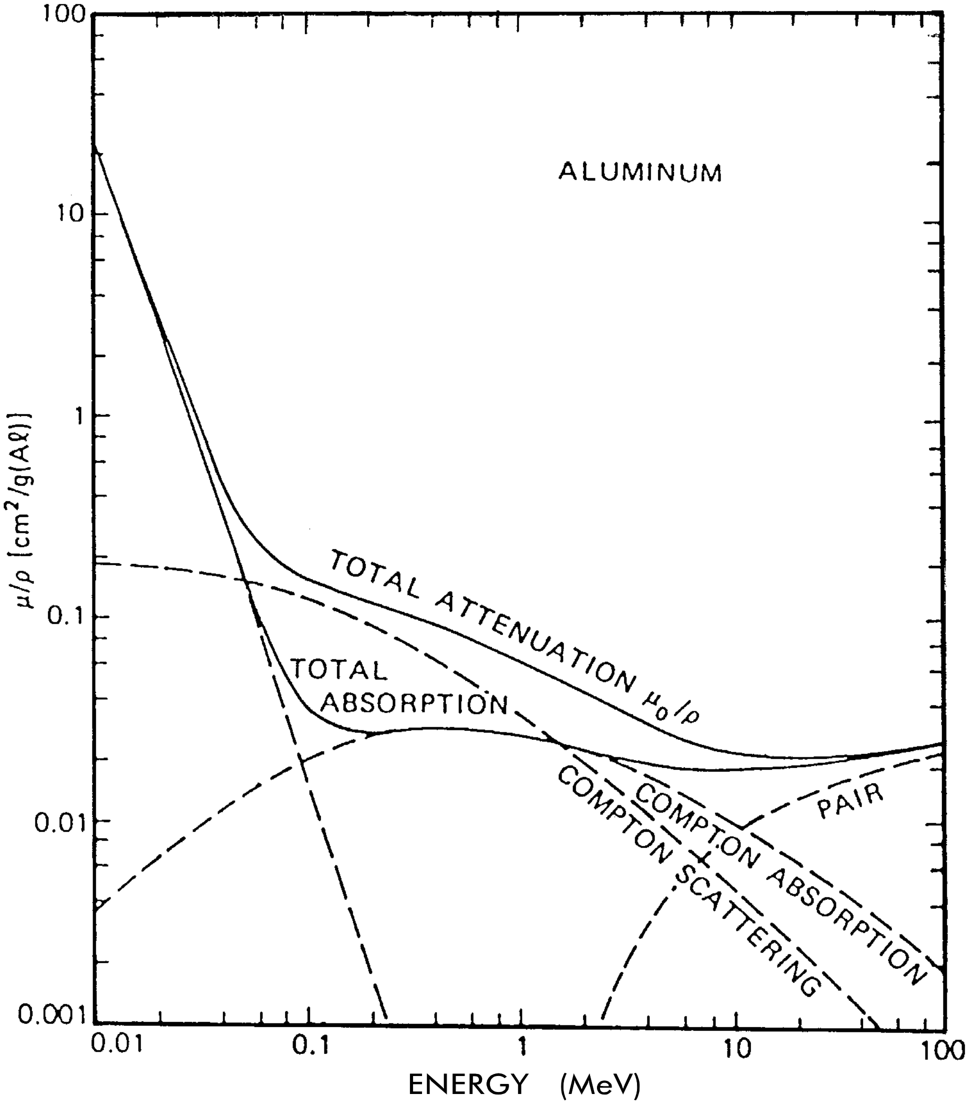
$$\mu_{ATTENUATION} = \mu_{SCATTERING} + \mu_{ABSORPTION} \tag{7.35}$$

and

$$\left(\frac{\mu_{ATTENUATION}}{\rho}\right) = \left(\frac{\mu_{SCATTERING}}{\rho}\right) + \left(\frac{\mu_{ABSORPTION}}{\rho}\right). \tag{7.36}$$

Some interaction processes will contribute to absorption, some contribute to scattering, and some contribute to both (an inelastic scattering process scatters some energy out of the beam, while some of the energy is deposited in the target medium). For example, Figure 7-5 shows the processes important to gamma ray scattering in aluminum in the energy range 10 keV to 100 MeV. The total (mass) attenuation coefficient is the sum of contributions from photoelectric absorption, Compton scattering and Compton absorption, and pair production. The total (mass) absorption coefficient lacks any contribution from Compton scattering but does have a contribution from the other three processes.[9]

Figure 7-5. Gamma ray interactions in aluminum and the mass absorption and mass attenuation coefficients.[9]



Computer codes exist which are capable of generating data on the total as well as individual contributions to the attenuation coefficient.[10] The National Institute of Standards and Technology has published (on the Internet) tabular and graphical data on the mass attenuation and mass energy absorption coefficients for all of the elements and a large number of materials of interest in radiation shielding, radiation measurement, and radiation health physics.[11] Figure 7-6 compares the mass attenuation coefficient and the mass energy absorption coefficient for three common materials (dry air, water, and iron). Note the influence of the K-edge (onset of photoelectric emission of K-shell electrons) on the iron coefficients. The K-edges of elements in air and water occur at photon energies below those plotted in the graphs.

If one is interested in determining the flux of radiation passing through a surface after an initial flux has passed through a mass of attenuating material, neither the attenuation coefficient nor the absorption coefficient will give the right answer. Use of the absorption coefficient will underestimate the "attenuation" of the radiation flux, but use of the attenuation coefficient will neglect the effect of small angle scattering and multiple scattering. A single scattering will cause a particle to be directed out of the beam, but additional scatterings may cause it to be directed back in the general direction

Figure 7-6. Comparison of the total gamma ray mass attenuation coefficient and the energy absorption coefficient for selected materials [11]:

a) dry air,

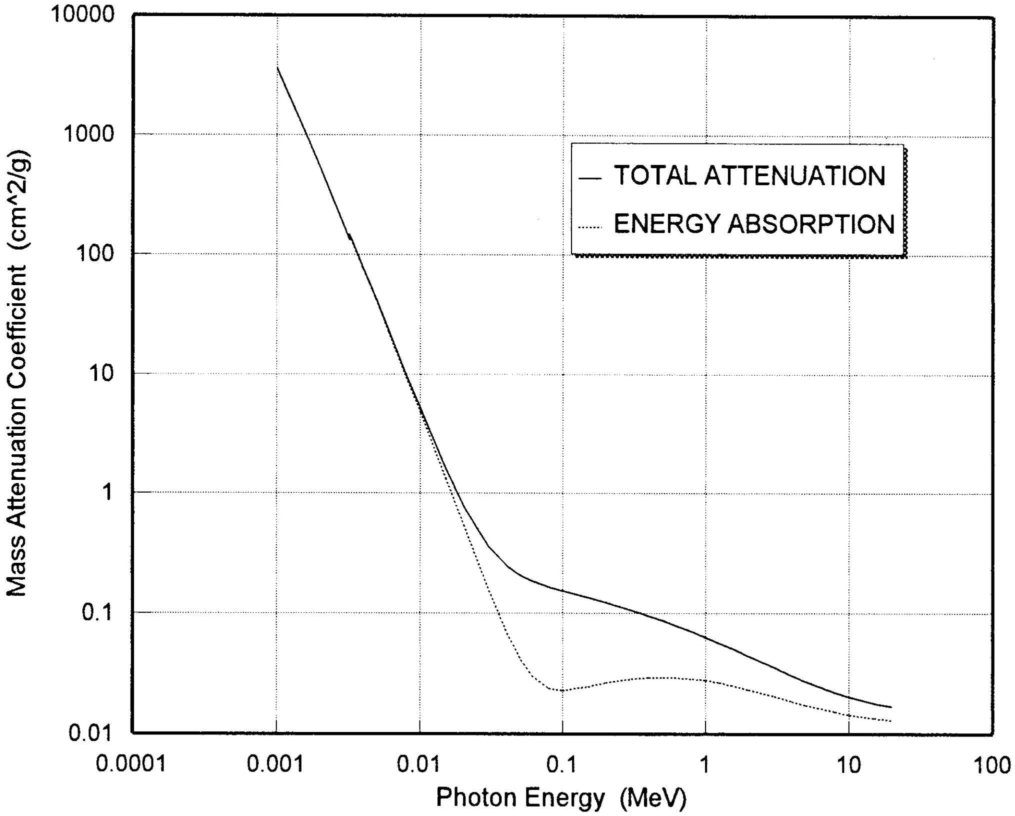
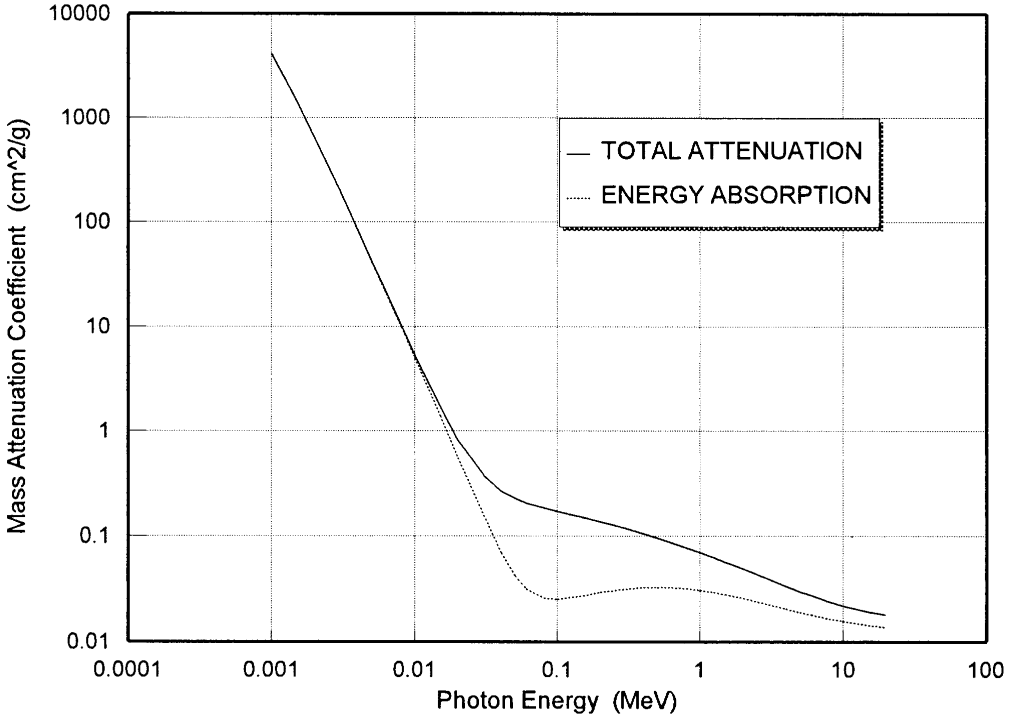
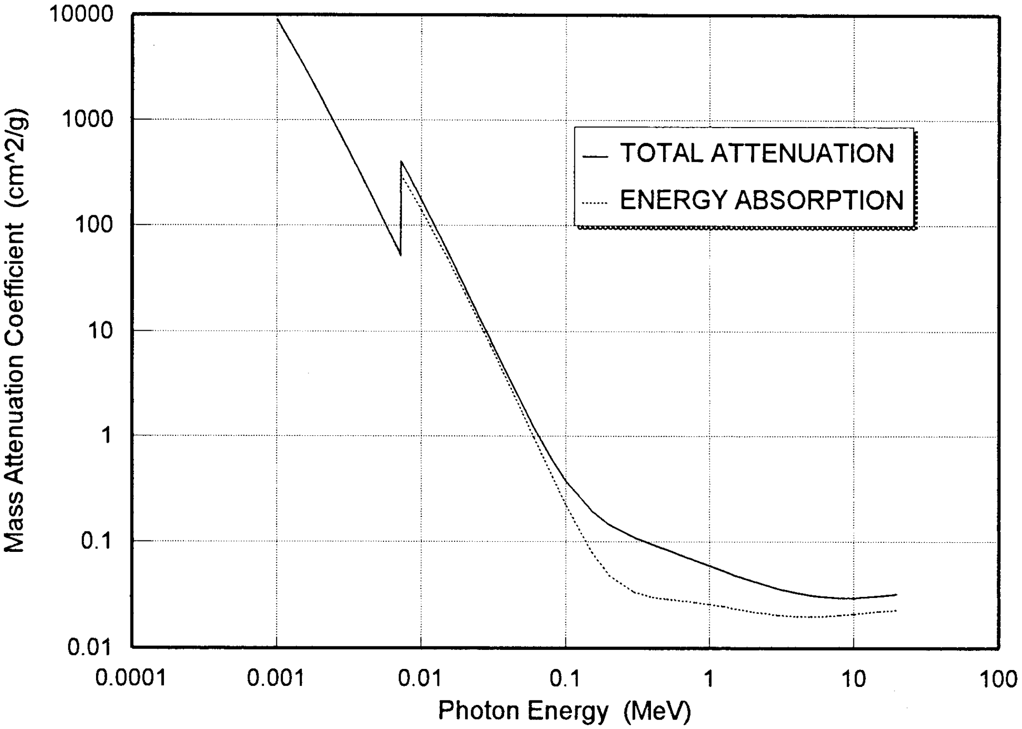


Figure 7-6 (continued). Comparison of the total gamma ray mass attenuation coefficient and the energy absorption coefficient for selected materials:

b) water,



c) iron.



of the measurement surface. If the particle is only scattered through a small angle, it may not move laterally by enough distance to miss the measurement surface. This results in a “buildup” factor *B* such that

$$\phi(x) = \phi(0) B(r, r_T) e^{-(\mu/\rho)\rho x} \tag{7.37}$$

where $B > 1$ is a function of the distance *r* from the scattering region and the transverse distance r_T away from the center of the beam. The farther the measurement point is from the attenuating layer, the less important either of these scattering effects becomes and *B* approaches unity. The interested reader is referred to texts on reactor engineering [1] and/or radiation shielding for detailed discussions of determining buildup factors.

The primary neutron reactions at energies below 10 MeV are elastic scattering, absorption (transmutation to a new isotope usually followed by gamma emission), and fission. Because the neutron only interacts via the short-range nuclear force, most neutron interactions are many-body problems and theoretical expressions for cross sections are usually suspect. However, the experimental nuclear physics community has compiled empirical data on virtually every element, isotope, and reaction of interest. There are a number of data compilations foremost among which may be the Evaluated Neutron Data Files (ENDF). Today these are readily accessible via the Internet in either tabular or graphical form.[12]

Figure 7-7 shows the dominant neutron cross sections for reactions with Aluminum-27 nuclei as generated by ENDFPLOT (one of the on-line graphical routines).[12] Elastic scattering and absorption (with subsequent gamma emission) are strong at all neutron energies between thermal energy (0.025 eV) through 10 MeV. Inelastic scattering (leaving the Al-27 nucleus in its first excited state rapidly becomes significant above a threshold of about 850 keV. The (n,α) transmutation reaction begins to become significant at neutron energies above 5 MeV. Figure 7-8 shows the neutron cross sections for the fissile nucleus Uranium-235. Being fissile, the fission cross section is greater than the absorption cross section at all energies. The extremely large number of resonances (narrow local peaks in the cross section) is evident in both the fission and absorption cross sections. The elastic scattering cross section becomes the dominant cross section at energies above a few keV. No other neutrons reactions have any appreciable cross section below 10 MeV.

An individual gamma ray or neutron tends to deposits much of its energy at individual points in a medium. That is, when a gamma ray is absorbed by the photoelectric effect, all of its energy is transferred to the photoelectron. Charged particles tend to deposit their energy gradually as they move through a medium. The **linear energy transfer (LET)** of a charged particle is the amount of energy deposited locally in a medium per unit length of travel. The LET may be determined from the ionization and excitation cross section expressions.

Figure 7-7. Neutron cross sections for Aluminum-27.[12]

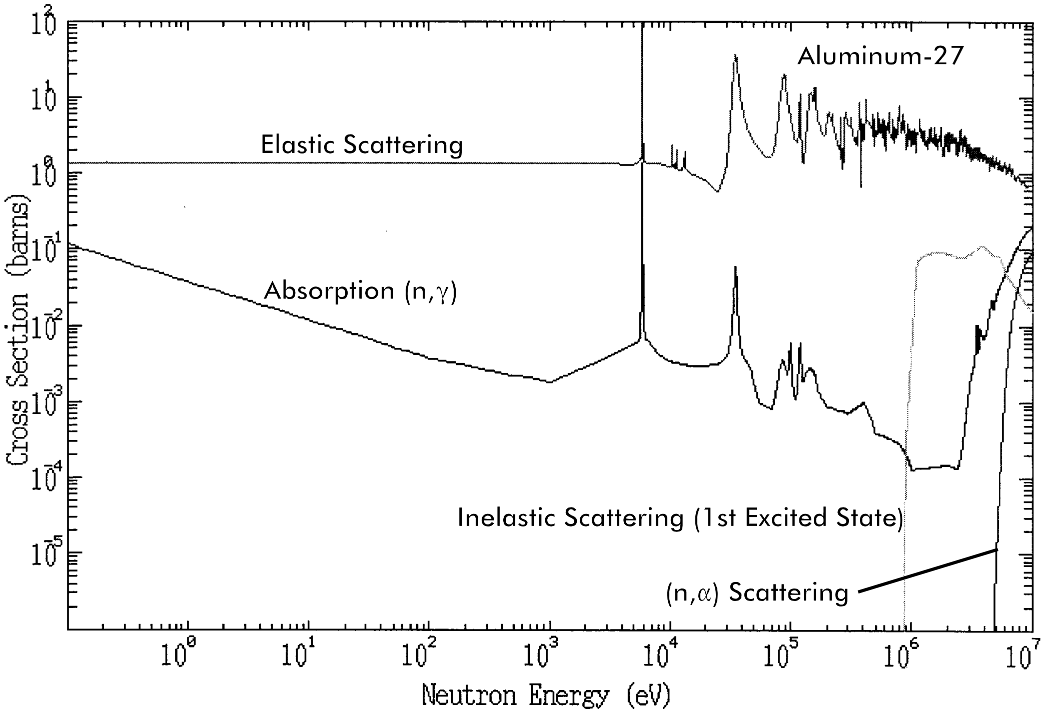
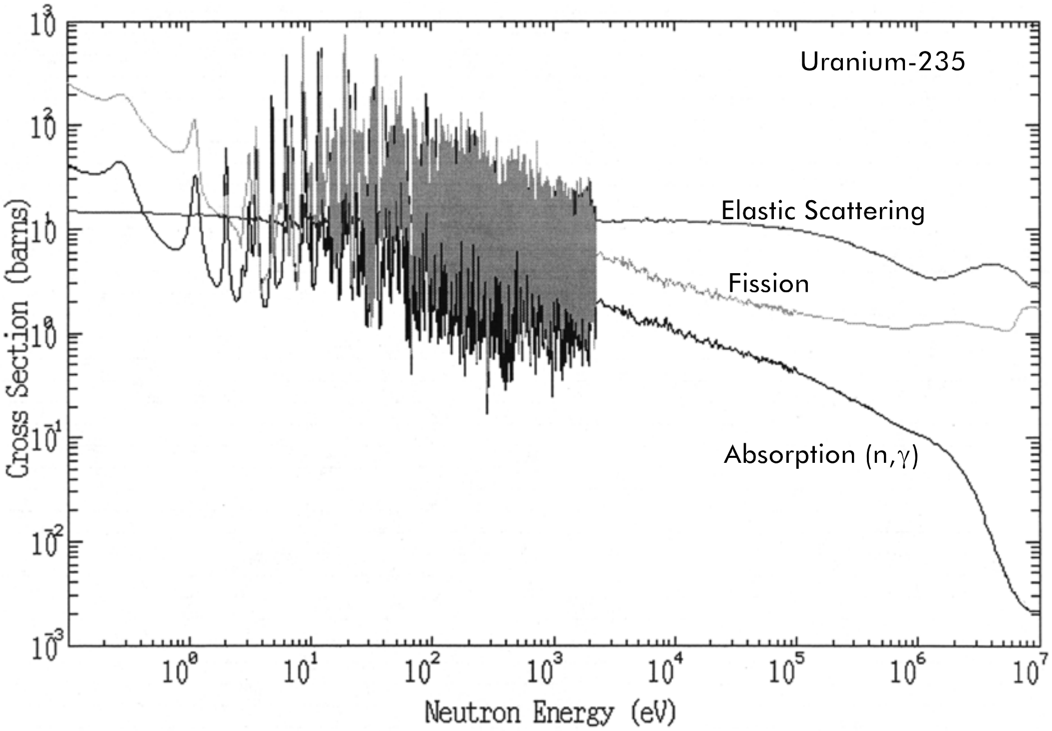


Figure 7-8. Neutron cross sections for Uranium-235.[12]



The rate of excitation of atoms on the transition n→m is given by

$$\left.\frac{dN_m}{dt}\right|_{n\to m} = \int dE\ N_n\ \sigma_{n\to m}(E)\,\Phi(E) \tag{7.38}$$

where $\Phi(E)$ is the charged particle flux as a function of kinetic energy. $\Phi(E)$ will take different forms depending on the source and the collisional history of the charged particles. The rate of ionization will look similar.

For example, consider a single non-relativistic, high-energy electron (energy $m_ec^2 >> E_0 >> \Delta E$ for all possible transitions). The average rate at which the electron collides inelastically with a collection of atoms of number density ρ (cm$^{-3)}$ is

$$\text{Rate} = \rho\bar{\sigma}v \tag{7.39}$$

where v is the electron velocity and $\bar{\sigma}$ is the cross section integrated over all possible transitions

$$\bar{\sigma} = \sum_n \frac{N_n}{N_{TOTAL}} \sum_m \frac{\int dE\ \sigma_{n\to m}(E)\,\Phi(E)}{\int dE\ \Phi(E)} \tag{7.40}$$

If, on the average, the electron loses energy $\bar{E}$ in each collision, whether by excitation or ionization, the time rate of energy loss is given by

$$\frac{dE}{dt} = -\,\bar{E}\rho\bar{\sigma}v \tag{7.41}$$

Since the electron moves at velocity v, the spatial variation of the energy loss is

$$\frac{dE}{dx} = \frac{dE}{d(vt)} = -\,\bar{E}\rho\bar{\sigma} \tag{7.42}$$

Because each $\sigma_{n\to m}$ is inversely proportional to E at high energies (we will ignore the *ln E* term in the cross section), $\bar{\sigma}$ must also be inversely proportional to E. Letting $\bar{\sigma} = \bar{\sigma}_A / E$, we find

$$\frac{dE}{dx} = -\bar{E}\rho\bar{\sigma}_A / E \tag{7.43}$$

Upon integration we obtain

$$E(x) = \left(E_0^2 - 2\bar{E}\rho\bar{\sigma}_A x\right)^{1/2} \tag{7.44}$$

or

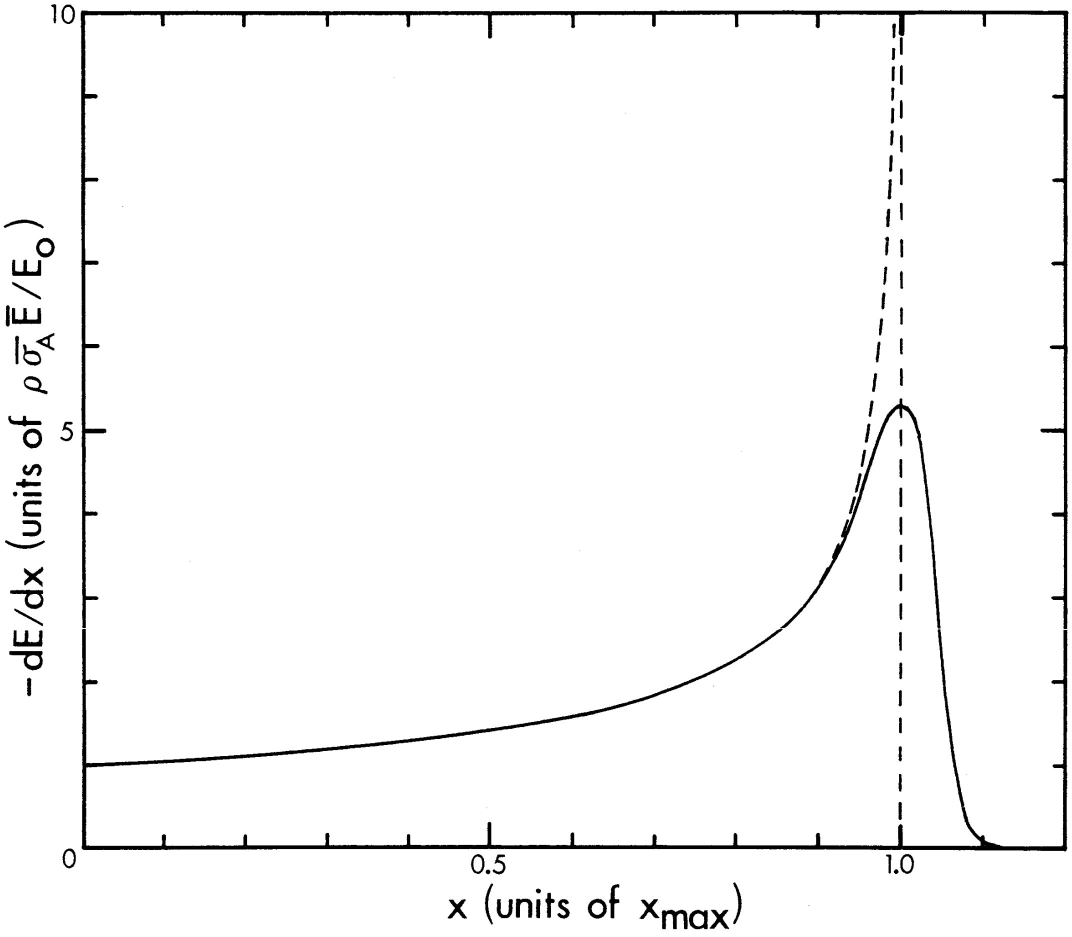
$$\frac{dE}{dx} = \frac{-\bar{E}\rho\bar{\sigma}_A}{\left(E_0^2 - 2\bar{E}\rho\bar{\sigma}_A x\right)^{1/2}} \tag{7.45}$$

The spatial dependence of the energy deposition is shown in Figure 7-9. The dashed curve shows the result of Eq. (7.45), while the solid curve is indicative of the behavior in a real system. Nature usually figures out a way to eliminate singularities. In this case the discrepancy between the real and calculated results is due to the combined effects of neglected processes including elastic scattering, electroionization, and the probabilistic nature of the scattering process. In any case, the energy loss rate (equivalent to the total excitation rate) is relatively constant at first, then rises quickly to a mazimum near the value

$$x_{max} = E_0^2 / 2\bar{E}\rho\bar{\sigma}_A \tag{7.46}$$

and then falls rapidly to zero. The peak exhibited by Figure 7-9 is commonly called the **Bragg** peak. The result Eq. (7.46) is obtained by setting *E(x)=0* in Eq. (7.44) and solving for *x*. If we begin with a nearly monoenergetic beam of high energy electrons, such as may be produced by a particle accelerator or high voltage electron beam source, then for much of the way into the medium on the way towards x_{max}, Φ*(E)* will remain nearly monoenergetic with the energy being given by Eq. (7.44). As the beam nears x_{max} it will lose its monoenergetic character, gradually becoming thermal in character somewhat past x_{max}. The LET of the charged particle is given by Eq. (7.45). The **range** of the charged particle in the material is given by Eq. (7.46).

Figure 7-9. Energy deposition by a charged particle.



Radiation Exposure and Dose

Exposure is the amount radiation present at the site of an individual or piece of equipment. We measure exposure in terms of Roentgens. One Roentgen (R) produces 2.08 x 10^9 ion pairs (an ion pair is an electron and a positive charged ion) per cm^3 of dry air or a total deposited charge of 2.58 x 10^{-4} Coulombs per kg of dry air.

$$1\ R = 2.08 \times 10^{9}\ \text{ion pairs / cm}^{3} = 2.58 \times 10^{-4}\ \text{C / kg (dry air)} \qquad (7.47)$$

There is no SI unit of exposure. Technically, Roentgen units apply only to x-rays or gamma radiation. However, this does not prevent them from being used for other forms of ion-producing radiations. Exposure is an indication of the potential for hazardous conditions to exist. A radiation source of a certain specified strength will always produce the same certain exposure when present for a specified period of time at a specified distance from the volume of interest. Because time is of interest in an exposure, a more common source description is in Roentgens per hour (R/hr). When objects or personnel are given an exposure to a certain amount of radiation, they absorb a fraction of the total energy. What doesn't get absorbed cannot produce an effect. The amount of radiation absorbed per unit volume is called the **dose**. Exposure is a measure of what effects might be, dose is a measure of what the effects will likely be. **Dose rate** is the dose accumulated per unit time.

Dose is measured in SI units of Grays. One Gray (Gy) is 1 J of absorbed energy per kilogram of material. A older unit, the rad, is still used and is equal to 100 ergs absorbed per gram of material. One Gray equals 100 rads. For x-rays, gamma radiation, and beta radiation, it has be found that an exposure of 1 Roentgen will produce a dose of 1 rad (1 cGy) in a person. Exposure to so many Roentgens or a dose of so many rads will produce a specific level of effect. However, if the radiation is something other than gamma or beta radiation, it has been found that different levels of effects are observed. Indeed, it has been found that fast neutrons are far more destructive than gamma rays. This is due to the heavy mass of the neutron imparting considerable recoil energy (relative to gamma rays) to nuclei in the vicinity of its interaction. Recoiling nuclei can break chemical bonds and cause substantial cellular damage. For this reason exposure to one "Roentgen" of neutron radiation has the same biological effect as exposure to perhaps ten Roentgens of x-rays. This has led to the development of the concept of **dose equivalent**. If absorbed dose is measured in rads, then the dose equivalent is measured in rems (Roentgen equivalent man) where

$$\textit{Dose Equivalent}\ \text{(rems)} = \textit{Absorbed Dose}\ \text{(rads)} \times \textit{RBE} \qquad (7.48a)$$

or

$$\textit{Dose Equivalent}\ \text{(Sieverts)} = \textit{Absorbed Dose}\ \text{(Grays)} \times \textit{RBE} \qquad (7.48b)$$

where *RBE* is the relative biological effectiveness. Many texts now use the term "**quality factor (QF)**" instead of *RBE*, but the meaning is the same same. The SI unit of dose equivalent is the Sievert (Sv). In general, the *RBE* is a function of the particle type and the particle energy. *RBE* values can be estimated from Table 7-3. [13]

Table 7-3. Relative Biological Effectiveness (RBE) or Quality Factors (QF) for radiation. [13]

PARTICLE TYPE	CONDITIONS		RBE or QF
X-Rays, Gamma Rays	All		1
Electrons, Positrons	All		1
Alpha Particles or Heavy Charged Particles	LET Values in MeV/cm in Water:	< 35	2
		35	2
		70	4
		230	10
		530	20
		1750	40
		> 1750	40
Neutrons	Energy in MeV:	Thermal	5
		0.0001	4
		0.01	5
		0.1	15
		0.5	22
		1.0	22
		10	13

The absorbed dose can be calculated from the expression

$$\frac{1}{\rho}\frac{d\Phi}{dx} = -\left(\frac{\mu}{\rho}\right)\Phi \tag{7.49}$$

where Φ is the energy fluence (J/cm^2). The quantity $d\Phi/dx$ is the energy deposited per unit volume. Division by the density ρ converts this to the energy deposited per unit mass. The energy fluence can be estimated from the particle flux density by multiplying by the average particle energy and the exposure time. A more accurate value of absorbed dose can be obtained by integrating the product of the particle energy times the mass absorption coefficient (as a function of energy) times the flux density as a function of energy times the exposure time. That is,

$$Absorbed\ Dose = -\int dE \left(\frac{\mu}{\rho}\right) ET\phi(E) \tag{7.50}$$

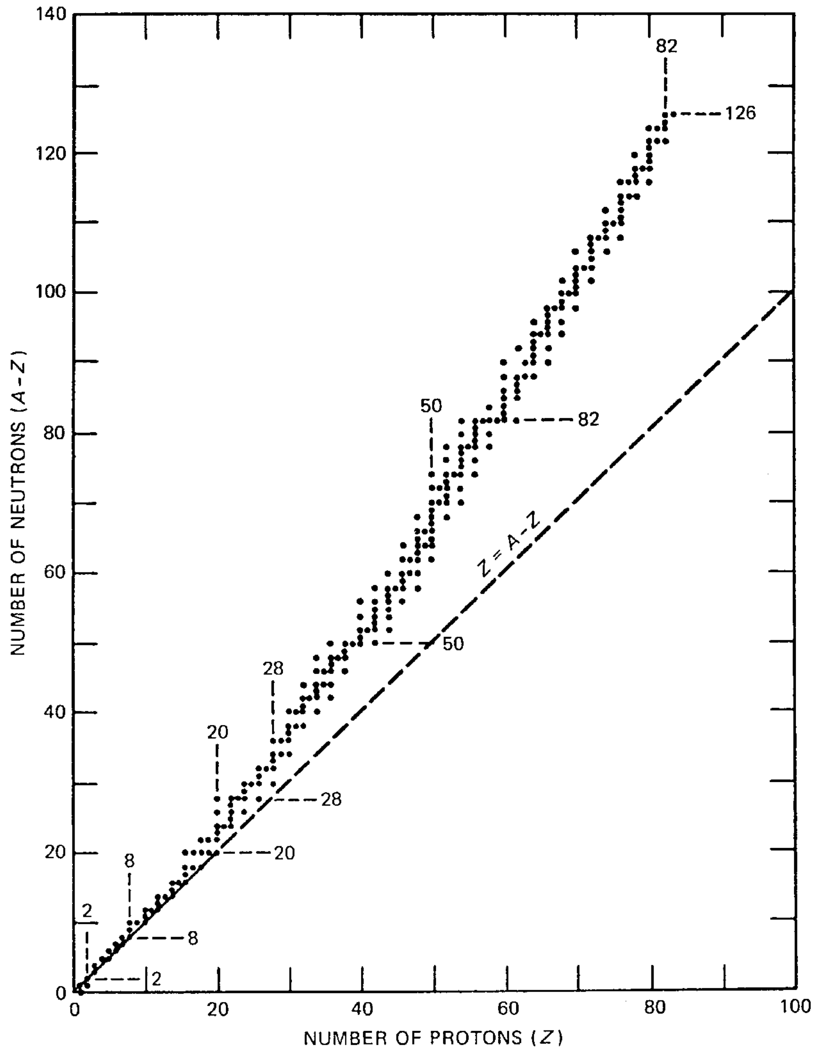
where E is the particle energy, T is the exposure time, and $\phi(E)$ is the particle flux per unit area per unit time per unit energy interval.

Radioactive Decay

Many of the atomic nuclei we have observed on earth are stable. That is, they have existed in their current form without change since they were formed in a supernova explosion long before the formation of our solar system. However, many nuclei are unstable. They have excess energy relative to some different but achievable configuration of protons and neutrons. Unstable nuclei transform themselves into more stable nuclei through any one of a number of processes collectively referred to as radioactive decay.

Stable nuclei have characteristics that are moderately well defined. As illustrated in Figure 7-10, they tend to have a number of neutrons that is slightly larger than the number of protons. The

Figure 7-10. The relationship between neutron number and proton number in stable nuclei. [1]



trend illustrated in Figure 7-10 continues well beyond the last stable nucleus shown. If a curve were drawn through the stable nuclei and extrapolated to Z=120 or so, then the radioactive nuclei lying nearer this extrapolated curve would have longer half-lives than those lying considerable distances from the curve. For example, Thorium-232 (Z=90, N=142) lies exactly where the extrapolated curve would intercept the upper end of Figure 7-5. It has a half-life of 1.4 x 10^{10} years, long enough that even though Th-232 is radioactive, much of it has still not decayed in the period since its creation (long before the birth of the solar system). Curium-247 (Z=96, N=151) has a half-life of 1.56 x 10^7 years and lies near the extrapolated line. If we were to plot the binding energy per nucleon as a function of Z and N, we would find that the maximum binding energies would superimpose almost perfectly on this stability chart. Since binding energies are actually negative energies, the curve is sometimes referred to as the "valley of stability". Several other empirical observations are worth mentioning. Nuclei with even numbers of protons and even numbers of neutrons tend to be more stable than nuclei with odd numbers of neutrons and protons. This appears to result from pairing of spins. For example, a spin up proton can pair up with a spin down proton to produce a lower energy state. Furthermore, there are certain values of N and Z that seem to have extra stability. These values are usually called "magic numbers" and are given by

N or Z = 2, 8, 20, 28, 50, 82, 126, and 184.

At these values of N and Z theoretical models (so-called "shell models" of nuclear structure predict the formation of closed shells with complete pairing between all nuclear particles and all quantum numbers.

Alpha decay is the spontaneous emission of an alpha particle. The daughter nucleus has 2 fewer neutrons, 2 fewer protons, and an atomic weight 4 less than the parent nucleus. The decay reaction may be represented as

$$ {}_Z^A T^N \xrightarrow{\alpha \text{ Decay}} {}_{Z-2}^{A-4} T-2^{\;N-2} + {}_2^4\mathrm{He}^2\left(\alpha\right) \qquad (7.51)$$

Alpha decay is rare in nuclides with $Z \leq 82$ (lead). It is common near the "valley of stability" for elements with $Z > 82$. Given the shape of the binding energy curve, one can mathematically ascertain that alpha decay is not energetically possible for ground state nuclei with A less than some number. Empirically, this value appears to be roughly $A = 200$ which corresponds roughly to $Z = 82$.

Beta-minus (β^-) decay results from the conversion of a neutron into a proton with the subsequent production and emission of an electron and an electron antineutrino. The decay reaction is

$$ {}_Z^A T^N \xrightarrow{\beta^- \text{ Decay}} {}_{Z+1}^{A} T+1^{\;N-1} + \mathrm{e}^- + \bar{\nu}_e \qquad (7.52)$$

Beta-minus decay is common in neutron-rich nuclides (those lying on the neutron-abundant side of the valley of stability).

Beta-plus (β^+) decay results from the conversion of a proton into a neutron with the subsequent production and emission of a positron and an electron neutrino. The decay reaction is

$$ {}^{A}_{Z}T^{N} \xrightarrow{\beta^{+}\text{ Decay}} {}^{\;\;A}_{Z-1}\,T-1^{\;N+1} + \mathrm{e}^{+} + \nu_{e} \qquad (7.53)$$

Beta-plus decay is common in neutron-deficient nuclides (those lying on the neutron-poor side of the valley of stability).

Electron capture is the conversion of a proton into a neutron by absorbing and inner shell electron and emitting a neutrino. The decay reaction is

$$ {}^{A}_{Z}T^{N} + \mathrm{e}^{-}\left(\text{K Shell}\right) \xrightarrow{\text{Electron Capture}} {}^{\;\;A}_{Z-1}\,T-1^{\;N+1} + \nu_{e} \qquad (7.54)$$

Electron capture is common in neutron-deficient nuclides. It frequently competes with beta-plus decay for decays in the same nuclide.

Spontaneous fission is the splitting (in the absence of any initiating particle) of a nucleus into two smaller nuclei plus the emission of a small number (typically 2 or 3) free neutrons. The decay reaction is

$$ {}^{A}_{Z}T^{N} \xrightarrow{\text{Spontaneous Fission}} {}^{A1}_{Z1}T1^{\,N1} + {}^{A-A1-X}_{\;\;Z-Z1}T2^{\;N-N1-X} + X\;{}^{1}_{0}\mathrm{n}^{1} \qquad (7.55)$$

With the exception of Beryllium-8 which fissions instantaneously into two alpha particles, spontaneous fission is not known to occur below thorium *Z=90*. It becomes more and more likely as Z increases. Although it competes with alpha decay, spontaneous fission usually is the much weaker process.

Deuteron, triton, and Helium-3 emission are rare modes of decay in which the parent nucleus emits a deuterium, tritium, or Helium-3 nucleus instead of an alpha particle. The decay reactions are

$$ {}^{A}_{Z}T^{N} \xrightarrow{\text{Deuteron Emission}} {}^{A-2}_{Z-1}\,T-1^{\;N-1} + {}^{2}_{1}\mathrm{H}^{1}\left(\mathrm{d}\right) \qquad (7.56)$$

$$ {}^{A}_{Z}T^{N} \xrightarrow{\text{Triton Emission}} {}^{A-3}_{Z-1}\,T-1^{\;N-2} + {}^{3}_{1}\mathrm{H}^{2}\left(\mathrm{t}\right) \qquad (7.57)$$

and

$$ {}^{A}_{Z}T^{N} \xrightarrow{\text{Helium-3 Emission}} {}^{A-3}_{Z-2}\,T-2^{\;N-1} + {}^{3}_{2}\mathrm{He}^{1} \qquad (7.58)$$

Isomeric transition is the decay of a metastable (very-long-lived) excited nuclear state to the ground state by emission of a gamma ray. The decay reaction is

$$ {}^{A}_{Z}T^{N}\Big|^{*} \xrightarrow{\text{Isomeric Transition}} {}^{A}_{Z}T^{N}\Big|^{\text{gnd}} + \gamma \qquad (7.59)$$

Isomeric transitions differ from other decay schemes in that its decay product is the same nuclide as the parent. The product atom is usually electrically neutral.

Internal conversion is the decay of an excited nuclear state by ejection of an inner shell electron rather than emission of a gamma ray. The decay reaction is

$$ {}^{A}_{Z}T^{N}\Big|^{*} + \mathrm{e}^{-}\left(\mathrm{K\ Shell}\right) \xrightarrow{\text{Internal Conversion}} {}^{A}_{Z}T^{N}\Big|^{+} + \mathrm{e}^{-}\left(\mathrm{free}\right) \qquad (7.60)$$

Internal conversion differs from isomeric transition in that the product atom is an electrically positive ion. Neither changes the basic nuclide.

Details on the characteristics of isotope can be found in several sources. Summary data on decay modes and half-lives can be found in the Table of the Nuclides.[8],[14] This document exists in wall chart, pamphlet [8], and electronic form [14]. Details on nuclear energy levels, decay modes including secondary particle emissions, and half-lives, can be found in the Table of Isotopes.[15]

Activity is the radioactive disintegration rate. The SI unit of activity is the **Becquerel (Bq)** which is equal to one (1) disintegration per second. An older unit of activity that is still in common use is the **Curie (Ci)**. One Curie is equal to 3.7 x 10^{10} disintegrations per second.

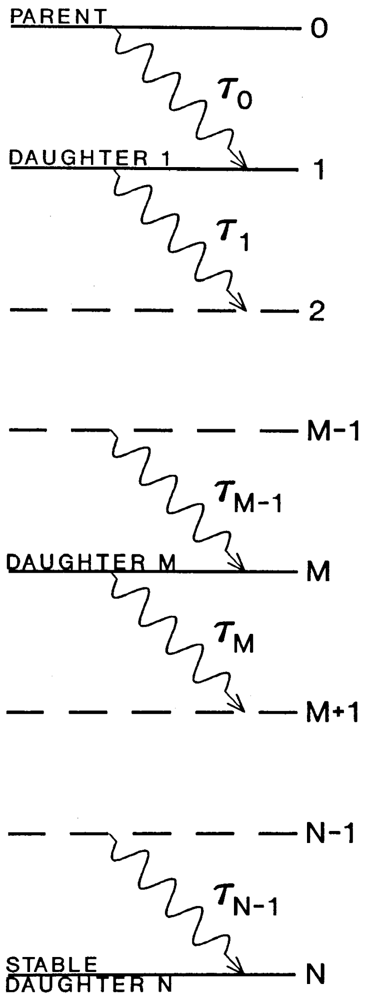
$$1 \text{ Ci} = 3.7 \text{ x } 10^{10} \text{ Bq}$$

Unstable (radioactive) species decay to other unstable species until ultimately one of the decay products is stable. This is called a **radioactive decay chain**.

Let us examine the hypothetical decay chain shown in Figure 7-11. Nuclide 0 is the species formed initially in some production system, or is naturally occurring. It is referred to as the parent nuclide of the decay chain. The parent decays with a characteristic time constant τ_0 to nuclide 1. Nuclide 1 is called a daughter as are the products of all subsequent decay processes. Once produced by decay of the parent, nuclide 1 can randomly decay to nuclide 2 with a characteristic time constant τ_1. The decay chain continues with nuclide M decaying to nuclide M+1 with characteristic time constant τ_M.

The quantity of each nuclide present as a function of time requires solution of the rate equations for the system. Rate equations are sets of equations that balance the rate at which a species is created or produced against the rate at which that species decays. Adjacent to each level in Figure 7-6 we have given the appropriate rate equation. We assume that we start with a known quantity of nuclide 0, N_0 that was either collected from a production source or purified from ore, or other source. Since there is no further production, the rate equation contains only a decay term. The rate of decay is assumed to be proportional to the number of N_0 present and inversely proportional

Figure 7-11. A typical radioactive decay chain and the rate equations associated with each nuclide.



$$\frac{dN_0}{dt} = -\frac{N_0}{\tau_0} \tag{7.61}$$

$$\frac{dN_1}{dt} = -\frac{N_1}{\tau_1} + \frac{N_0}{\tau_0} \tag{7.62}$$

$$\frac{dN_M}{dt} = -\frac{N_M}{\tau_M} + \frac{N_{M-1}}{\tau_{M-1}} \tag{7.63}$$

$$\frac{dN_N}{dt} = +\frac{N_{N-1}}{\tau_{N-1}} \tag{7.64}$$

to the decay constant τ_0. The rate equation for the first daughter and subsequent radioactive daughters has a production term (assumed due solely to decay from nuclide 0) and a decay term. The final stable daughter has a production term but lacks a decay term. In general, it is difficult to solve coupled differential equations in closed form. Regardless, we will attempt solutions using Laplace transforms.

Consider first the decay of the parent nuclide. Consulting Appendix E on Laplace transforms we find that Eq. (7.61) can be transformed to yield

$$s\mathcal{N}_0(s) - \mathcal{N}_0(0) = -\mathcal{N}_0(s)/\tau_0 \tag{7.65}$$

Solving this equation for $\mathcal{N}_0(s)$ yields

$$\mathcal{N}_0(s) = \mathcal{N}_0(0) / \left(s + \frac{1}{\tau_0} \right) \tag{7.66}$$

Using the inverse transforms gives the final result

$$N_0(t) = N_0(0) e^{-t/\tau_0} \tag{7.67}$$

The parent decays with simple exponential form and characteristic decay time τ_0. The time constant τ is related to the half-life $t_{1/2}$, the time it takes for half of the nuclide to decay, by the relation

$$t_{1/2} = (\ln 2)\, \tau \cong 0.69315\, \tau\ . \tag{7.68}$$

Using the half-life instead of τ or vice versa is a common mistake. Forewarned is forearmed.

The rate equation for the first daughter nuclide can also be Laplace transformed to yield

$$s\mathcal{N}_1(s) - \mathcal{N}_1(0) = -\frac{\mathcal{N}_1(s)}{\tau_1} + \frac{\mathcal{N}_0(s)}{\tau_0} \tag{7.69}$$

Simple combining of terms yields

$$\left(s + \frac{1}{\tau_1} \right) \mathcal{N}_1(s) = \mathcal{N}_1(0) + \mathcal{N}_0(s) / \tau_0 \tag{7.70}$$

which is easily solved for $\mathcal{N}_1(s)$

$$\mathcal{N}_1(s) = \frac{\mathcal{N}_1(0)}{\left(s + \frac{1}{\tau_1} \right)} + \frac{\mathcal{N}_0(s)}{\tau_0 \left(s + \frac{1}{\tau_1} \right)} \tag{7.71}$$

We have already obtained a solution for $\mathcal{N}_0(s)$, i.e., Eq. (7.66), which is substituted into Eq. (7.71).

$$\mathcal{N}_1(s) = \frac{\mathcal{N}_1(0)}{\left(s + \frac{1}{\tau_1} \right)} + \frac{\mathcal{N}_0(0)}{\tau_0 \left(s + \frac{1}{\tau_0} \right)\left(s + \frac{1}{\tau_1} \right)}\ . \tag{7.72}$$

This equation can be solved by applying the appropriate inverse Laplace transforms to yield

$$N_1(t) = N_1(0)e^{-t/\tau_1} + \left(\frac{\tau_1}{\tau_0 - \tau_1}\right) N_0(0)\left(e^{-t/\tau_0} - e^{-t/\tau_1}\right). \tag{7.73}$$

The rate equations for the second daughter and all succeeding daughters (except the last stable one) follow a similar procedure. The rate equation for the M$^{\text{th}}$ daughter is easily seen to be Laplace transformed to yield

$$\mathcal{N}_M(s) = \frac{\mathcal{N}_M(0)}{\left(s + {}^1\!/_{\tau_M}\right)} + \frac{\mathcal{N}_{M-1}(s)}{\tau_{M-1}\left(s + {}^1\!/_{\tau_M}\right)}. \tag{7.74}$$

Earlier results are repeatedly substituted into the expression

$$\begin{aligned} \mathcal{N}_M(s) &= \frac{\mathcal{N}_M(0)}{\left(s + {}^1\!/_{\tau_M}\right)} + \frac{1}{\tau_{M-1}\left(s + {}^1\!/_{\tau_M}\right)} \\ &\quad \times \left[\frac{\mathcal{N}_{M-1}(0)}{\left(s + {}^1\!/_{\tau_{M-1}}\right)} + \frac{\mathcal{N}_{M-2}(s)}{\tau_{M-2}\left(s + {}^1\!/_{\tau_{M-1}}\right)}\right] \end{aligned} \tag{7.75}$$

$$= \cdots \quad \text{until all } \mathcal{N}_{M-n}(s) \text{ are replaced by } \mathcal{N}_{M-n}(0)$$

The end result has the form

$$\begin{aligned} \mathcal{N}_M(s) &= \frac{\mathcal{N}_M(0)}{\left(s + {}^1\!/_{\tau_M}\right)} + \frac{\mathcal{N}_{M-1}(0)}{\tau_{M-1}\left(s + {}^1\!/_{\tau_{M-1}}\right)\left(s + {}^1\!/_{\tau_M}\right)} \\ &\quad + \frac{\mathcal{N}_{M-2}(0)}{\tau_{M-1}\tau_{M-2}\left(s + {}^1\!/_{\tau_{M-2}}\right)\left(s + {}^1\!/_{\tau_{M-1}}\right)\left(s + {}^1\!/_{\tau_M}\right)} \\ &\quad + \cdots \\ &\quad + \frac{\mathcal{N}_0(0)}{\tau_{M-1}\tau_{M-2}\cdots\tau_0\left(s + {}^1\!/_{\tau_0}\right)\cdots\left(s + {}^1\!/_{\tau_{M-1}}\right)\left(s + {}^1\!/_{\tau_M}\right)} \end{aligned} \tag{7.76}$$

In theory, inverse transformation of this equation will yield the time dependent solution for $N_M(t)$.

The specific result for the second daughter is given below

$$\begin{aligned}\mathcal{N}_2(s) = &\frac{\mathcal{N}_2(0)}{\left(s+{}^{1}\!/\!_{\tau_2}\right)} + \frac{\mathcal{N}_1(0)}{\tau_1\left(s+{}^{1}\!/\!_{\tau_1}\right)\left(s+{}^{1}\!/\!_{\tau_2}\right)} \\ &+ \frac{\mathcal{N}_0(0)}{\tau_1\tau_0\left(s+{}^{1}\!/\!_{\tau_0}\right)\left(s+{}^{1}\!/\!_{\tau_1}\right)\left(s+{}^{1}\!/\!_{\tau_2}\right)}\end{aligned} \tag{7.77}$$

with inverse form

$$\begin{aligned}N_2(t) = &N_2(0)e^{-t/\tau_2} + \left(\frac{\tau_2}{\tau_1-\tau_2}\right)N_1(0)\left(e^{-t/\tau_1}-e^{-t/\tau_2}\right) \\ &-\left[\frac{N_0(0)}{\tau_1\tau_0\left(\frac{1}{\tau_0}-\frac{1}{\tau_1}\right)\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\left(\frac{1}{\tau_2}-\frac{1}{\tau_0}\right)}\right] \\ &\times\left[\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)e^{-t/\tau_0} + \left(\frac{1}{\tau_2}-\frac{1}{\tau_0}\right)e^{-t/\tau_1} + \left(\frac{1}{\tau_0}-\frac{1}{\tau_1}\right)e^{-t/\tau_2}\right]\end{aligned} \tag{7.78}$$

It would be nice to be able to continue this process for subsequent daughters. However, a closed form inverse transform to Eq. (7.76) does not exist for $M \geq 3$.

The analysis can be carried a little further if we make an approximation. We will assume steady state behavior. In essence we will assume that so much parent exists and its half-life is so long that it will essentially never disappear. We will assume that all other half-lives are short compared to the parent half-life. In the steady state all time derivatives may be assumed to be equal to zero. There are two realistic initial conditions. In the first, the radioactive parent has been freshly processed. For example, it has been freshly mined, isolated chemically from other elements, and enriched to eliminate other nuclides. In this case, at t=0, only the parent is present. That is,

$$N_1 \approx N_2 \approx \cdots \approx N_N \approx 0 \tag{7.79}$$

The total activity is given by

$$\text{Activity} \approx -\frac{dN_0}{dt} \approx \frac{N_0}{\tau_0} \tag{7.80}$$

Now let us consider a raw ore of the parent. Since chemical processing has not occurred except on geological time scales, all of the daughters are present. In the steady state assumption,

$$\frac{dN_M}{dt} \approx 0 \qquad \text{in Steady State} \tag{7.81}$$

Thus

$$\frac{dN_M}{dt} = -\frac{N_M}{\tau_M} + \frac{N_{M-1}}{\tau_{M-1}} \approx 0 \tag{7.82}$$

This is readily solved to find

$$\frac{N_M}{\tau_M} = \frac{N_{M-1}}{\tau_{M-1}} \qquad \text{for any } M \tag{7.83}$$

The total activity is the sum of the activities of all of the nuclides

$$\text{Activity} \approx \sum_{M=0}^{N-1} \frac{N_M}{\tau_M} \approx N\,\frac{N_0}{\tau_0} \tag{7.84}$$

Thus, per unit mass of parent nuclide, the activity of the raw ore is *N* times stronger than the activity of the purified parent nuclide.

References

[1] Glasstone, Samuel and Sesonske, Alexander, Nuclear Reactor Engineering 3rd Ed. (Van Nostrand Reinhold, New York NY, 1981).

[2] Turner, James E., Atoms, Radiation, and Radiation Protection (McGraw-Hill Book Co., New York NY, 1992) pp. 115-119, 132.

[3] Schiff, Leonard I., Quantum Mechanics 3rd Ed. (McGraw-Hill Book Co., New York NY, 1968) pp. 420-422.

[4] Heitler, W., The Quantum Theory of Radiation 3rd Ed. (Oxford University Press, London UK, 1960) pp. 256-264.

[5] Mott, N. F. and Massey, H. S. W., The Theory of Atomic Collisions 3rd Ed. (Oxford University Press, London UK, 1971) pp. 475-504.

[6] Bethe, Hans A. and Salpeter, Edwin E., Quantum Mechanics of One- and Two-Electron Atoms (Plenum Press, New York NY, 1977) pp. 255-262.

[7] Massey, H. S. W. and Burhop, E. H. S., Electronic and Ionic Impact Phenomena (Oxford University Press, London UK, 1956) pp.137-141.

[8] Walker, F. William, Parrington, Josef R., and Feiner, Frank, Nuclides and Isotopes14th Ed. (General Electric Company Nuclear Energy Operations, San Jose CA, 1989).

[9] Van Lint, V. A. J., Flanagan, T. M., Leadon, R. E., Naber, J. A., and Rogers, V. C., Mechanisms of Radiation Effects in Electronic Materials, Vol. 1 (John Wiley & Sons, New York NY, 1980) p. 46.

[10] PHOTCOEF©, “Photon Interaction Coefficients of the Elements”, AIC Software, Inc. (23 October 1998). Available on the Internet at http://www.photcoef.com/2121.html.

[11] Hubbell, J. H., and Seltzer, S. M., “Tables of X-Ray Mass Attenuation Coefficients and Mass Energy-Absorption Coefficients”, National Institute of Standards and Technology, NISTIR 5632 (April 1997). Available on the Internet at http://physics.nist.gov/PhysRefData/XrayMassCoef/cover.html.

[12] Chang, Jonghwa, “ENDFPLOT-2.0", Korean Atomic Energy Research Institute (undated). Available on the Internet at http://atom.kaeri.re.kr/endfplot.shtml.

[13] Shapiro, Jacob, Radiation Protection: A Guide for Scientists and Physicians 3rd Ed. (Harvard University Press, Cambridge MA, 1990) p. 47.

[14] Chang, Jonghwa, “Table of the Nuclides”, Korean Atomic Energy Research Institute (undated). Available on the Internet at http://www.dne.bnl.gov/CoN/index.html.

[15] Lederer, C. Michael and Shirley, Virginia S. (Eds.), Table of Isotopes 7th Ed. (John Wiley & Sons, New York NY, 1978).

Problems

7-1. One will encounter six to eight fundamental particles in discussing ordinary nuclear reactions. List these particles in order of their likelihood of being encountered.

7-2. The mass energy of a Uranium-238 nucleus is 238.050785 amu. What is the binding energy (in MeV) per nucleon of Uranium-238? Compare this result with that for Uranium-235.

7-3. Estimate the energy released in a single fission reaction of the form

$${}_0n^1 + {}_{92}U^{235} \text{–>} 3\ {}_0n^1 + {}_{44}Ru^{110} + {}_{48}Cd^{123}$$

This energy should not depend strongly on exactly which fission products are produced.

7-4. The purpose of the moderator in a nuclear reactor is to slow down the fast neutrons produced by fission. Elastic scattering is the process used for slowing the neutrons. Which is likely to make the best moderator: polyethylene, aluminum, or lead? That is, which will produce the largest energy loss per elastic scattering event?

7-5. Which reaction is most likely to dominate the attenuation of gamma rays in the 10 keV energy region? In the 1 MeV energy region? In the 100 MeV energy region?

7-6. A hypothetical material has a nuclear reaction cross section of 10^{-24} cm^2 and a density of 10^{22} atoms/cm^3. How thick does a foil of this material have to be to cause at least 10% of the particles striking it to undergo the nuclear reaction?

7-7. The mean free path is the average distance a particle can travel before it undergoes a reaction. The mean free path l is given by

$$l = 1/N\sigma$$

where N is the atomic density and σ is the cross section for the particle interaction with the medium. Calculate the mean free path of 1 keV neutrons in aluminum.

7-8. Calculate the mean free path for neutron-induced fission reactions in Uranium-235. The energies of fission neutrons

7-9. A Cobalt-60 radiography source has an activity of 1000 Curies. Each disintegration produces a 1.5 MeV gamma ray. How thick must a lead shield be to reduce the human dose rate to 1 mrem/hr if the source is 1 m from the radiographer's control station?

7-10. Using the Table of Isotopes or the Chart of the Nuclides or similar reference, describe the radioactive decay chain for Uranium-238. That is, list the decay modes, daughters, and half-lives of all decays until a stable isotope daughter is reached. Does this decay chain have any weak side chains?

7-11. A nucleus has ten more neutrons than an isotope of the same element lying near the valley of stability. What radioactive decay mechanism is this isotope likely to employ?

7-12. An element heavier than lead lies near the valley of stability. What radioactive decay mechanism is this isotope likely to employ?

7-13. A sample of uranium ore is determined to contain one mole of Uranium-238. If no chemical or nuclear processing of this ore has occurred in the last billion years, what weight of Radium-226 (a daughter of U-238 decay) should this ore sample contain?

CHAPTER 8

RADIOMETRY, PHOTOMETRY, & RADIOMETRIC ANALYSIS

Radiometry

Radiometry is the science of electromagnetic radiation measurement. Our interest in radiometry lies in two areas. First, anyone involved with sensor design, selection, or integration needs to be familiar with the terminology used in radiometry. These terms permeate sensor design discussions. Second, radiometric analysis is a powerful tool that often lets one determine the characteristic range equation of a sensor.

The nomenclature of radiometry is governed by international standards [1]-[3]. The fundamental quantities, their symbols, essential definition, and units are summarized in Table 8-1. The fundamental quantity is **energy** with SI units of Joules. The **energy density** is the energy contained in a unit volume. This is equivalent to the derivative (change) of total energy with respect to volume. It has units of J/m^3. Power or **flux** is the time rate of change of the total energy. Flux has units of Watts. Flux is usually measured at a surface. **Flux density** is the flux passing through a unit area. This is equivalent to the derivative of the flux with respect to surface area. It has units of W/m^2. Flux density either comes out of a surface (it is emitted by the surface) in which case it is called **radiant exitance** or it is incident on a surface in which case it is called **irradiance**. Note that radiant exitance and irradiance are given different symbols to facilitate discrimination between the direction of energy flow at the surface. If the power flows through a surface at an angle, the

Table 8-1. Standardized radiometric quantities and units.

QUANTITY	SYMBOL	DEFINING EQUATION	SI UNIT NAMES	UNIT SYMBOLS
ENERGY	Q_e	---	JOULE	J
ENERGY DENSITY	w_e	dQ_e/dV	---	J/m^3
FLUX (POWER)	Φ_e	dQ_e/dt	WATT	W [J/s]
FLUX DENSITY (AT A SURFACE) - RADIANT EXITANCE (EMITTED) - IRRADIANCE (INCIDENT)	 M_e E_e	$d\Phi_e/dA\cos\theta$	---	W/m^2
RADIANT INTENSITY	I_e	$d\Phi_e/d\Omega$	---	W/sr
RADIANCE	L_e	$d^2\Phi_e/d\Omega\,dA\cos\theta$	---	W/m^2-sr

actual surface area must be corrected by the cosine of the angle θ with respect to the normal to the surface to yield the projected surface area. This accounts for the $\cos\theta$ term in the denominator of the defining equation. **Radiant intensity** is the flux per unit solid angle (i.e., the derivative of total flux with respect to solid angle). It has units of W/sr. Solid angle is a product of azimuthal and elevation angular extent. It relates the directionality in three-dimensional space of flux away from a point or flux towards a point. The magnitude of a solid angle is the area of the surface encompassing the flux projected on a unit sphere (the unit sphere has a radius = 1). The entire unit sphere has an area of $4\pi(1)^2 = 4\pi$ and is therefore defined to have 4π steradians (sr) of solid angle. **Radiance** is a quantity which describes the derivative of total energy with respect to time, surface area, and solid angle. It has units of W/m^2-sr.

Table 8-1 uses energy as the basic unit for defining the radiometric quantities. There is an alternative formulation which uses the total number of electromagnetic quanta as the defining element. In this formulation, “Joules” would be universally replaced by “Number of Photons”. Otherwise the definitions remain the same. The subscript “e” in Table 8-1 explicitly indicates that the quantities use the energy-based formulation. Quantities using the quantum-based formulation customarily have a subscript “q” rather than “e”. If no subscript is used, then the energy-based formulation should be assumed.

A subscript “λ” is used to denote a spectral quantity, i.e., a quantity with explicit wavelength dependence

$$X_{e\lambda} = \frac{\text{Quantity } X_e(\lambda)}{\text{Unit wavelength interval centered at } \lambda} \tag{8.1}$$

The spectral and spectrum-integrated quantities are interrelated by the expressions

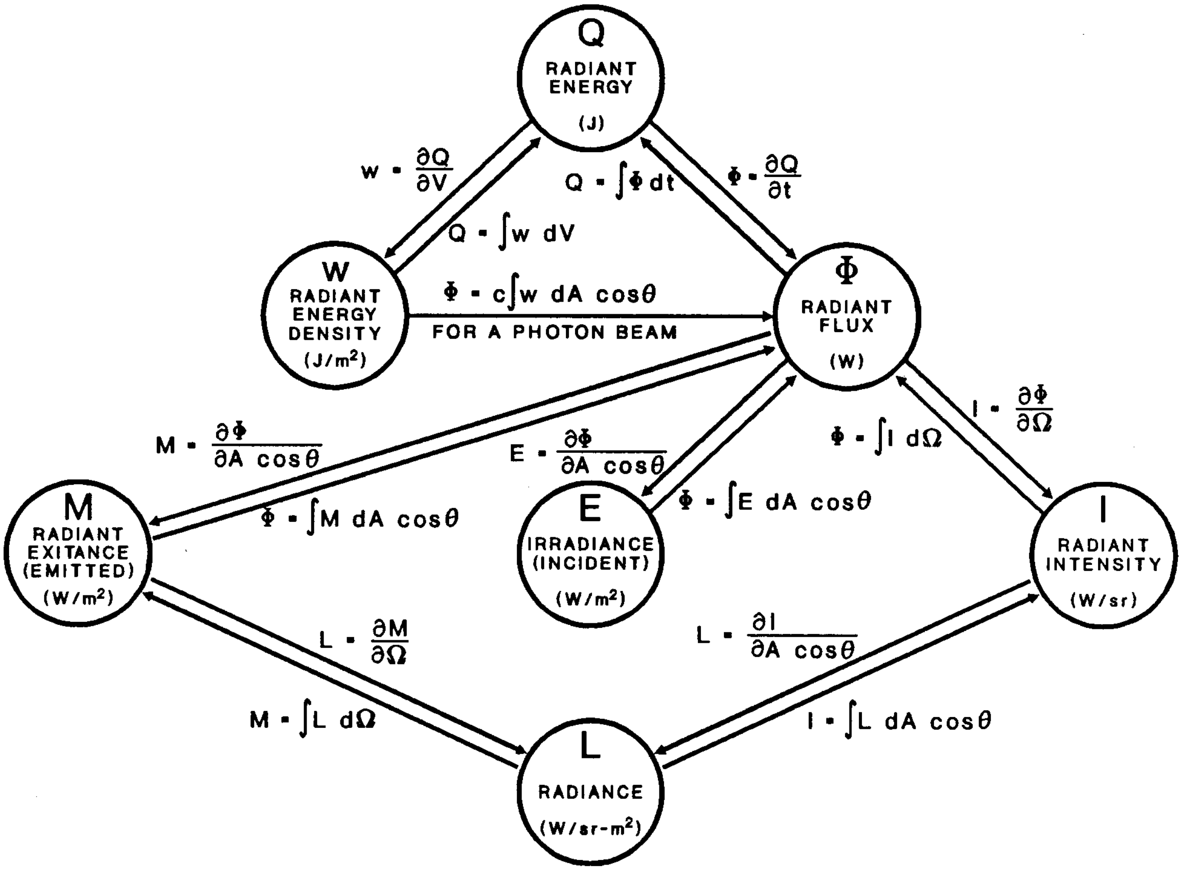
$$X_{e\lambda} = \frac{dX_e}{d\lambda} \tag{8.2}$$

and

$$X_e = \int_0^\infty d\lambda \; X_{e\lambda} \tag{8.3}$$

In radiometric analyses, it is often necessary to convert from one radiometric quantity to another radiometric quantity. For example, we may have a measurement of radiant exitance but desire to know the flux. Figure 8-1 shows the interrelations between each of the radiometric quantities. To convert from one to another follow any path from the initial quantity to the final quantity performing the indicated operations as you go. For example, to convert from radiance to flux, one could integrate first over area to get radiant intensity and then integrate over solid angle to get flux. Alternatively, one could integrate over solid angle to get radiant exitance and then integrate over area to get flux.

Figure 8-1. Relationships between radiometric quantities.



So far the discussion has addressed only quantities associated with the radiation. Obviously, in any radiometric analysis, material properties must be considered. There are four material properties relevant to radiometry: absorptance, reflectance, transmittance, and emissivity. The absorptance α of a material is the ratio of radiant energy absorbed by the material to the radiant energy incident on the material

$$\alpha = \frac{Q_{ABSORBED}}{Q_{INCIDENT}} . \tag{8.4}$$

The reflectance ρ of a material is the ratio of radiant energy reflected by the material to the radiant energy incident on the material

$$\rho = \frac{Q_{REFLECTED}}{Q_{INCIDENT}} . \tag{8.5}$$

The transmittance τ of a material is the ratio of radiant energy transmitted through a material to the radiant energy incident on the material

$$\tau = \frac{Q_{TRANSMITTED}}{Q_{INCIDENT}} . \qquad (8.6)$$

The fourth quantity, emissivity is slightly different in general nature. The emissivity ε of a material is the ratio of the energy radiated (emitted) from a material to the energy that would have been emitted had the material been a blackbody

$$\varepsilon = \frac{Q_{EMITTED}}{Q_{BLACKBODY}} = \alpha . \qquad (8.7)$$

In a later section of this chapter we will discuss blackbody radiation in detail. At this point it should suffice to say that blackbodies have unique properties that make them suitable as absolute (primary) standards for radiation measurements. By definition, blackbodies have $\varepsilon = 1$.

Emissivity is sometimes called emittance in analogy with absorptance, reflectance, and transmittance, although emissivity remains the much more common term. In the same light, some individuals will use the terms absorptivity, reflectivity, and transmissivity, instead of absorptance, reflectance, and transmittance. Widespread usage of these three alternative forms has eliminated any serious connotations of error in their use, nevertheless, **emissivity, absorptance, reflectance, and transmittance are the officially accepted terms**.

In Eq. (8.7) we have also alluded to the fact that emissivity must equal absorptance. This fact is required in order to establish thermal equilibrium in material systems in contact with a radiation field. Since reflection and transmission do not add or subtract energy from a material, they cannot change its temperature. If absorption exceeded emission, then a material isolated from everything but radiation would continue to absorb energy at a faster rate than it could emit it. The temperature would steadily rise towards infinity. If emission exceeded absorption, then the object would radiate energy away faster than it could absorb it. The temperature would steadily fall towards absolute zero. Since neither extreme is physically realistic, it follows that absorption must at some point be equal to emission.

The relation

$$\alpha + \rho + \tau = \varepsilon + \rho + \tau = 1 \qquad (8.8)$$

is just as easy to justify. Since energy must be conserved, if it isn't reflected, and it isn't absorbed, and it's not allowed to transmit unaffected through the material, then it must not have existed in the first place since there are no other alternatives.

Another powerful concept in radiometry is the **Lambertian** source or Lambertian surface. A Lambertian surface has the characteristic that the radiant intensity I_θ in a direction θ relative to the normal to the surface has a smaller value by a factor of $cos\,\theta$ than the radiant intensity normal to the surface I_n. That is,

$$I_\theta = I_n \cos\theta \qquad (8.9)$$

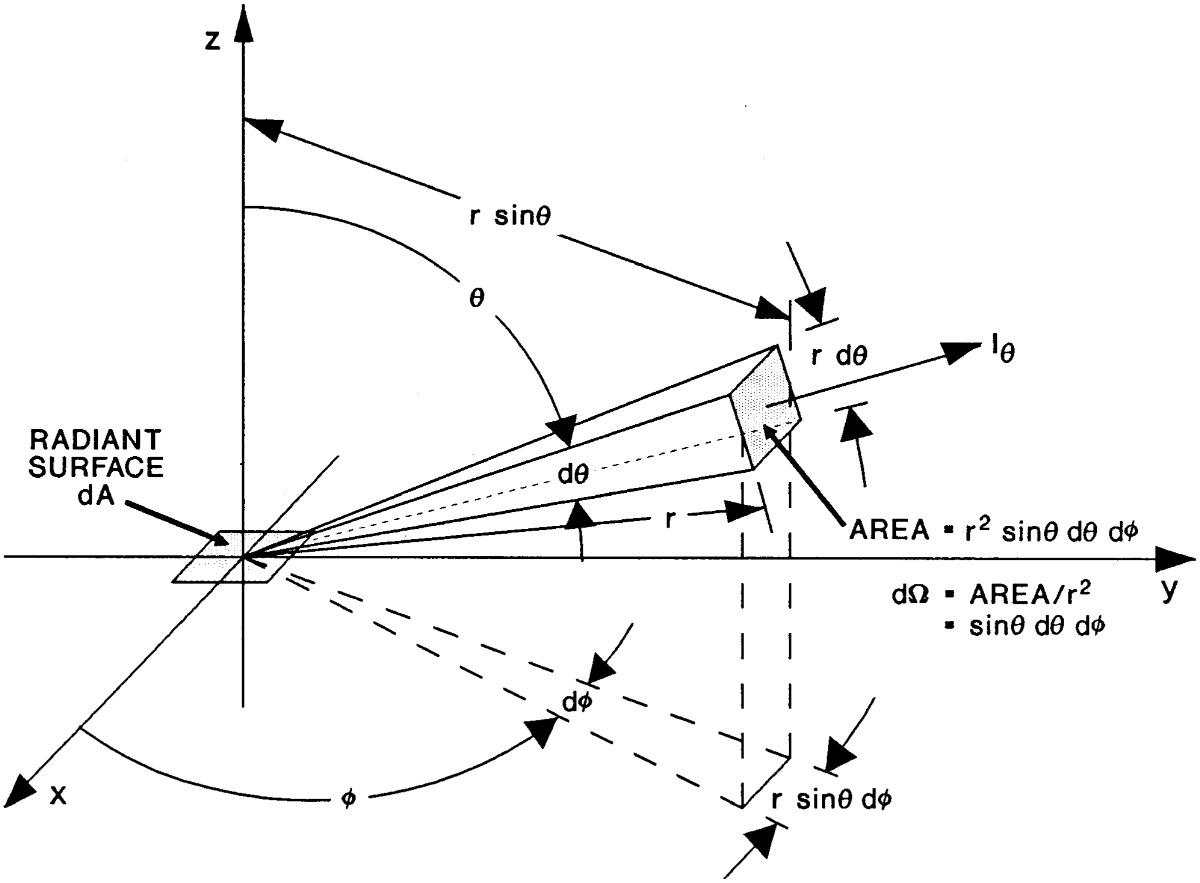
Eq. (8.9) is known as Lambert's Law for diffuse surfaces. The $cos\theta$ dependence holds whether the radiant intensity arises from emission or reflection. If due to reflection it is unaffected by the direction of illumination. The total radiant flux into a hemisphere around a Lambertian surface is found by integration to be

$$\Phi = \int d\Omega \, I_\theta = \int_0^{\pi/2} d\theta \cos\theta \sin\theta \int_0^{2\pi} d\phi \, I_n = \pi I_n \tag{8.10}$$

That is, the flux is effective radiated into π steradians (or half of the solid angle of a hemisphere. Essentially, Lambertian surfaces have no preferred directivity for either emission or reflection.

The geometry associated with the integration above is shown in Figure 8-2. The relation between the emitting element of area and the measurement surface element are clearly shown.

Figure 8-2. Geometry of emission from a Lambertian surface.



Another interesting property of Lambertian surfaces is that the radiance is independent of angle. Using our earlier definitions we find

$$L_\theta = \frac{\partial^2 Q}{\partial\Omega\,\partial A\cos\theta} = \frac{\partial I}{\partial A\cos\theta} = \frac{I_\theta}{dA\cos\theta} = \frac{I_n\cos\theta}{dA\cos\theta} = \frac{I_n}{dA} = L \quad (8.11)$$

Finally, for Lambertian surfaces, simple relationships exist between radiance, radiant intensity, radiant flux, and radiant exitance. That is,

$$L = \frac{I_n}{dA} = \frac{\Phi}{\pi\,dA} = \frac{M}{\pi}\,. \quad (8.12)$$

Because of this assumption of a Lambertian surface will always simplify radiometric calculations.

It is instructive to question what makes a surface Lambertian and are any real surfaces Lambertian? It is evident from the reflection characteristics that mirror-like surfaces are not Lambertian. Lambertian surfaces diffuse any light that they reflect. Matte surfaces are more likely to be Lambertian. Physically, if the surface height is random and rough on the scale of a wavelength of the radiation of interest, and has a transverse correlation length that is less than the wavelength of the radiation, the surface will be Lambertian. Translated into common language, the peak-to-valley variation in height must be larger than a wavelength, height measurements taken at separated points must exhibit random variations, and you cannot move more than one wavelength along the surface before the height changes from a relative peak to a relative valley. It is difficult to produce a surface that truly possesses this character, however, most surfaces approximate this character if the wavelength is small enough. Although, we will not prove it, blackbody radiators are Lambertian. They exhibit the fundamental $cos\,\theta$ behavior.

Photometry

Photometry is simply radiometry complicated by the fact the ultimate detector is the human eye. Thus, photometric quantities must take into account the spectral response of the eye. As we shall see, this is a significant complication. Nevertheless, the combat systems engineer must be familiar with both radiometric measurements as well as photometric measurements.

For every radiometric quantity, there is a corresponding photometric quantity. These are summarized in Table 8-2. It is a standard convention to denote the photometric quantities by using the radiometric symbols with the subscript "*v*" instead of the subscript "*e*" or "*q*" or no subscript at all. The fundamental photometric quantity is the **luminous energy**. It is related to the radiant energy by the relation

$$Q_v = \int_{0.380\mu\text{m}}^{0.760\mu\text{m}} d\lambda\, K(\lambda)\, Q_{e\lambda} \tag{8.13}$$

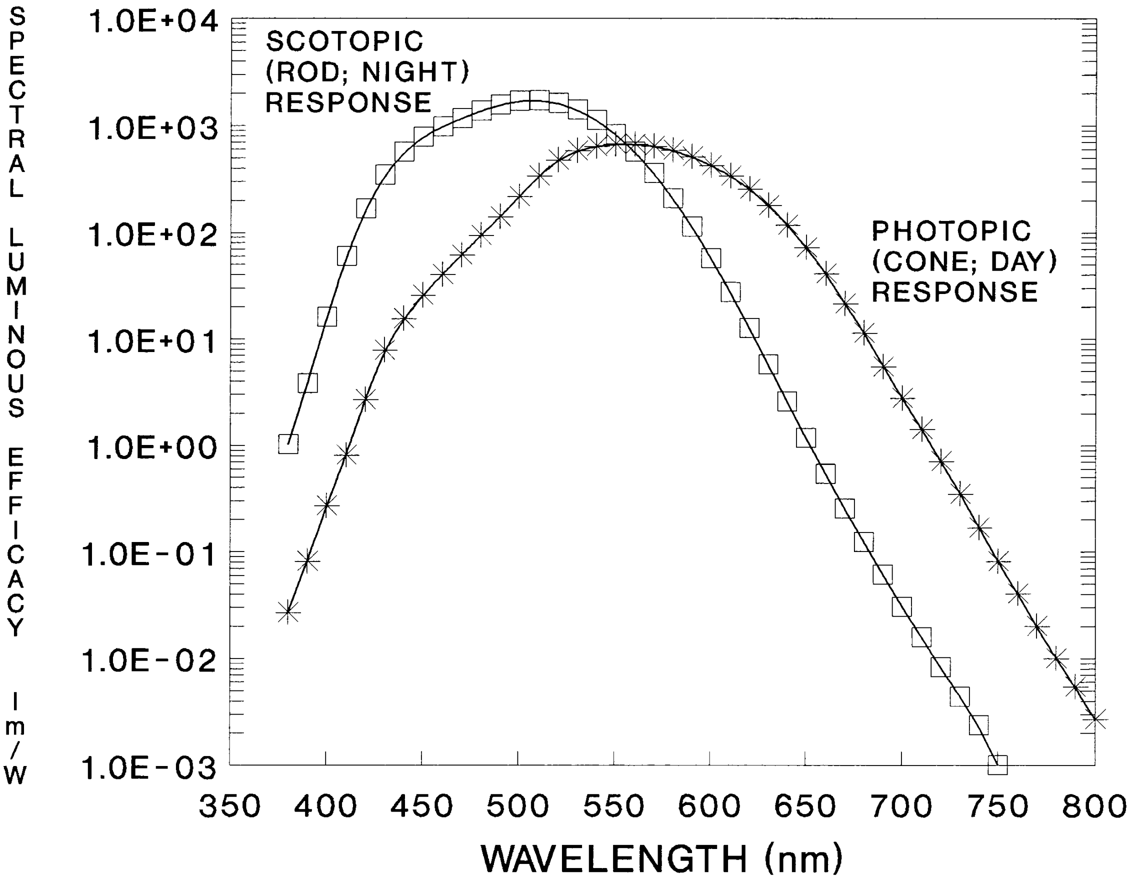
where *K(λ)* is the spectral luminous efficacy is a factor that relates radiant energy (in J) at a specific

Table 8-2. Standardized photometric quantities and units.

QUANTITY	SYMBOL	DEFINING EQUATION	UNIT NAMES	UNIT SYMBOLS
LUMINOUS ENERGY (QUANTITY OF LIGHT)	Q_v	$\int_{0.380\,\mu m}^{0.760\,\mu m} K(\lambda)\, Q_{e\lambda}\, d\lambda$	TALBOT	--- [lm s]
LUMINOUS ENERGY DENSITY	w_v	dQ_v/dV	---	lm s/m^3
LUMINOUS FLUX	Φ_v	dQ_v/dt	LUMEN (SI)	lm
LUMINOUS FLUX DENSITY - LUMINOUS EXITANCE (EMITTED) - ILLUMINANCE (INCIDENT)	 M_v E_v	$d\Phi_v/dA$	LUX (SI) PHOT FOOT CANDLE	lx [lm/m^2] ph [lm/cm^2] ft c [lm/ft^2]
LUMINANCE (PHOTOMETRIC BRIGHTNESS)	L_v	$d^2\Phi_v/d\Omega\, dA \cos\theta$ VALID FOR LAMBERTIAN SOURCES ONLY {	NIT STILB FOOT LAMBERT LAMBERT APOSTILB	nt [lm/m^2 sr] [cd/m^2] sb [cd/cm^2] ft L [(1/π) cd/ft^2] L [(1/π) cd/cm^2] asb [(1/π) cd/m^2]
SPECTRAL LUMINOUS EFFICACY (PHOTOPIC RESPONSE FUNCTION)	K(λ)	$\Phi_{v\lambda}/\Phi_{e\lambda}$	---	lm/W
SPECTRAL LUMINOUS EFFICIENCY	V(λ)	$K(\lambda)/K_{max}$	---	---
RETINAL ILLUMINANCE	E_t	$L_v\, A_{pupil}$	TROLAND	td [mm^2 cd/m^2]

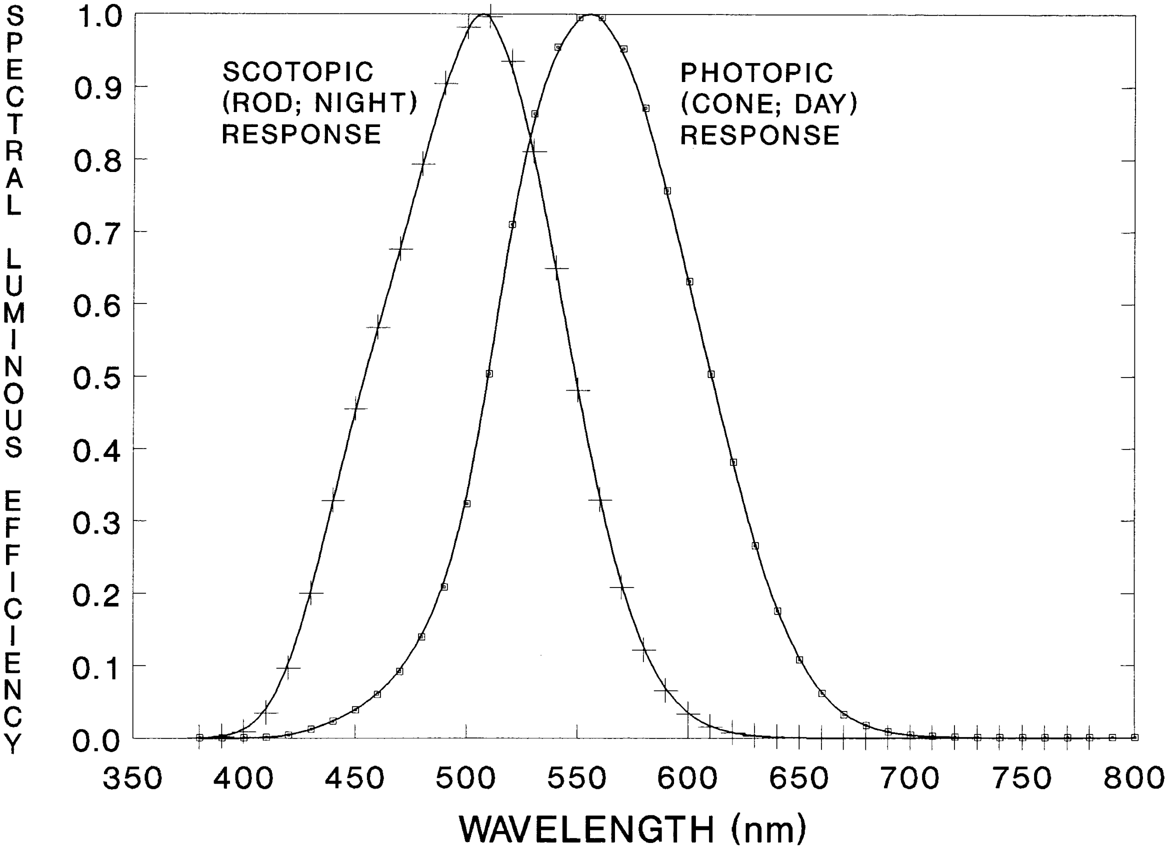
wavelength to the luminous energy (in Talbots or lumen-seconds). In essence the luminous efficacy is visual response of the eye. The conversion from radiant energy to luminous energy only needs to consider energy at wavelengths in the visible spectrum. Hence, the limits of integration. The luminous efficacy is plotted in Figure 8-3. There are two curves: one for scotopic vision (the black & white nighttime vision produced by the rod cells in the human retina) and one for photopic vision (the color daytime vision produced by the cone cells).

Figure 8-3. Spectral luminous efficacy curves for the two form of human visual response.



Several of the photometric quantities have units that are valid SI units. Most do not. Almost all of the quantities have alternate units that are not SI. In general these alternate units are obtained by substituting feet or centimeters for the SI meters. Some of the alternate units of luminance have an additional factor of 1/π that accounts for Lambertian emission into π steradians. Two additional quantities are listed in Table 8-2 that have no radiometric counterparts. The spectral luminous efficiency (as opposed to efficacy) is the relative response of the eye. The photopic and scotopic spectral luminous efficiency curves are presented in Figure 8-4. The final photometric quantity is

Figure 8-4. Spectral luminous efficiency of human vision.



retinal illuminance. This quantity is the luminance multiplied by the area of the pupil (of the eye). It is a measure of how much luminous intensity actually illuminates the retina. It is commonly encountered in studies of external degradation of the human visual system such as flashblindness. All of the photometric quantities with radiometric counterparts are used in vision system analyses exactly like their radiometric counterparts are used in other analyses. Photometric quantities can be transformed between themselves exactly like the corresponding radiometric quantities using the relations in Figure 8-1.

Radiometric Analysis – Range Equations

Radiometric analysis is the science of accounting for the flows of energy involved in the operation of any sensor system using radiation (electromagnetic, acoustic, or nuclear) as a propagator. In some instances, radiant energy may flow from the sensor itself or some other source, interact with the target, and flow back to the sensor. In others the target may produce the energy that ultimately flows to the sensor. When coupled with an expression for the intrinsic noise of the sensor, the radiometric analysis result produces the carrier-to-noise-ratio equation for the sensor. This is the single most important equation involved in design.

Many factors affect the flow of radiation. Electronic components may produce gain or loss of energy (power). Losses may come reflection or absorption of energy. Either mechanism has an adverse effect on performance, but reflection may produce unwanted side effects. Propagation through the medium intervening between sensor and target can cause attenuation through scattering or absorption. Imperfect collimation and/or diffraction may cause the intensity to decrease through spreading. Turbulence or multipath effects will cause some of the energy to go in unwanted directions. Interaction of external radiation with a target will not reflect isotropically. Antennas used for collected signal radiation will have angle-dependent collection efficiency. Consideration of all of these effects and factors is the province of radiometric analysis.

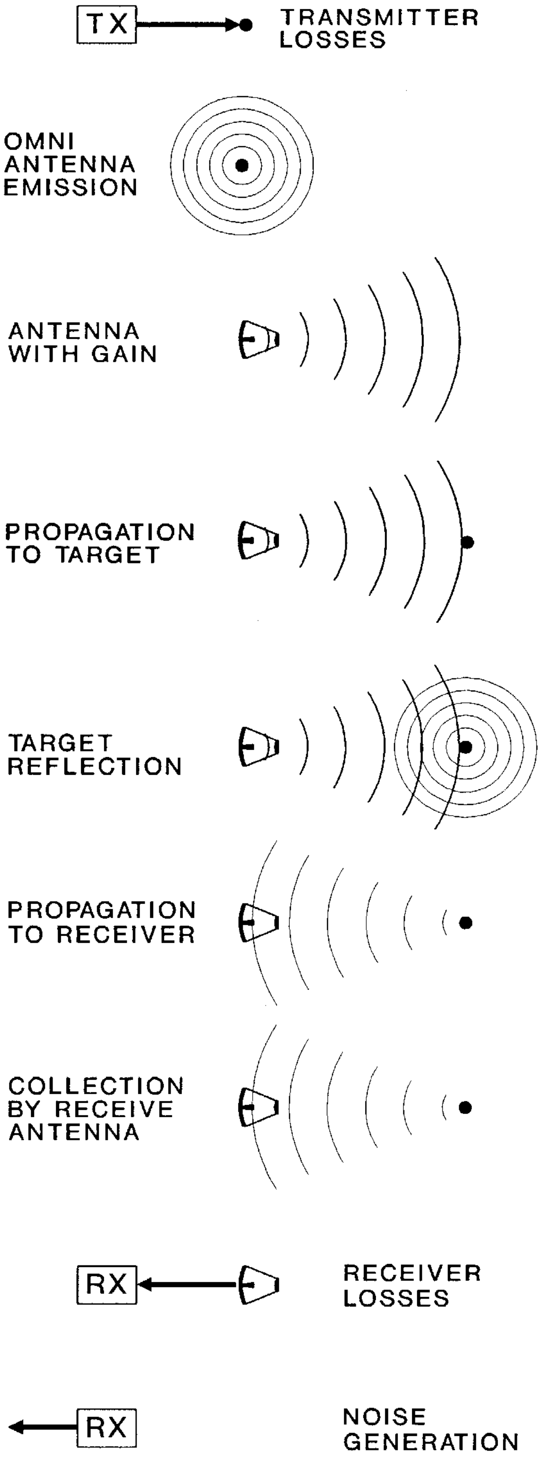
The basic approach is to begin at the point most distant from the sensor output in terms of propagation. Note: in an active sensor, this most distant point may in fact be spatially close to the output point. The flow of radiation is then traced element by element accounting for all gains and losses, reflections, redirections, and spreading. The power remaining at the output point is the average signal power. Dividing the expression representing this result by the expression for the sensor noise power gives the CNR equation.

Figure 8-5 illustrates the radiometric analysis that yields the CNR equation for a conventional microwave radar. This result is usually called the radar range equation.[4] The analysis begins at the transmitter unit with output power P_T. The waveguide and other transmit components introduce a loss factor L_T (defined in this instance as the ratio of output power to input power). Thus the antenna is seen to radiate an effective power P. If this power were to be radiated by an omnidirectional antenna (one that radiates its power into 4π sr), the power density (intensity) at distance R is the effective power divided by the factor $4\pi R^2$. The real antenna is assumed to have an antenna gain defined by

$$G_T = 4\pi A_T / \lambda^2 \tag{8.14}$$

The antenna gain is the ratio of the solid angle of a sphere divided by the effective solid angle into which the antenna radiates. Thus, the effective power density of the radar at range R is increased by the antenna gain. As the power density propagates to the target, which is assumed to be at range R, it suffers atmospheric extinction of magnitude $e^{-\alpha R}$. The target presents an effective area, called the radar cross section σ, to the incident radiation. The total reflected power is the product of the

Figure 8-5. Radiometric derivation of the radar equation.



Power at Antenna

$$P = P_T L_T$$

Power Density from Omnidirectional Antenna

$$I = P_T L_T \frac{1}{4\pi R^2}$$

Power Density from High-Gain Antenna

$$I = P_T L_T \frac{1}{4\pi R^2} G_T$$

Power Density at Target

$$I = P_T L_T \frac{1}{4\pi R^2} G_T e^{-\alpha R}$$

Power Reflected from Target

$$P = P_T L_T \frac{1}{4\pi R^2} G_T e^{-\alpha R} \sigma$$

Power Density at Receiver

$$I = P_T L_T \frac{1}{4\pi R^2} G_T e^{-\alpha R} \sigma \frac{e^{-\alpha R}}{4\pi R^2}$$

Power Collected by Antenna

$$P = P_T L_T \frac{1}{4\pi R^2} G_T e^{-\alpha R} \sigma \frac{e^{-\alpha R}}{4\pi R^2} A_R$$

Power at Receiver Detector

$$P = P_T L_T \frac{1}{4\pi R^2} G_T e^{-\alpha R} \sigma \frac{e^{-\alpha R}}{4\pi R^2} A_R L_R$$

Carrier-to-Noise Ratio out of Detector

$$CNR = \frac{P}{kTBF} = \frac{P_T L_T L_R G_T A_R \sigma}{kTBF} \frac{e^{-2\alpha R}}{\left(4\pi R^2\right)^2}$$

incident power density and this effective area. The power is assumed to be reflected into 4π sr. Any directivity in the reflection is already accounted for in the definition of the radar cross section. As the reflected radiation spreads out from the target and propagates back to the receiver, it suffers another $1/4\pi R^2$ spreading loss and another $e^{-\alpha R}$ atmospheric attenuation loss. The resulting power density is then incident on the receive antenna. The product of the power density and the area of the receive antenna is the total received power. This power is reduced by any losses in the receiver electronics L_R. At this point we now have an expression for the overall signal power. Dividing by the known receiver noise *kTBF* (where *T* is the electronics temperature, *B* is the noise bandwidth of the electronics, and *F* is an excess noise term that must be included to describe real amplifiers), yields the radar range equation

$$CNR = \frac{P}{kTBF} = \frac{P_T L_T L_R G_T A_R \sigma}{kTBF} \frac{e^{-2\alpha R}}{\left(4\pi R^2\right)^2} \quad . \tag{8.15}$$

Radiometric analysis of any other sensor proceeds in the same straightforward manner as our radar example.

In summary, radiometric analysis is one of the powerful and universal tools available to the combat systems engineer. It lets him derive a CNR-based range equation for any radiation-based sensor.

Blackbody Radiation

It is not difficult to make a blackbody. The engineer merely needs to make something that will absorb all (or at least the vast majority of) radiation incident upon it. A large hollow ball with a small hole is a good example. Any radiation passing through the hole into the ball will strike the interior of the ball and reflect many times before it can escape out the small hole. Since no material is a perfect reflector, even a few percent absorption multiplied by dozens of reflections lead to almost complete absorption. Indeed, thirty years ago, teachers used to make excellent blackbodies by stacking a hundred (or so) double-edged razor blades side by side. The very shallow taper of the blades from body to sharp edge meant the any radiation striking the face with all the sharp edges would undergo many reflections off the shiny metal (from one blade to the adjacent blade and back again. Despite the high reflectivity of the metal, the edged side of the stack appeared to be a deep, velvety black. Heating the blade stack would provide a nice blackbody radiator for infrared experiments. Unfortunately, it is almost impossible to find double-edged razor blades today. Technology has made them obsolete for their original intended purpose. Scientists are interested in blackbodies because $\alpha = \varepsilon$ implies that a perfect absorber (a "black" body) is also a perfect emitter. The radiant exitance from a blackbody is a function only of the temperature and the wavelength of the radiation. The material characteristics are completely eliminated. The spectrum and "brightness" of one blackbody at temperature T is exactly the same as the spectrum and brightness of any other blackbody at temperature T.

For these reasons, not only are blackbodies used as emission standards in radiometric measurements, they have also been the subject of considerable study in and of themselves. Indeed, studies of blackbody radiation paved the way for quantum theory. Max Planck was forced to propose that radiation existed in quantized form (Einstein later called them photons) in order to derive an expression for the emission of a blackbody that matched experimental measurements. That expression, described by various people as Planck's Radiation Law, the Planck distribution, or the Planck function, takes slightly different forms depending on the exact radiometric quantity described. The author has found the expression for the spectral radiant exitance to be the most useful. In energy-based form it is given by [5]

$$M_\lambda = \frac{2\pi c^2 h}{\lambda^5}\left[\frac{1}{e^{hc/\lambda kT}-1}\right] = \frac{3.741832\times 10^8}{\lambda^5}\left[\frac{1}{e^{14387.86/\lambda T}-1}\right] \quad (8.16)$$

where M_λ has units of W/m^2-μm when the temperature T is expressed in Kelvins and the wavelength λ is expressed in μm. The quantum-based expression for the spectral radiant exitance is

$$M_{q\lambda} = \frac{2\pi c}{\lambda^4}\left[\frac{1}{e^{hc/\lambda kT}-1}\right] = \frac{1.883652\times 10^{27}}{\lambda^4}\left[\frac{1}{e^{14387.86/\lambda T}-1}\right] \quad (8.17)$$

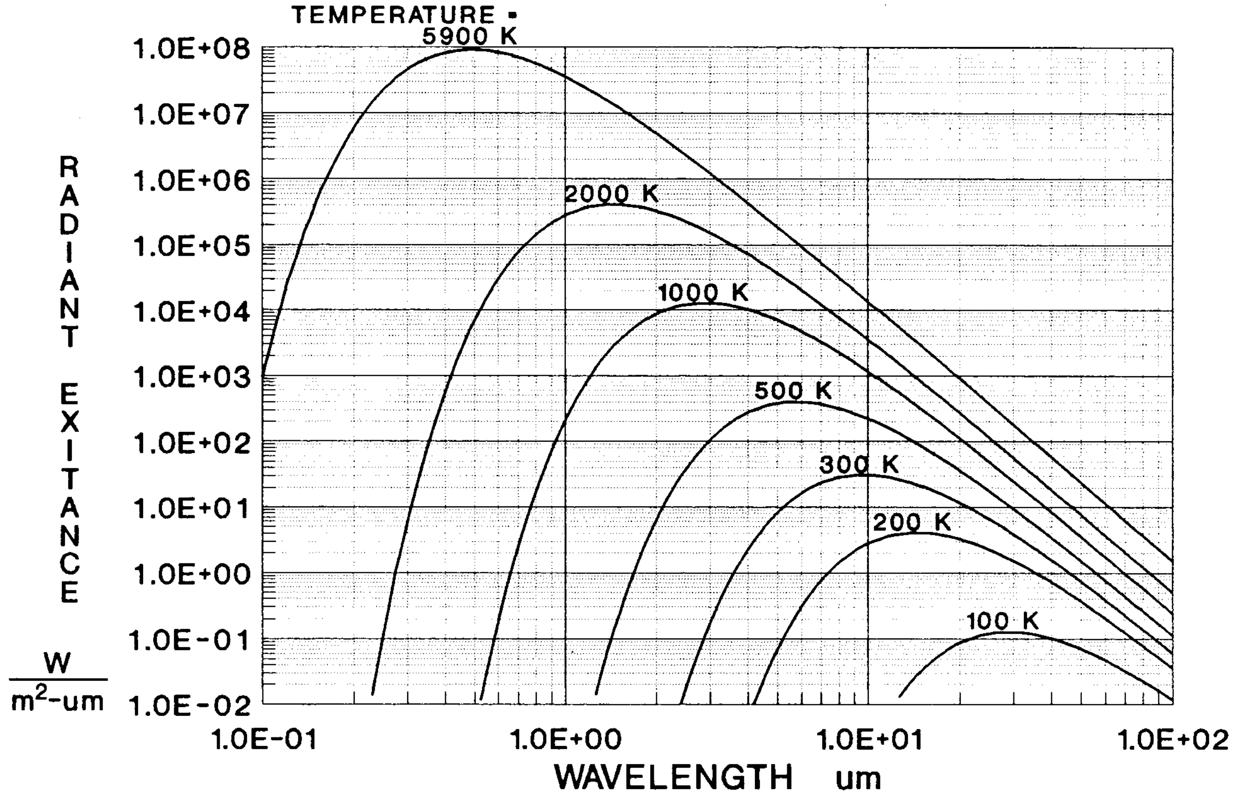
where $M_{q\lambda}$ has units of photons/sec-m^2-μm. The primary difference between the two forms is the factor hc/λ, the photon energy.

The behavior of the Planck distribution is shown in Figure 8-6. At any specific material temperature, the radiant exitance rises rapidly with increasing wavelength to a peak and then falls off much more slowly as wavelength continues to increase. All points on a curve for one value of temperature T_0 always lie below the corresponding points on any curve with temperature greater than T_0. All points on that first curve always lie above the corresponding points on any curve with temperature lower than T_0. The wavelengths (λ_{MAX}) corresponding to the peak value of radiant exitance are related to the temperature T of the surface that produced them by Wien's Displacement Law

$$\lambda_{MAX} T = 2897.756 \ \mu\text{m - K} \qquad (8.18)$$

Thus, a surface with a temperature of 1000 K has its peak value of radiant exitance at a wavelength of 2.897756 μm. A Planck distribution with its peak value at 1.0 μm is produced by a temperature of 2897.756 K.

Figure 8-6. The Planck distribution of radiant exitance for blackbodies.



There are times when knowledge of the total radiant exitance M is desired. This can be obtained by integrating Eq. (8.16). The result is commonly referred to as the Stefan-Boltzmann Law,

$$M = \int_0^\infty d\lambda \ M_\lambda = \frac{2\pi^5 k^4}{15h^3c^2} T^4 = \sigma T^4 = 5.67032 \times 10^{-8} \ T^4 \tag{8.19}$$

where M has units of W/m^2, if T and λ are expressed as above. The constant σ is called the Stefan-Boltzmann constant. In quantum-based form the Stefan-Boltzmann Law takes the form

$$M_q = \int_0^\infty d\lambda \ M_{q\lambda} = 2.40406 \frac{2\pi k^3}{h^3c^2} T^3 = 1.52040 \times 10^{15} \ T^3 \tag{8.20}$$

where M_q has units of photons/sec-m^2.

In some infrared sensor analyses it is desirable to determine how much the radiant exitance of a surface will change if the temperature of that surface is varied. This can be determined by taking the temperature derivative of Eq. (8.16). The evaluated form of this derivative is given as

$$\begin{aligned} \frac{\partial M_\lambda}{\partial T} &= \frac{2\pi c^3 h^2}{kT^2 \lambda^6} \frac{e^{hc/\lambda kT}}{\left[e^{hc/\lambda kT} - 1\right]^2} \\ &= \frac{5.383695 \times 10^{12}}{T^2 \lambda^6} \frac{e^{14387.86/\lambda T}}{\left[e^{14387.86/\lambda T} - 1\right]^2} \end{aligned} \tag{8.21}$$

where $\partial M_\lambda / \partial T$ has units of W/m^2-µm-K. The quantum based version takes the form

$$\begin{aligned} \frac{\partial M_{q\lambda}}{\partial T} &= \frac{2\pi c^2 h}{kT^2 \lambda^5} \frac{e^{hc/\lambda kT}}{\left[e^{hc/\lambda kT} - 1\right]^2} \\ &= \frac{2.710172 \times 10^{31}}{T^2 \lambda^5} \frac{e^{14387.86/\lambda T}}{\left[e^{14387.86/\lambda T} - 1\right]^2} \end{aligned} \tag{8.22}$$

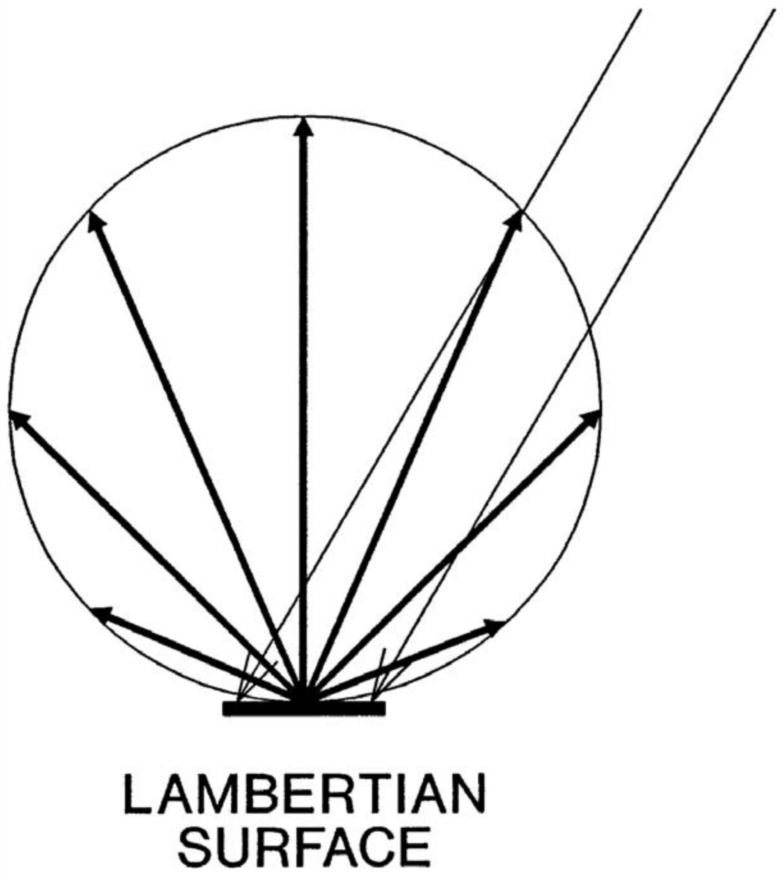
where $\partial M_{q\lambda} / \partial T$ has units of photons/m^2-µm-K. The product of the temperature derivative and the temperature change gives an estimate of the change in spectral radiant exitance.

Reflectance

Along with atmospheric attenuation and diffraction, reflection from the target is among the critical factors that must be considered in radiometric analyses. As defined earlier, the reflectance ρ of a material is the ratio of the total radiant energy reflected by the material to the total radiant energy incident on the material. This importance is only partially due to the potential for loss of energy that is posed by non-unity reflectance. It is also due to the angular dependence of the reflected radiation.

There are two ideal reflectors. The first is the Lambertian surface discussed in an earlier section. The Lambertian surface reflects energy with the same angular dependence regardless of the angle of incidence of the radiation. As shown in Figure 8-7, the reflected energy always obeys a cosine of the reflection angle (relative to the surface normal). Lambertian reflectors are often called **diffuse** reflecting surfaces. The descriptive terms matte finish or flat (as in flat black) may also be used. When illuminated with coherent radiation, Lambertian surfaces exhibit the phenomenon of laser speckle (the optical analog of Rayleigh cross section fluctuations in radar – both phenomena will be discussed in later volumes of this work.

Figure 8-7. Directionality of reflectance from a Lambertian surface.

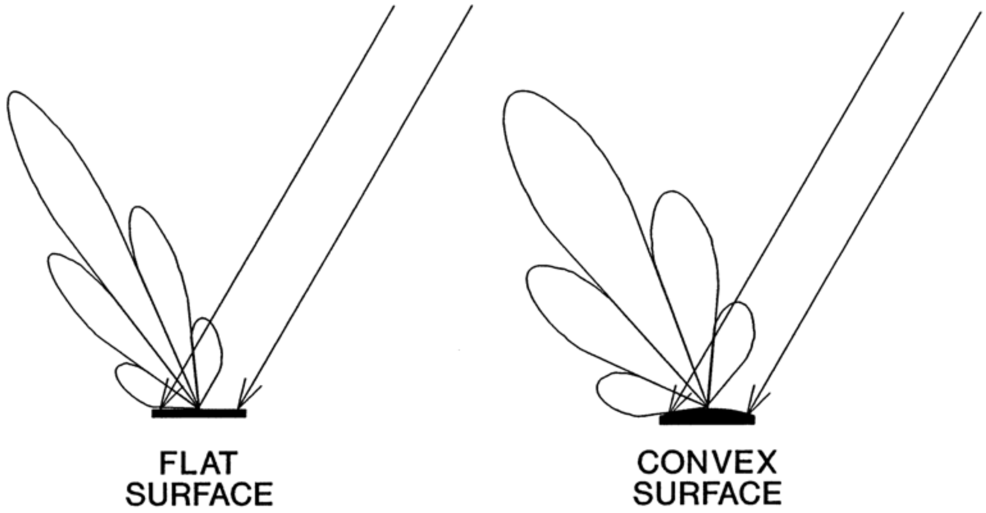


The second ideal reflector is the perfectly smooth surface. In the perfectly smooth surface, the reflecting surface has no scratches, dents, bumps, dust motes, chemical deposits, composition fluctuations, or any other roughness on any scale that approaches within several orders of magnitude of the wavelength of electromagnetic radiation being reflected. In practical terms this means that surface height variations cannot exceed more than a few atomic diameters or noticeable deviation from ideal behavior will occur. The surface may be curved but it cannot have the slightest roughness. The angular dependence of the radiation from the perfectly smooth surface is determined solely by diffraction. The reflecting surface will act like an aperture, whose size and shape are those of the original surface projected onto a plane perpendicular to the initial propagation direction. The resulting diffraction pattern will then be directed into the angle of the ordinary reflection. If the reflecting surface is curved the diffraction pattern will be further altered by the geometric reflection characteristics of the surface.

The perfectly smooth surface exhibits mirrorlike reflections. Such surfaces (and any reflections from them) are called **specular** surfaces or specular reflections. When such surfaces are oriented at just the right angle with respect to a distant bright source, observers will see a bright reflection. Since few situations have stationary reflectors and truly motionless observer, the bright reflection appears as a short flash of light often called a **glint**. Consequently, some people call the surfaces glint reflectors. Another descriptive term sometimes used is the term gloss finish.

Few real surfaces have characteristics that match either ideal reflector. They usually possess some mixture of both. The angular characteristics are broad and smooth without obvious diffraction characteristics, but there is a strong peak at normal incidence. That is, the surface exhibits some reflection at all angles, but there is a noticeable glint at the mirror reflection angle. Dull surfaces have a larger percentage of diffuse (Lambertian) character; shiny surfaces have a larger percentage of specular (smooth surface) character.

Figure 8-8. Directionality of reflectance from a smooth surface.



The bidirectional reflectance distribution function (BRDF) $\rho(\theta_i, \phi_i; \theta_r, \phi_r)$ describes the complex angular dependence of reflections from materials with less than ideal behavior. The BRDF is defined as

$$\rho(\theta_i, \phi_i; \theta_r, \phi_r) \tag{8.23}$$
$$= \frac{\text{Power reflected in a unit solid angle about direction } (\theta_r, \phi_r)}{\text{Power incident from direction } (\theta_i, \phi_i)}$$

where the subscript *i* denotes incidence and the subscript *r* denotes reflection. The BRDF has units of "per steradian". As a result, BRDF is not limited to a maximum value of 1. A specular reflector may have a BRDF in the specular reflection direction of 10^6 per steradian or higher. Bidirectional reflectance distribution functions are not easy to measure, requiring very sophisticated equipment. However, a limited number of measurements on important materials can be found in the literature. [6]

Most materials have no preferred directions (other than the surface normal). For these isotropic materials there is no specific dependence on ϕ_i. There will be a dependence on ϕ_r except in the case of the Lambertian reflector. A few materials do have preferred directions. For example, fabrics have threads that run parallel or perpendicular to each other. Machined surfaces have parallel machining marks. Diffraction gratings have parallel rows of microscopic reflecting surfaces. All of these materials will have explicit ϕ dependences that must be considered. This possibility is usually only considered in analysis if there is evidence that such anisotropic materials are present.

If the radiation is reflected through 180 degrees directly back along the direction of incidence, we obtain the monostatic reflectance distribution function (MRDF)

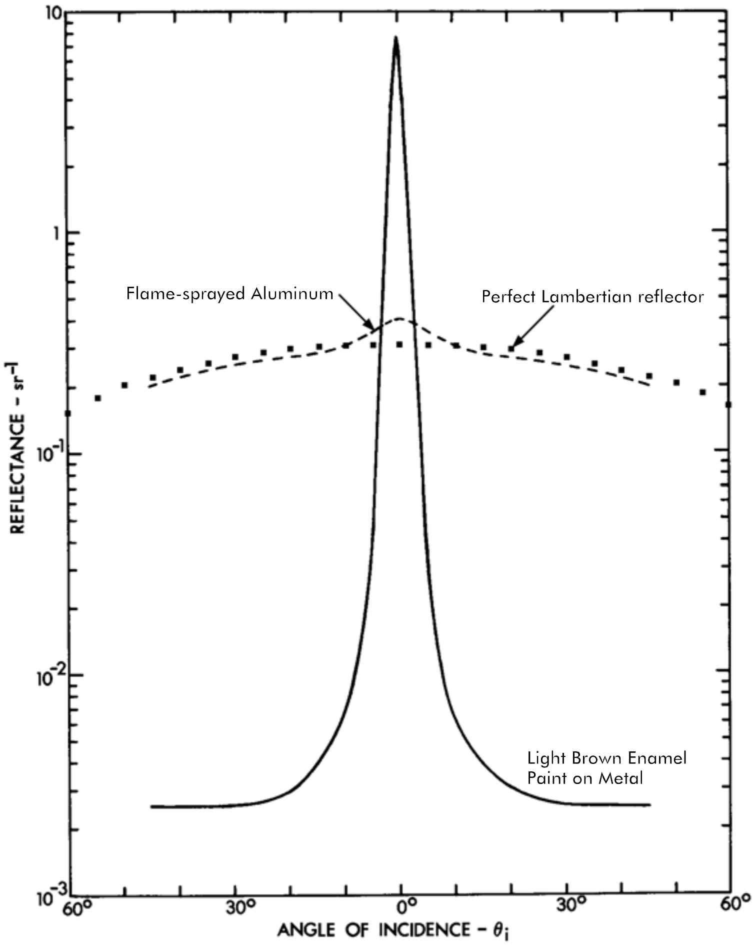
$$\rho_M(\theta_i, \phi_i) = \rho(\theta_i, \phi_i; \theta_i, \phi_i) \tag{8.24}$$

This is the reflectance distribution that directly affects the performance of laser sensors. Figure 8-9 shows the monostatic reflectance distribution function for two real materials: flame-sprayed aluminum (made from aluminum powder which is melted by a flame or plasma and blown onto an aluminum surface where it solidifies retaining much of its powder-like character while still being cemented onto the base plate) and light brown gloss enamel paint on smooth metal. These materials are isotropic so we need only plot the θ-dependence. Isotropic materials obey the relation

$$\rho_M(\theta_i, \phi_i) = \rho_M(\theta_i) . \tag{8.25}$$

Also shown is the MRDF of a Lambertian surface. As is evident, the flame-sprayed aluminum is almost Lambertian, but a very small amount of specular character remains. The painted surface exhibits a high degree of specular character.

Figure 8-9. Monostatic directional reflectance distribution function.



The total reflectance ρ (the reflectance associated with the transmittance and absorptance) can be determined by integrating either the BRDF or MRDF over all possible reflection angles, i.e.,

$$\rho = \int_0^{2\pi} d\phi \int_0^{\pi/2} d\theta \sin\theta \; \rho(\theta,\phi) . \tag{8.26}$$

We will discuss only the MRDF from now on, however, any of the results can be generalized to use the BRDF by substitution. If we assume that the Lambertian reflector has an MRDF of the form

$$\rho_{DIFFUSE}(\theta,\phi) = \frac{\rho_0 \cos\theta}{\pi} \quad \text{sr}^{-1} \tag{8.27}$$

where ρ_0 is the total reflectance. Performing the integration over both angles yields

$$\rho = \int_0^{2\pi} d\phi \int_0^{\pi/2} d\theta \sin\theta \, \frac{\rho_0 \cos\theta}{\pi} = \pi \frac{\rho_0}{\pi} = \rho_0 \tag{8.28}$$

which verifies that ρ_0 is the total reflectance.

If we consider a small flat reflecting surface, then it is obvious that reflection can only occur into a hemisphere (at most). Reflection cannot occur into the material. Because the total reflectance contains only components that have been reflected into a hemisphere, the total reflectance of opaque materials is sometimes called the **hemispherical reflectance**.

Most materials have MRDF functions with both diffuse and specular characteristics. In an approximation useful for some materials, the MRDF may be separated into a Lambertian component and a specular component

$$\rho(\theta,\phi) = \rho_{SPECULAR}(\theta,\phi) + \rho_{DIFFUSE}(\theta,\phi). \tag{8.29}$$

The specular component could, in principle, be approximated by any convenient function. The log-normal function provides a qualitatively accurate fit in many instances, but does not lend itself to easy evaluation of integrals. The author is unaware of a good functional approximation for $\rho_{SPECULAR}$. It is possible to state that the product of the peak value of the MRDF, $\rho(0,0)$, multiplied by the effective solid angle $\Omega_{1/2}$ (nominally measured at the half-intensity point) should be less than unity, i.e.,

$$\rho(0,0)\Omega_{1/2} \approx \rho(0,0)\,\pi\theta_{1/2}^2 \approx \rho_S \le 1 \ , \tag{8.30}$$

where $\theta_{1/2}$ is half-angle of the cone with solid angle equal to the solid angle of the reflection. The specular component of the total reflectance is given by ρ_S. The more the specular component approaches the ideal smooth surface result, the narrower the value of $\Omega_{1/2}$, and the larger the permissible value of $\rho(0,0)$.

A cosine-theta curve fit to the MRDF for large angles (well away from the specular peak) will yield the diffuse component

$$\rho_{DIFFUSE}(\theta,\phi) \approx \frac{\rho_D \cos\theta}{\pi} . \tag{8.31}$$

The diffuse component of the total reflectance is given by the coefficient of the curve fit ρ_D. Obviously, the total reflectance is approximately equal to the sum of the diffuse and specular components and must be less than unity

$$\rho_D + \rho_S \approx \rho \le 1 . \tag{8.32}$$

All of the preceding discussions have addressed materials with no transmittance. Radiation is either reflected or absorbed. However, the same concepts can be applied to transparent or semi-transparent materials. The reflected component should now be assumed to be all radiation that is neither absorbed nor transmitted without directional change. Thus, scattering is considered to be an element of reflection, as is Fresnel reflection. However, in a transparent material, radiation may be scattered into a sphere, rather than only a hemisphere. The MRDF (and BRDF) for transparent materials is defined the same as for opaque materials, but the expression relating MRDF to total reflectance must involve the integral over a sphere (rather than a hemisphere)

$$\rho = \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta \; \rho(\theta,\phi) . \tag{8.33}$$

The total reflectance is obviously not related to the hemispherical reflectance in transparent materials.

The reflectances described above are functions of wavelength. Two wavelength effects dominate. First, the total reflectance will vary with wavelength because the absorptance varies with wavelength. Because transmittance, absorptance, and reflectance must sum to unity, any change in one quantity is likely to change both of the others. Second, even if the total reflectance does not change appreciably with wavelength over a wide span of wavelengths, the MRDF may change significantly. Consider flame-sprayed aluminum. This material is a metal with a total reflectance near unity. Most metals reflect strongly from visible through radio wavelengths. However, the surface roughness of flame-sprayed aluminum has a spatial scale of the order of 0.1 mm. At infrared wavelengths, the surface is very rough and has a diffuse (almost Lambertian) MRDF. However, at microwave wavelengths, the surface is smooth (compared to the wavelength) and the MRDF will have a strong specular character. The total reflectance is near unity at both extremes, but the reflection properties are vastly different. As wavelength is scanned from the infrared to the microwave, the MRDF will transition from diffuse to specular character. At submillimeter wavelengths, the MRDF will have elements of both characters.

References

[1] International Standards Organization, "International System of Units (SI)", International Standard ISO 1000 (1971).

[2] American National Standards Institute, "Illuminating Engineering, Nomenclature & Definitions for", ANSI/IES RP16-1986 (1986).

[3] American National Standards Institute, "Nomenclature & Definitions for Illumination Engineering", ANSI Z7.1-1967 (1967).

[4] Skolnik, Merrill I., Introduction to Radar Systems 2nd Ed. (McGraw-Hill, New York NY, 1980) Chap. 2.

[5] Wolfe, William L., "Radiation Theory", Chapter 1 in Wolfe, William L. and Zissis, George J., (Eds.), The Infrared Handbook, Revised Edition, (Environmental Research Institute of Michigan, Ann Arbor MI, 1985).

[6] Environmental Research Institute of Michigan, "Target Signature Analysis Center: Data Compilation, Eleventh Supplement, Volume 1 - Bidirectional Reflectance: Definition, Discussion, and Utilization and Volume 2 - Bidirectional Reflectance: Graphic Data", AFAL-TR-72-226 (1972).

Problems

8-1. A blackbody has a temperature of 3 K. Plot the radiant exitance of this blackbody over the wavelength range from 100 μm to 10 cm. Ten to fifteen points is sufficient. What is the wavelength of peak emission? What is the radiant exitance at this wavelength (the peak)? Referenced to a standard blackbody at 300 K, what is the temperature increase in the standard blackbody that provides an increase in radiant exitance of that standard blackbody that is equal to the total radiant exitance of the 3 K blackbody? This temperature increase is the **signal-equivalent temperature difference.**

8-2. A material has a reflectance of 90%. If that material has a temperature of 300 K what is the radiant exitance from the surface?

8-3. A thin plate is perfectly absorbing on one side and perfectly reflecting on the other. The plate is in a vacuum and perfectly insulated from its environment except for radiation. Radiation from a 300 K blackbody is incident only on the absorbing side of the plate. What is the equilibrium temperature that the plate will reach? What temperature would the plate reach if both sides were perfectly absorbing and it is still illuminated from only one side?

8-4. A perfectly collimated laser beam is incident on a perfectly reflecting Lambertian surface at 45° angle of incidence. What is the angular distribution of the reflected radiant intensity?

8-5. Two blackbodies have variable temperatures that are exactly 1 K different. What is the temperature dependence of the difference between the radiant exitances of the two plates?

8-6. A laser emits 25 W of radiation at 514.5 nm. How many lumens does this laser emit? The laser radiation is shaped such that it uniformly illuminates a diffuse reflecting surface (100 cm^2 area) without loss of power. What is the luminous exitance (lux) coming from this surface?

8-7. A dedicated microwave communications link is established between two widely separated antennas. A transmitter is attached to one antenna while a receiver is attached to the other antenna. Using the nomenclature used in the radar equation example, determine an expression for the CNR of such a communications link as a function of antenna separation.

8-8. A specular reflector has a total reflectance of 1. If the peak value of the monostatic reflectance is 10^6 sr^{-1}, what the half-angle of the cone into which most of the radiation is reflected?

8-9. A mirror has a diffuse monostatic reflectivity component $\rho_{DIFFUSE}(\theta) = 6.366 \times 10^{-7} \cos\theta \, sr^{-1}$. What is the monostatic reflectance of the mirror if it is tilted at 60°? What is the maximum possible value of the specular component of the total reflectance?

8-10. A surface is covered with random bumps. The average bump height is 0.1 μm and the average spacing between adjacent bumps is 0.1 μm. What is the probable dominant character of the reflectance at a wavelength of 10 μm?

CHAPTER 9

SENSOR FUNCTIONS

Uses of Sensor Systems

There are few operations that are performed by modern combat platforms that do not involve sensors in one way or another. Sensors are used to navigate from the beginning to the end of a mission. They are used to detect, localize, track, and potentially identify possible targets. They are used to guide and control the weapons that are used to attack those targets. They are used to detect, localize, track, and possibly identify potential threats to the platform and to guide and control the weapons that are used to defend against those threats. They are used to determine the status of systems on board the platform. They are used to determine the status of the environment around the platform. There are so many functions that it is not possible to address them all in any specific fashion.

Figure 9-1 lists a number of the functions that combat sensor systems are commonly called upon to perform. Most of the listed functions are self-evident. However, a few of the functions may not be familiar to everyone. For example, pseudo-imaging is the acquisition of high-resolution but under-sampled or partially sampled data on a scene (usually containing a target). For example, a rosette scan will sample a scene only where the scan pattern covers it. The sampled data could be used to generate an image of sorts, however, it is seldom used to generate an observable image. It is more often used to track a target or possibly identify a target. Range profiling is the acquisition of high-resolution range data on a target without any accompanying angular resolution data. It has limited use in target identification. The meanings of terms like target recognition, target identification, target orientation estimation, etc., are discussed in great detail in Chapter 12. Situational awareness is the determination of the location and apparent motion of all platforms (whether friendly, hostile, or neutral; and air, surface, or subsurface) in a spherical volume out to some range surrounding one’s own platform. It is useful in permitting a commander to determine how a complex situation is evolving. Terrain following is straight line flight along a vertical path that attempts to maintain a constant altitude above the ground. Terrain avoidance is flight along a three dimensional path that attempts to maintain a constant altitude above the ground but takes advantage of the ability to minimize altitude variations (with respect to sea level) by going around taller terrain features. Obstacle avoidance is the ability to detect vertical and horizontal obstacles (such as radio towers and or power transmission lines and maneuver over, under, or around them while maintaining normal terrain following or terrain avoidance flight. Bathymetry is the measurement/mapping of ocean bottom depth or the depth of targets beneath the ocean surface. Vibratometry is the measurement of the vibrations of an object. It can be useful in determining the status of bearings on rotating equipment. It can also be used to identify tactical targets by their unique vibrational characteristics.

Table 9-1. Common sensor functions.

SEARCH

TARGET DETECTION

TARGET LOCATION
- RANGE
- AZIMUTH
- ELEVATION

VELOCITY/RANGE RATE DETERMINATION

TRACKING
- ANGLE ONLY
- RANGE & ANGLE
- RANGE, ANGLE, & VELOCITY

IMAGING

PSEUDO-IMAGING

RANGE PROFILING

OBJECT RECOGNITION

OBJECT IDENTIFICATION

OBJECT ORIENTATION ESTIMATION

MAPPING

RECONNAISSANCE

SURVEILLANCE

SITUATIONAL AWARENESS

NAVIGATION
- TERRAIN CONTOUR MAPPING (TERCOM)
- FEATURE-BASED
- DOPPLER
- INERTIAL
- CELESTIAL

SYSTEM STATUS
- CONSUMABLES (FUEL LEVEL, OXYGEN, etc.)
- PROPULSION (ENGINE RPM, TEMP., etc.)
- ELECTRICAL (VOLTAGE, AMPERAGE, etc.)
- HYDRAULIC (PRESSURE)
- DAMAGE CONTROL (TEMPS, WATER LEVELS, etc.)

COMMUNICATIONS RECEPTION

AIR SPEED MEASUREMENT

GROUND SPEED MEASUREMENT

ALTIMETRY

TERRAIN FOLLOWING

TERRAIN AVOIDANCE

OBSTACLE AVOIDANCE

HOMING GUIDANCE
- ACTIVE
- SEMI-ACTIVE
- PASSIVE

AIMPOINT SELECTION

MODULATION/DEMODULATION

TARGET DESIGNATION

MUNITION FUZING
- PROXIMITY
- IMPACT
- ALTITUDE (BAROMETRIC)
- ALTITUDE (RADAR)

DAMAGE/KILL ASSESSMENT

ENVIRONMENTAL REMOTE SENSING
- ANEMOMETRY (WIND SPEED)
- TURBULENCE DETECTION
- THERMOMETRY
- ATMOSPHERIC CONSTITUTION
- CONTAMINATION DETECTION
- SOIL/ROCK/VEGETATION TYPE

BATHYMETRY

VIBRATOMETRY

The Six "Elemental" Sensor Functions

The preceding section has presented a fairly comprehensive catalog of functions that modern combat systems are performed. As stated earlier, the list is too long to discuss each function in detail. However, after performing a detailed analysis of exactly what is meant by most of the functions listed in Table 9-1 the author has come to believe that virtually all of them can be decomposed into combinations of six "elemental" sensor functions: search, detection, estimation, modulation/ demodulation, image-based perception, and tracking. That is, by performing several of the elemental functions in the proper sequence, with the proper signal conditioning between functions, the most complex function can be synthesized.

Search is the collection of activities involved in providing sensor coverage of areas of interest. This function concerns itself with providing the opportunity for sensors to perform other subsequent functions such as target detection or tracking. It is a separate function from those subsequent functions. It is possible to detect an object without search although this requires consideration in the sensor design. It is also possible to successfully and efficiently search an area without accomplishing any detections (or indeed without any possibility of successful subsequent detection). The search function involves coordination of sensor scanning and platform motions to cover a designated area in minimum time and with maximum efficiency of coverage and sensor utilization. Chapter 10 elaborates on the nature of the search function and provides useful results that can impact sensor system design and performance.

Detection is the simple act of determining whether an object of potential interest is present or absent in the sensing volume of a system. It says nothing about the object other than it is present. It does not indicate whether it is a hostile target, a friendly platform, or even that it is a military platform. Detection simply says whether or not something other than "noise" is present in the sensor signal. The ability with which a sensor can detect a target will depend on the carrier-to-noise ratio achieved in the sensor, the false alarm rate that is acceptable, and the inherent characteristics of the targets. Chapter 11 discusses the elemental function of detection in detail.

Estimation is the process of determining the value of some property of a "target". Range, velocity, altitude, azimuth, elevation angle, acceleration, signal "size" are examples of properties that are "estimated" by processing the signals from appropriate sensors. The precision with which the value of a parameter may be estimated will be shown to be limited by the resolution of the measuring sensor and the signal-to-noise ratio achieved. Chapter 12 discusses the elemental function of estimation in detail.

Modulation and demodulation are seldom end functions of sensors but one or the other or both are often critical intermediates in the accomplishment of desired end functions such as parameter estimation or tracking. In some instances the sensor will perform both modulation and demodulation. For example, a radar may transmit a frequency-modulated waveform and receive a modified frequency-modulated waveform in the pursuit of estimating the range and velocity to a target. In others, one sensor will establish a modulation for another sensor to utilize, such as a coded laser designator signal for a laser guided weapon. In others the interaction of a sensor signal with the target may produce a modulation that is useful for another function, as in a conical scan tracking system. Many different kinds of modulation are utilized. These include amplitude modulation, frequency modulation, phase modulation, polarization modulation, and various pulse modulations. Chapter 13 discusses modulation and demodulation and its relation to sensor systems.

Imaging, or more properly, **image-based perception**, is an elemental function that is more complex than detection or estimation. Imaging itself is merely a way in which sensor data is collected and organized. However, when an image is presented to a human observer to process, the perceptual processes that ensue are critical functions in many weapons systems. For example, in an electro-optical targeting system, a “television” sensor gathers a near infrared image of a scene. The image is presented to the weapons operator who examines it and may or may not detect and/or identify a target, based only on his examination of the image. The human perceptual system has been characterized both empirically and semi-empirically in terms of describable target and image characteristics. Chapter 14 discusses imaging and image-based perception in considerable detail.

Tracking is the process by which the motions of detected objects are followed over time and predicted into the future. This elemental function is also somewhat more complex than simple detection or estimation. Tracking can be achieved by dedicated sensor and processor hardware. It can also performed entirely as a processing function if an adequate detection sensor is present. The former dedicated systems are basically servo systems. Target motion produces an error signal which the servo acts upon to reduce. As the error signal is kept within bounds, the position of the target is kept within known limits. Tracking systems which do nothing but process data from detection systems rely on tracking filters. This filters average measurements and predict motions. Subsequent measurements are compared against predictions and the prediction is continuously corrected. Chapter 15 describes both types of tracking. Principles and implementation of dedicated tracking systems are described. The theory of tracking filters is presented and the performance of different filters is compared.

The author is not completely convinced that this list of elemental sensor functions is complete, yet so far has not identified others that seemingly fulfill all of his criteria. He may change his mind in the future. In establishing this list, he restricted it to functions that have direct utility, that is, the output of the elemental function is of immediate interest to the user or the function itself is mission critical. He also restricted the list to those functions for which considerable theoretical analysis is required to determine their impact on sensor performance. He tried to eliminate functions that are implemented in single components and are simply considered tools of the sensor designer.

Incidental Sensor Functions

In light of the preceding section where six elemental sensor functions were identified, a number of incidental functions were also identified. These are functions that sensor designers incorporate to improve performance, permit use of specific technologies, or perform necessary signal processing functions. They are not elemental in that there are usually alternative sensor implementations that can avoid their incorporation, yet can perform the same top level functions for the user. For example, analog-to-digital conversion may simplify the processing of sensor signals, yet, analog processing can almost always yield comparable results (most sensor technologies predate the invention of the digital computer). In Table 9-2 we list some of the more significant incidental sensor functions.

Figure 9-2. A Partial List of Incidental Sensor Functions.

TARGET ILLUMINATION
DISPLAY
TIMEKEEPING
AMPLIFICATION
FILTERING
HETERODYNE MIXING
ANALOG-TO-DIGITAL CONVERSION
DIGITAL-TO-ANALOG CONVERSION
SIGNAL MULTIPLEXING/DEMULTIPLEXING
FAST FOURIER TRANSFORM (FFT)
SIGNAL INTEGRATION
SENSOR POINTING
DETECTOR COOLING
INERTIAL LINE-OF-SIGHT STABILIZATION
ELECTRICAL POWER CONDITIONING
ENVIRONMENTAL CONDITIONING
DATABASE MANAGEMENT
DIGITAL DATA PROCESSING

Problems

9-1. Identify a sensor function that is not included in Table 9-1.

9-2. Pick a complex sensor function from Table 9-1 and identify how the elemental functions combine and interact to create that complex function. Identify any critical subfunctions that are not covered by the six elemental functions.

9-3. Nominate a candidate for a seventh elemental sensor function. Justify your nomination. This function need not be listed in Table 9-1. Does your nomination suggest a need for more than seven elemental functions? That is, is there some related function identified that is neither part of the six original functions nor your nominated seventh function?

CHAPTER 10.

SEARCH

Introduction to Search

Search is the collection of activities involved in providing sensor coverage of areas of interest. This function concerns itself with providing the opportunity for sensors to perform other subsequent functions such as target detection or tracking. Search involves the coordination of sensor scanning and platform motions to cover a designated area in minimum time and with maximum efficiency of coverage and sensor utilization. Search and detection are separate functions – it is possible to search with sensors that are malfunctioning or are otherwise incapable of adequate detection performance. Nevertheless, search and detection are interdependent. Without effective search, the detection abilities of even the best sensors are underutilized. Conversely, without adequate detection ability, even an optimum search strategy may produce unsatisfactory results.

Depending on the mission being performed and the conditions under which it is performed, search may be carefully planned and executed, it may be simple, regular, and repetitive, or it may be random. It may involve continuous observation during a continuous sweep or it may involve discrete looks at discrete positions. It may be carried out by stationary platforms or it may require platform motion to accomplish. The detection range and coverage of the sensors utilized will affect the choice of search type and the speed at which it can be accomplished. The interrelation of sensor and search type are the subject of this Chapter.

The search problems commonly encountered in combat systems applications can usually be sorted into one of two general types. One search type involves the one-time search of a large area using a sensor of limited range operated from a mobile platform. We will refer to problems of this type as **area search** problems. Typical area search applications include anti-submarine patrol by long-range aircraft, target acquisition by antiship missiles, mine-hunting, and “SCUD hunting” by attack aircraft. The shape of the search area, the range of the detection sensor, the mobility and degree of reaction expected from the target being searched for, and the time available for the search will all affect the nature and potential for success of area searches. A second type of search involves repeated search of a designated solid angle out to some required range by a scanning sensor operated from a platform that may be assumed to be stationary. This type of problem will be called **volume search**. Air defense search, ballistic missile early warning, and target acquisition by anti-aircraft missile seekers are typical examples of volume search. The solid angle to be searched, the required detection range, the revisit rate required, and the angular resolution of the search sensor, among other things, will all impact volume search performance. In subsequent sections we will address the peculiarities of area search and volume search in more detail.

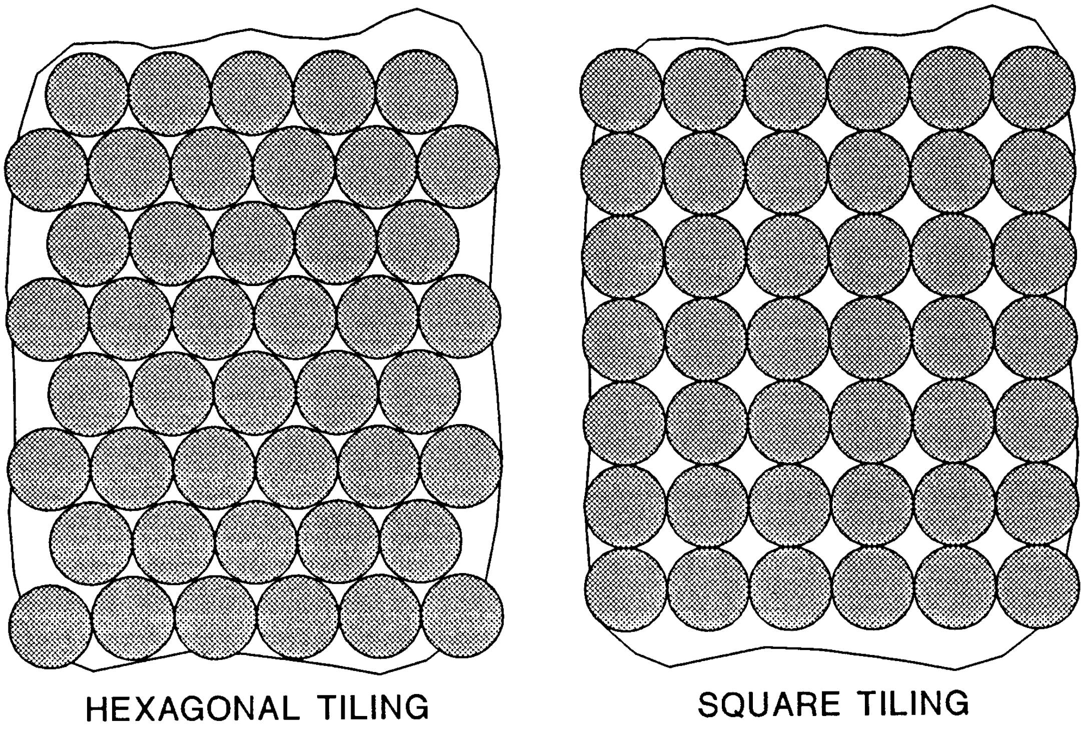
Regardless of search type, some search systems will utilize a continuous scan of a suitable sensor system while others will use measurements obtained at discrete positions, the discrete positions being scanned over the search area. Discrete measurements complicate the search process in ways that are described in the following section.

Finally, the ultimate goal of search is target detection. Detection is addressed in Chapter 11. The effects of non-unity detection probability on the choice of search patterns and of search parameters are discussed in the last section of this chapter.

Circle Packing

Discrete measurement search involves the one-at-a-time observation of limited portions of the search region. It is very common that each observation covers a region that is roughly circular. For example, if the measurement is made by a single pulse from a radar, the beam of the radar is likely to be circular. If the “measurement” is an azimuthal scan by a radar or a conical scan by an infrared sensor, the area covered by the scan will be circular. Using this circular assumption (extension of any circular region result to an elliptical region result is trivial) we observe that it is impossible to “tile” any search region with circular tiles without either leaving gaps (reduced degree of coverage) or having significant overlaps (reduced efficiency of coverage). Figure 10-1 shows two tiling patterns (hexagonal and square) drawn with no overlap between the circular tiles. Both patterns show that there is considerable uncovered area. To reduce the uncovered area to zero will require overlap between tiles and a reduction in the efficiency of coverage. The relationship between completeness of coverage and the efficiency of coverage is analyzed in the **circle packing** problem. It should be noted that the two tiling patterns studied here (hexagonal packing and square packing) are the only two patterns of packing equally sized circles observed in nature. Hexagons, squares, and triangles are the only regular polygons that can be tiled without overlap and without gaps (and triangles can only be tiled if the orientation of the triangles is alternated.

Figure 10-1. Two regular tiling patterns using circular “tiles”.
The patterns as drawn have zero tile overlap.



Consider first the case where the circular coverages are sparsely (yet still regularly) tiled. That is, the separation between the centers exceed the radii of the circles. Let *a* denote the radius of a circular coverage area and let *2d* denote the separation between nearest neighbor circle centers. In this case it is relatively easy to calculate that the fraction of the total area covered by circles is given by

$$\eta_{HEX} = A_{CIRCLE} / A_{HEXAGON} = \pi a^2 / 2\sqrt{3}\, d^2 = \left(\pi / 2\sqrt{3}\right)\xi^{-2} \tag{10.1}$$

for hexagonal packing and by

$$\eta_{SQ} = A_{CIRCLE} / A_{SQUARE} = \pi a^2 / 4d^2 = \left(\pi / 4\right)\xi^{-2} \tag{10.2}$$

for square packing. We have defined a dimensionless separation parameter $\xi = d/a$. The coverage is relatively poor (roughly 90% for hexagonal packing and 79% for square packing) even if $\xi = 1$.

Improved coverage requires overlap. Figure 10-2 shows the geometry of coverage and overlap in the two forms of packing. Figure 10-3 shows the geometry of hexagonal packing in a higher level of detail. Determination of the area that is not covered (the small triangular shaped region at the center of the figure) is not amenable to a simple mathematical expression. It requires calculation of the area of more regular geometric objects and adding and subtracting the areas of those objects until the residual corresponds to the area of interest.

Figure 10-2. Geometry used in calculating overlap and coverage gaps in a) hexagonal packing (or tiling) and b) square packing. Overlap regions are shown by +45° diagonal shading. Uncovered regions are shown by -45° diagonal shading.

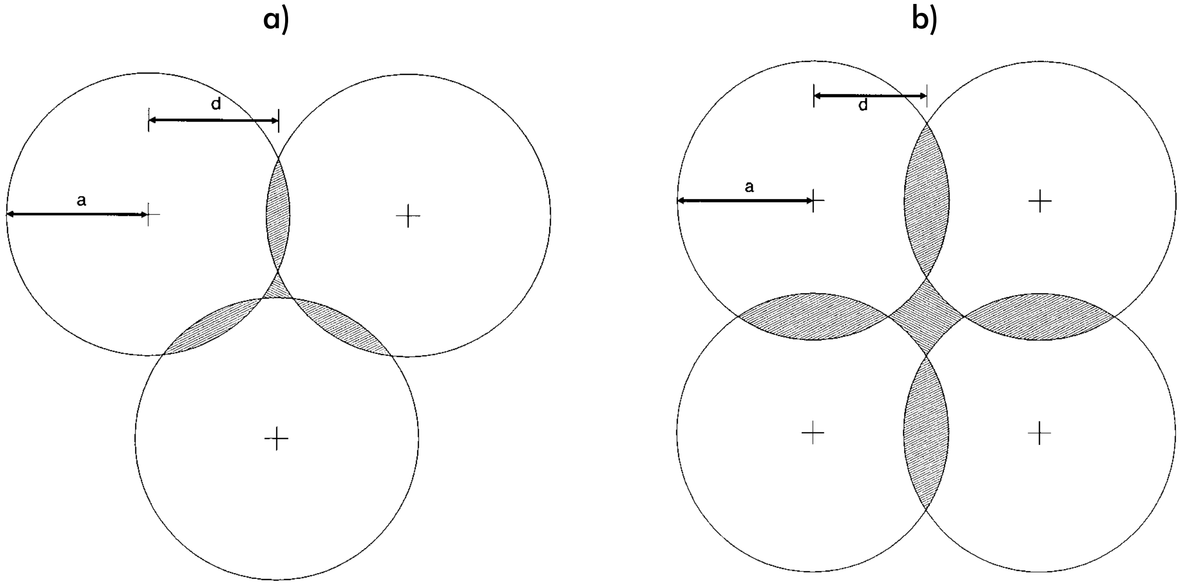
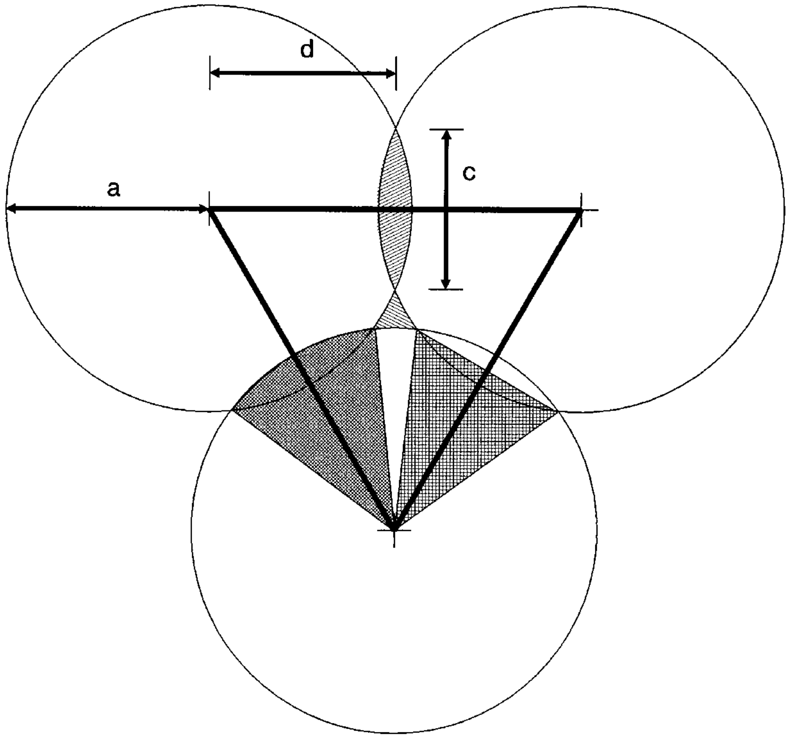


Figure 10-3. Geometry of overlap and non-coverage calculations in hexagonal packing showing the different geometric regions used in the calculation.



The area of the small triangular region (-45° striped) that is not covered by a circle can be calculated using the following general expression

$$\begin{aligned} \textit{Area Not Covered} &= \textit{Area of Large Triangular Region} \\ &\quad - 3 \text{ x } \textit{Area of a } 60° \textit{ Sector} \\ &\quad + 3/2 \text{ x } \textit{Area of The Lens-Shaped Overlap Region} \\ &= A_{LTR} - 3A_{60S} + 1.5A_{LSOR} \end{aligned} \tag{10.3}$$

which can be further decomposed into

$$\begin{aligned} \textit{Area Not Covered} &= \textit{Area of Large Triangular Region} \\ &\quad - 3 \text{ x } \textit{Area of a } 60° \textit{ Sector} \\ &\quad + 3/2 \text{ x } 2 \text{ x } (\textit{Area of Cross-Hatched Sector} \\ &\qquad\qquad - \textit{Area of Checked Triangle}) \\ &= A_{LTR} - 3A_{60S} + 3(A_{CHS} - A_{CT}) . \end{aligned} \tag{10.4}$$

The area of the large equilateral triangle is easily seen to be

$$A_{LTR} = \sqrt{3}d^2 \tag{10.5}$$

as is the area of the three 60° sectors

$$3\,A_{60S} = 0.5\pi a^2\,. \tag{10.6}$$

The area of the cross-hatched sector is given by

$$A_{CHS} = 0.5\theta a^2 = a^2 \cos^{-1}\left(d/a\right). \tag{10.7}$$

The area of the checked triangle is given by

$$A_{CT} = 0.5cd = d\sqrt{a^2 - d^2}\,. \tag{10.8}$$

The area not covered by the circles can now be calculated by

$$A_{NC} = \sqrt{3}d^2 - 0.5\pi a^2 + 3\left(a^2 \cos^{-1}\left(d/a\right) - d\sqrt{a^2 - d^2}\right). \tag{10.9}$$

For each circular spot added to the hexagonal pattern, two small triangular areas with no overlap are produced. Each added circular spot covers an effective hexagonal area

$$A_{HEX} = 2.59808x^2 = 2\sqrt{3}d^2 \tag{10.10}$$

where *x* is the side of the hexagon. Thus the effective fractional coverage in hexagonal packing is

$$\eta_{HEX} = 1 - \frac{\left(\sqrt{3}d^2 - 0.5\pi a^2 + 3\left(a^2 \cos^{-1}\left(d/a\right) - d\sqrt{a^2 - d^2}\right)\right)}{\sqrt{3}d^2} \tag{10.11}$$

or

$$\eta_{HEX} = \frac{0.5\pi a^2 - 3\left(a^2 \cos^{-1}\left(d/a\right) - d\sqrt{a^2 - d^2}\right)}{\sqrt{3}d^2}\,. \tag{10.12}$$

If we define a dimensionless ratio $\xi = d/a$, then the fractional coverage becomes

$$\eta_{HEX} = \left(1/2\sqrt{3}\right)\pi\xi^{-2} - \sqrt{3}\left(\xi^{-2}\cos^{-1}(\xi) - \xi^{-1}\sqrt{1-\xi^2}\right) \tag{10.13}$$

Eq. (10.13) is essentially the same as Eq. (10.1) with a correction accounting for overlap. The efficiency in covering the area is given by

$$\varepsilon_{HEX} = 1 - \frac{3\left(a^2 \cos^{-1}(d/a) - d\sqrt{a^2 - d^2}\right)}{\pi a^2} \tag{10.14}$$

$$\varepsilon_{HEX} = 1 - (3/\pi)\left(\cos^{-1}\xi - \xi\sqrt{1-\xi^2}\right) \tag{10.15}$$

Following the general approach taken above, but without showing the details, the square packing problem yields the following results. In the square coverage pattern, the diamond shaped area not covered has area

$$A_{NC} = 4d^2 - \pi a^2 + 4\left[a^2 \cos^{-1}(d/a) - d\sqrt{a^2 - d^2}\right] \tag{10.16}$$

The effective fractional coverage is given by

$$\eta_{SQ} = 1 - \frac{4d^2 - \pi a^2 + 4\left[a^2 \cos^{-1}(d/a) - d\sqrt{a^2 - d^2}\right]}{4d^2} \tag{10.17}$$

or

$$\eta_{SQ} = \frac{\pi a^2 - 4\left[a^2 \cos^{-1}(d/a) - d\sqrt{a^2 - d^2}\right]}{4d^2} \tag{10.18}$$

or

$$\eta_{SQ} = (\pi/4)\xi^{-2} - \left[\xi^{-2} \cos^{-1}(\xi) - \xi^{-1}\sqrt{1-\xi^2}\right]. \tag{10.19}$$

The efficiency in covering the area is given by

$$\varepsilon_{SQ} = 1 - \frac{4\left[a^2 \cos^{-1}(d/a) - d\sqrt{a^2 - d^2}\right]}{\pi a^2} \tag{10.20}$$

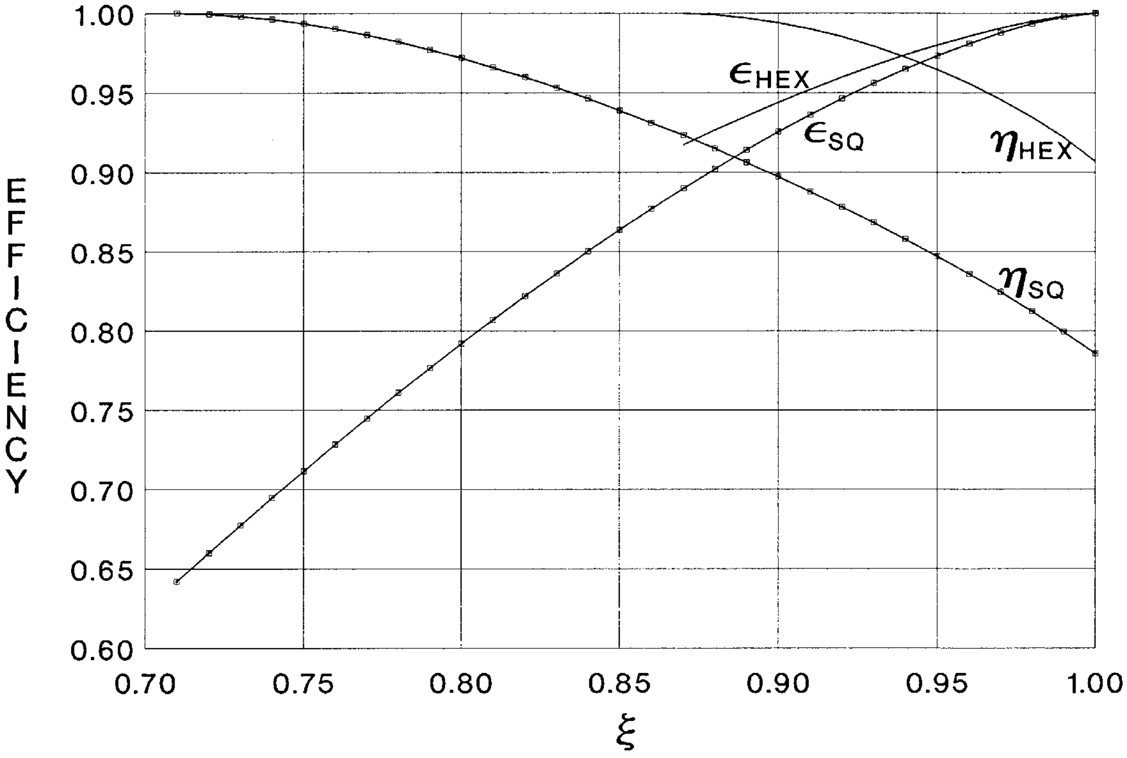
or

$$\varepsilon_{SQ} = 1 - (4/\pi)\left[\cos^{-1}(\xi) - \xi\sqrt{1-\xi^2}\right] . \tag{10.21}$$

The results for the fractional coverage and the coverage efficiency are summarized in graphical form in Figure 10-4. Only the critical region between zero overlap (*ξ =1*) to complete overlap are plotted. It is obvious that hexagonal packing is superior to square packing. Considerably less overlap is required to achieve complete coverage. The worst case efficiency is only about

92%. Square packing requires considerably more overlap to achieve complete coverage. In square packing, the worst case efficiency can be as low as 64%. However, hexagonal packing is considerably harder to accomplish in practice than is square packing. That is why both types of packing have analyzed here.

Figure 10-4. Fractional coverage (η) and coverage efficiency (ε) for hexagonal and square circle packing as function of the overlap parameter ξ.

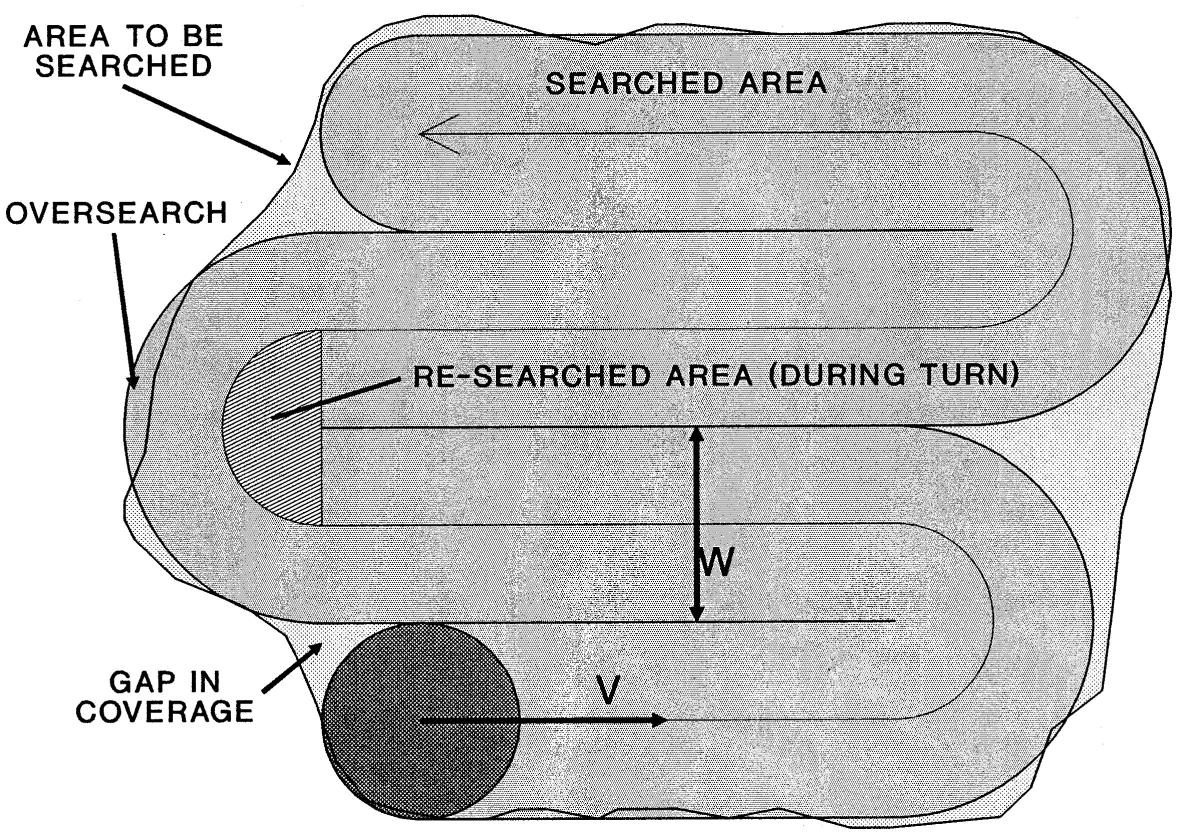


Area Search

Let us consider searching an irregular area with a search sensor with finite area coverage. The configuration envisioned is similar to that shown in Figure 10-5. The search process is desired to regular, that is, it can be accomplished without overly complex and frequently changing motions. Such a search will exhibit three kinds of inefficiencies. The regular motion will cause some small areas to be missed, causing gaps in the search coverage. It will also cause some search effort to be expended outside the search area, an inefficiency this author calls oversearch. The regular motion may also cause the finite-sized instantaneous search area to turn back upon itself. This invariably happens in back and forth search motions. The small area which is scanned more slowly or scanned more than once can be called re-search.

All three processes are undesirable. Gaps in coverage can lead to reduced detection probability. Oversearch and re-search do not contribute to detection (ostensibly any targets are contained within the designated search area) but cause extra time to be consumed. This delays detection which gives the target more time to react by attempting to escape, by attacking the searcher, or by accomplishing more of its original mission.

Figure 10-5. Inefficiencies in area search.



In ideal exhaustive search it is assumed that the area to be searched is completely covered once (with no gaps in coverage), without wasted effort searching outside the designated area (over-search) and without searching any area twice (re-search). Let A_T be the actual area of the region to be searched. Let the search be accomplished by a sensor system capable of search a swath of width *W* carried by a platform moving at velocity *V*. An ideal exhaustive search can be completed in a time

$$T = A_T / WV \; . \tag{10.22}$$

The fraction of the total area that can be searched in time *t* is given by

$$A(t) = WVt \tag{10.23}$$

or

$$A(t) / A_T = WVt / A_T = t / T \; . \tag{10.24}$$

In a time *T* the entire area can be searched. We will assume a "cookie cutter" detection process. That is if the target is within an area that has been exhaustively searched, it will have been detected with unit probability. If it is outside the area that has been searched, it will not be detected. That is, detection occurs when the swath of the search sensor passes over the location of the target. As a function of elapsed time we have

$$P(t) = WV\,t / A_T = \gamma\, t \tag{10.25}$$

where *P*(*t*) is the cumulative detection probability. If a target is located at a random position within the search area, then the probability density of detecting the target is a uniform function of time with a range from 0 to *T*. The expected value of the time required for target detection is *T/*2.

Let us consider a random search of the area rather than an exhaustive search. A random search covers the search area at a fixed rate of area coverage per unit time, but rather than visiting each "segment" of area one at a time in order until the area is covered, the "segments" are selected and searched at random. Each segment has the same probability of being selected next as any other segment, including those already searched. Thus, there is a small but finite probability that some segment will never be searched.

Assume the target has not yet been detected. Because the target is as likely to be in the next small volume to be searched as it is in any other small volume, the probability density that the target will be detected in a unit interval of time $t \rightarrow t+\Delta t$ is given by

$$p(t)\Delta t = \gamma\,\Delta t \; . \tag{10.26}$$

Let the quantity *q*(*t*) represent the probability that the target has not been detected by time *t*. Then

$$q(t) = 1 - P(t). \tag{10.27}$$

The probability that the target is still not detected by the time $t+\Delta t$ later is given by the probability that it has not been detected at time t multiplied by the probability that the target is not detected in the interval $t \rightarrow t+\Delta t$. That is,

$$q(t+\Delta t) = q(t)\left[1 - \gamma \Delta t\right] \tag{10.28}$$

Rearranging terms we have

$$\frac{q(t+\Delta t) - q(t)}{\Delta t} = -\gamma q(t) \tag{10.29}$$

Since Δt is arbitrary, we may take the limit as it approaches zero. In this case Eq.(10.29) reduces to a differential equation

$$\frac{dq(t)}{dt} = -\gamma q(t) \quad \text{or} \quad \frac{dq(t)}{q(t)} = d \ln q(t) = -\gamma dt\ . \tag{10.30}$$

The latter form of the equation is easily recognized as the logarithmic derivative with solution

$$q(t) = e^{-n(t)} = e^{-\gamma t} \tag{10.31}$$

The cumulative probability of detection as a function of time then becomes

$$P(t) = 1 - e^{-\gamma t} = 1 - e^{-WVt/A_T} \tag{10.32}$$

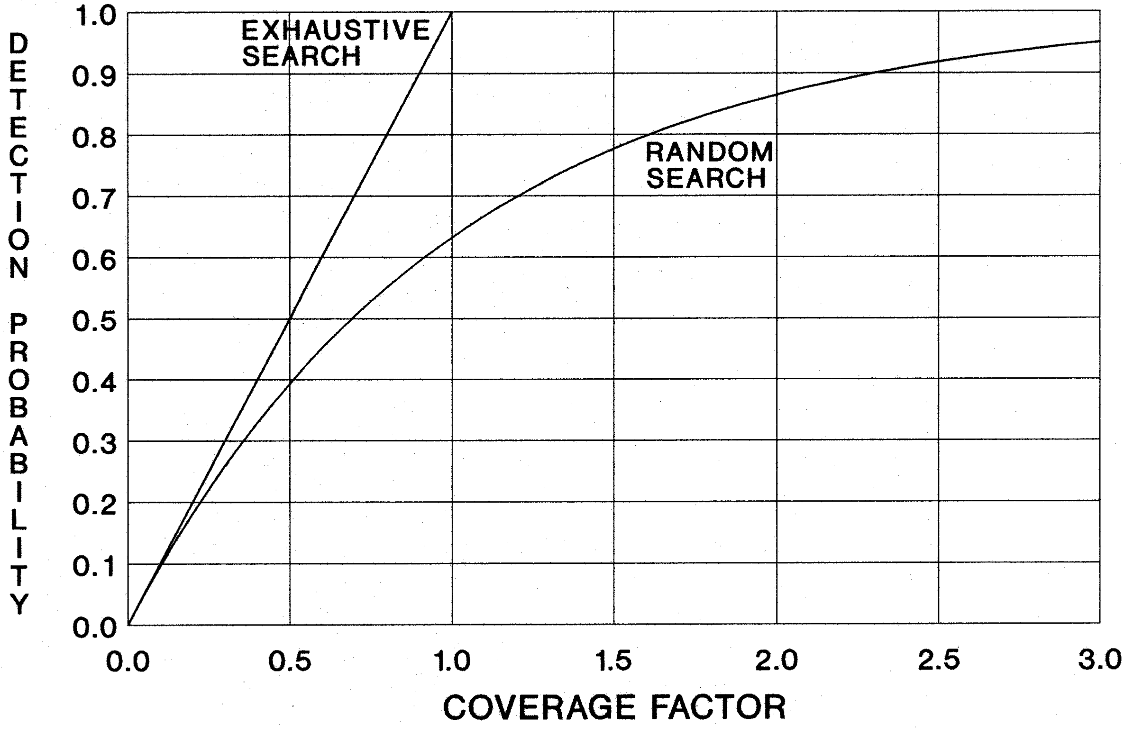
The cumulative probability for random search (Eq.(10.32)) is compared with that of exhaustive search (Eq.(10.25)) in Figure 10-6. Clearly exhaustive search is superior. Note: We have assumed that γ is constant in the derivation above. If for some reason γ is a function of time, i.e., $\gamma = \gamma(t)$, then the exponent ($-n(t) = -\gamma t$) in Eq.(10.32) can be replaced by the integral of $\gamma(t)$ over time

$$-n(t) = -\gamma t \rightarrow -\int_0^t dt\, \gamma(t) \tag{10.33}$$

with no loss in validity.

The question arises as to why bother to derive the random search expression? Before answering this, let us ask another question. What if the target (implicitly assumed to be stationary

Figure 10-6. Comparison of detection probability versus time for exhaustive and random search.



in the analysis above) is free to move about within the search area and attempt to evade the searcher? Empirically, it has been found that evading targets will turn an exhaustive search problem into a random search problem. This does not mean that the searcher should use a random search pattern. The searcher should almost certainly pursue one of the exhaustive search patterns, but that searcher will need to search the area repeatedly. The probability of detection as a function of time will more likely follow Eq.(10.32) rather than Eq.(10.25).

A special case of area search results when a target is initially known to be at a precise location, but the "searching" platform is located a considerable distance away. At the time of initial detection, the target leaves that position at a substantial speed in a random direction. The search sensor must now try to find the target in an area that is constantly expanding in time. A prime example of this can be found in undersea warfare. When a submarine attacks a ship, the radio signal (SOS) from the ship results in the submarine's position being very precisely known at the time of attack. Anti-submarine warfare (ASW) patrol aircraft may be hundreds of km distant. They reach the position of the attack many minutes if not hours after the submarine has left the area. The aircraft does not know which direction the submarine has taken. It must then attempt an exhaustive search of an ever-expanding area. Let the target escape velocity be U and the time delay before beginning the search be τ. At time t the area containing all possible target positions is given by

$$A_T(t) = \pi U^2 (t+\tau)^2 \ . \tag{10.34}$$

If the swath width and speed of the search platform are the same as defined earlier, then the time dependent probability of detection

$$p(t) = \gamma(t) = VW / A_T = \frac{VW}{\pi U^2 (t+\tau)^2} \ . \tag{10.35}$$

If we evaluate the probability exponent

$$n(t) = \int_0^t dt\, \gamma(t) = \int_0^t dt \ \frac{VW}{\pi U^2 (t+\tau)^2} = \frac{VW}{\pi U^2} \left\{ \frac{1}{\tau} - \frac{1}{t+\tau} \right\} \tag{10.36}$$

we then have

$$P(t) = 1 - e^{-\frac{VW}{\pi U^2}\left\{\frac{1}{\tau} - \frac{1}{t+\tau}\right\}} \ . \tag{10.37}$$

If we let $t \rightarrow \infty$, then we find

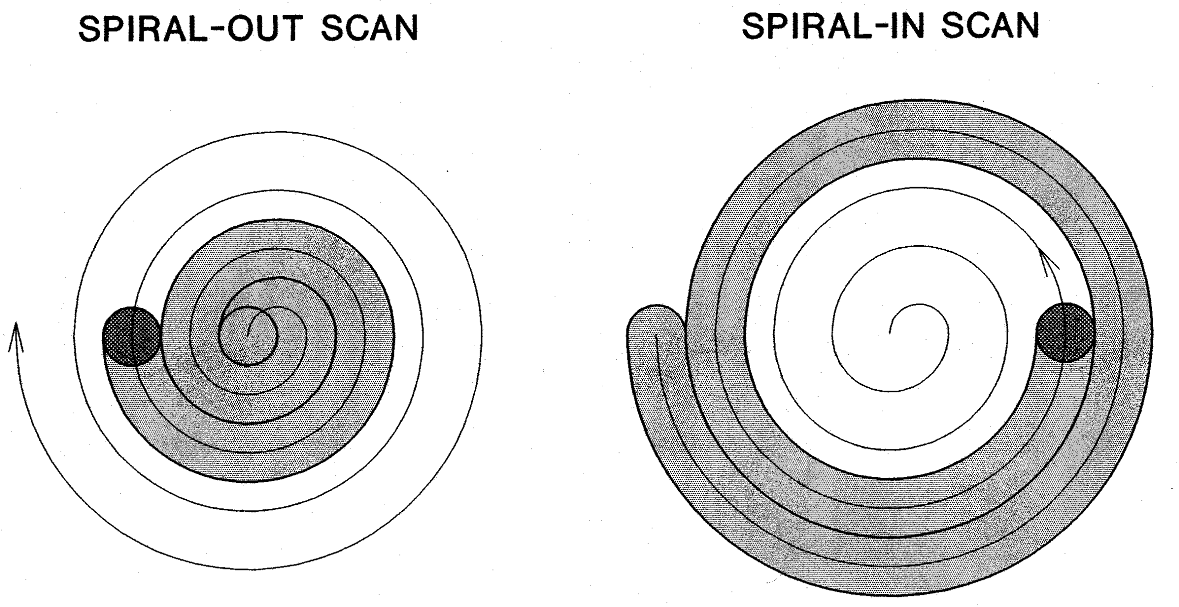
$$P(\infty) = 1 - e^{-\frac{VW}{\pi U^2 \tau}} < 1 \, . \tag{10.38}$$

That is, the detection probability does not go to unity no matter how long the platform searches. This is reasonable, because although the area searched grows linearly with time, the area that needs to be searched grows as the square of the elapsed time. The rate of probability growth is highest immediately after time τ and falls rapidly as t becomes large compared to τ. If the target is not detected quickly, it will probably never be detected.

However, there must be a limit to the validity of Eq.(10.37). If the distance the target can move $U\tau$ is smaller than one-half the swath width, i.e., $W/2$, then it is clear that if the search swath passes directly (centered) over the original target location, then it must also pass over the current location of the target. It must be possible for the detection probability to be unity for some choice of variables. We have just shown that $P(\infty) = 1$, if $2U\tau/W < 1$. It is probable that somewhat larger values of $2U\tau/W$ may also permit unity detection probability, but an exact result is not available.

What search pattern should be used in exhaustive search? If nothing is known about the target beforehand, except that the target is not likely to attempt evasion (e.g., because it is a fixed site), then a simple **back-and-forth** scan, as shown in Figure 10-5, may be considered optimum. It provides relatively complete coverage with minimal oversearch and minimal re-search. If on the other hand, it is known that a stationary target has a higher probability of being near the center of the search area than near an edge, then a "**spiral-out**" search pattern (as shown in Figure 10-7) is

Figure 10-7. Spiral scan patterns. The spiral-out pattern yields minimum search times for targets that are more likely near the center of a search area. The spiral-in pattern yields the maximum detection probability for maneuvering targets.



preferred. The search is started at the most probable location and the area is covered by an ever-growing spiral. The spiral out pattern has a search rate that is comparable to the back-and-forth pattern but on the average will provide earlier detection (the high probability area is searched first followed by the lower probability areas). There are many instances where it is known that the probability is higher that the target is near the center of the search pattern. For example, if a submarine is initially detected by a surveillance system to be located at a given location and traveling at known speed in a fixed direction, then the most probable location is the forward projection of the established track. This will be the true position if the submarine does not change speed or maneuver. The longer the submarine waits to maneuver, the closer it will be to the predicted location. Since a maneuver may be assumed to be equally likely in any interval of time, only the least likely maneuvers (those made immediately after the track was established) can produce targets at the edge of the uncertainty volume (which also comprises the largest volume element). The submarine is more likely to be within the inner 25% of the search area than it is in the outer 75%.

If it is suspected that the target will attempt to evade detection, but the search swath and velocity are large and the target velocity is small, then a "**spiral-in**" or **trapping** spiral (also shown in Figure 10-7) may prove advantageous. The search pattern circles the likely uncertainty area and moves in. If R_S is the average distance from the center of the spiral of one loop around the spiral, W is the swath width, V is the search velocity, and U is the target velocity, the time required to search that loop of the spiral is given by

$$T_S = 2\pi R_S / V \ . \tag{10.39}$$

The maximum radial distance R_R that the target can move in this time is

$$R_R = UT_S = 2\pi R_S U / V \ . \tag{10.40}$$

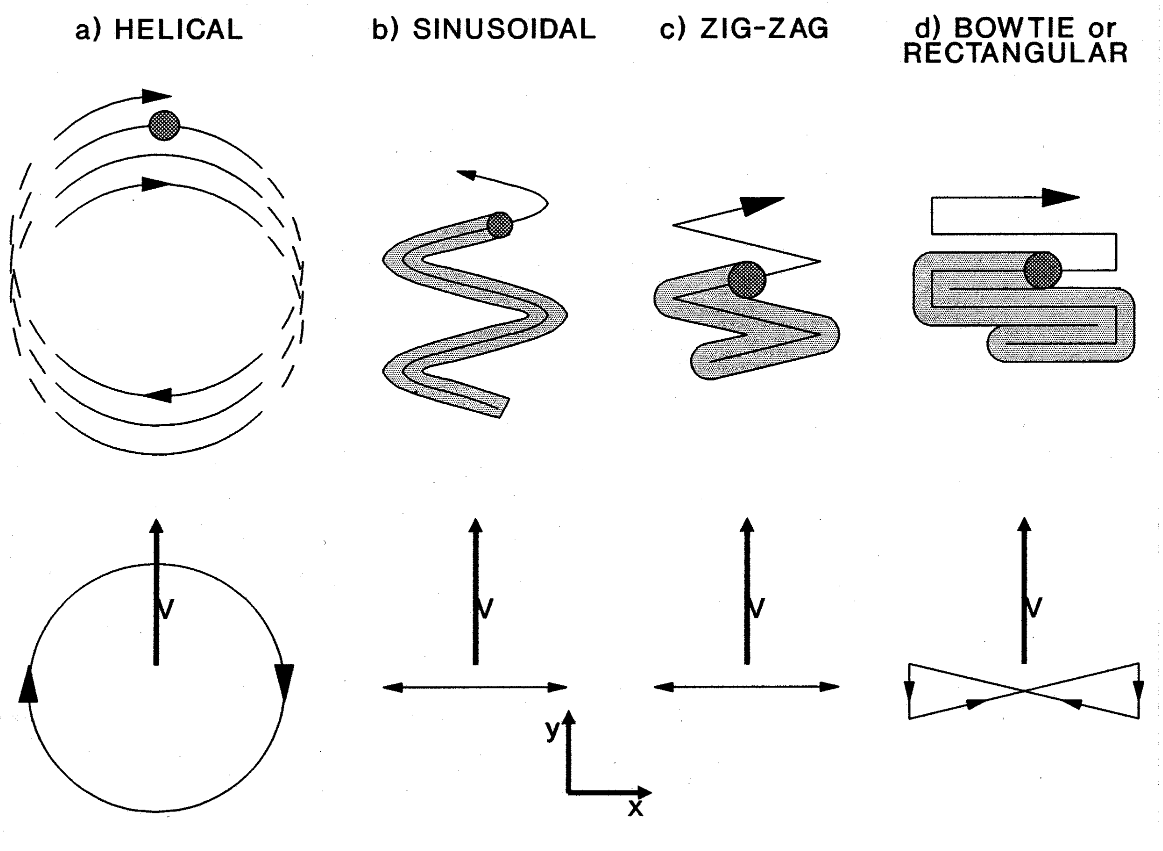
If R_R is smaller than W then the target cannot escape from the ever-narrowing spiral. Each succeeding loop around the spiral takes less time than the preceding loop. Thus, the target can only move a shorter and shorter distance before the next loop begins. If $R_R > W$ then the target has a significant but not unity probability of escaping detection. The spiral-in search pattern will maximize the probability of detecting an evading target. On the other hand, if the target does not attempt to evade, then detection will not occur until all of the area has been searched. A maximum search time will be consumed. Each possible search pattern has advantages and disadvantages depending on the behavior of the target.

The physical process of searching an area is usually accomplished by combining two kinds of motion. The swath width is usually generated by scanning a sensor with a relatively narrow field of view in some back and forth fashion. This is then coupled with platform motion in some desired pattern (back-and-forth, spiral-in, spiral-out, etc.) to accomplish the total area search. The scanning to generate the swath width can be accomplished in several ways. A **helical scan** (a simple conical scan coupled with forward platform motion) will produce two roughly parallel bar scans (one leading the motion and one trailing the motion). This is shown in Figure 10-7. The portions of the scan parallel to the platform motion are often discarded as there is considerable re-search which merely adds to the processing burden. This scan is easily implemented even relatively large apertures can be scanned at reasonable high rates, but is very inefficient as no more than 50% (and possible much less) of the scan provides useful area coverage. The rate of the scan must be high enough that the ground-projected distance covered by the field of view of the sensor is greater than the distance traveled by the platform between scan bars.

Back and forth line scans are a more efficient way to generate the swath. The back and forth motion usually takes a simple form. The transverse motion of the sensor field of view may be sinusoidal in time or it may be linear. Simple linear motion results in a **zigzag** or **sawtooth scan** (Fig. 10-7). The sawtooth scan suffers from the fact that if the area is covered completely (no gaps between zigs and zags) then each element of area is viewed twice (the re-search factor – the average number of times each element of search area is searched – is roughly 2). If the re-search factor is reduced, then the coverage must be reduced less than 100%. This can be ameliorated by using a **rectangular scan** (or **bowtie scan)**. In the bowtie scan a longitudinal motion is coupled to the transverse motion. Ignoring platform motion, the angular motion of the field of view will scan a closed pattern that has a bowtie shape (also shown in Figure 10-7). When platform motion is added, the search pattern is transformed into rectangular motion on the ground (thus the two different ways of referring to the same process). This scan is more difficult to implement but has minimum re-search and very little coverage loss.

Sinusoidal scan is relatively easy to produce and can have relative low re-search and relative low coverage loss. The major drawback to sinusoidal scan is a variable dwell time. Dwell time is the effective time that the sensor actually views each element of the search area. The signal-to-noise ratio of many sensors depends strongly on the dwell time. The dwell time is highest at the edges of the swath and shortest at the center of the swath. The variation in sensor performance as a function of scan position is undesirable and may not be acceptable from a detection perspective (detection in this instance is not cookie cutter and requires the more detailed analysis of the form described in Chapter 11).

Figure 10-7. Sensor scan patterns commonly employed in area search.



Volume Search

In volume search the sensor covers a limited angular area (usually but not always out to a fixed range) and attempts to determine if target is present within the conical volume or not. Radars tend to search in range as well as angle; simple imaging systems search only in angle (although a maximum range is often assumed). In every instance, the angular region must be repeatedly searched.

Search of the angular region is equivalent to an area search except that it almost always must be completed in a specified time. The revisit time is usually determined from system considerations. For example, if the search is performed by track-while-scan system (see Chapter 15 for discussion of tracking systems), then the search must be repeated at a rate that is consistent with the tracking accuracy required. This imposition of a specific value on the area search time *T* is a significant complication. If we assume a continuous search with an angular swath width θ, an angular scan velocity ω, and a total angular area Ω to be search, we have by analogy to Eq.(10.24)

$$T = \Omega / \theta \omega . \tag{10.41}$$

Eq.(10.41) is simple enough. The primary complication comes from the need to dwell on each point of the search area for a finite period of time. The performance of every sensor depends on the carrier-to-noise ratio. If the carrier-to-noise ratio goes to zero, the detection performance of a sensor goes to zero. In every instance, carrier-to-noise ratio can be related to the ratio of energy received from the target to noise energy intrinsic to the system. Since no target can radiate or reflect infinite power, then as the observation time goes to zero, the energy received from the target must go to zero. In practice, the carrier-to-noise ratio will go to zero as the observation time goes to zero.

Thus, every sensor must have a minimum **dwell time** τ over which energy for a single observation is collected. In the continuous scan can described above, the dwell time is provided by an angular width ϕ perpendicular to the swath width. The dwell time can be calculated to be

$$\tau = \phi / \omega \tag{10.42}$$

with

$$\omega = \phi / \tau . \tag{10.43}$$

It is the imposition of a maximum value on the scan rate ω that causes the problems. The maximum value for scan rate means that there is a minimum value for the scan time *T*. We have encountered one of those situations that makes systems engineering so interesting. It is desirable to make *T* as small as possible and to make ω as small as possible, but the two are inversely related. The trade between the two has no unique solution and any selection can be debated. There is a limited amount of flexibility available by increasing θ and/or ϕ, but sooner or later, the device designer will object that too much is being asked for.

In imaging systems, the value of ϕ is usually associated with the angular resolution (see Chapter 12 for a discussion of resolution) of the imaging system. The value of θ is usually related to the number of detector elements in the array N and the detector angular subtense θ_D

$$\theta = N\theta_D . \tag{10.44}$$

If the imager has only a single detector element then θ is usually associated with the angular resolution (as is ϕ).

In radar systems, the problem becomes even more complicated. In a radar, each measurement requires at least one radar pulse. In many systems, pulse integration may cause a large number of pulses to be required to make a single measurement. Thus, if the radar has a pulse repetition frequency *PRF* and requires N_I pulses per measurement, then the "dwell time" must be

$$\tau = N_I / PRF \tag{10.45}$$

and the maximum scan rate becomes

$$\omega = \phi / \tau = \phi\, PRF / N_I \quad . \tag{10.46}$$

The angle ϕ is most often associated with the full-angle beamwidth of the radar, although occasionally it is associated with the angular resolution of the radar – the actual selection is usually a subject of a trade concerning the effects of gaps in coverage and the radial dependence of the detection probability.

A second complication in radar systems arises from the requirement that each radar pulse must travel to the target and come back to the radar before the radar can scan to a new direction. This requirement imposes a limit on the *PRF*

$$PRF_{MAX} = c / 2R \tag{10.47}$$

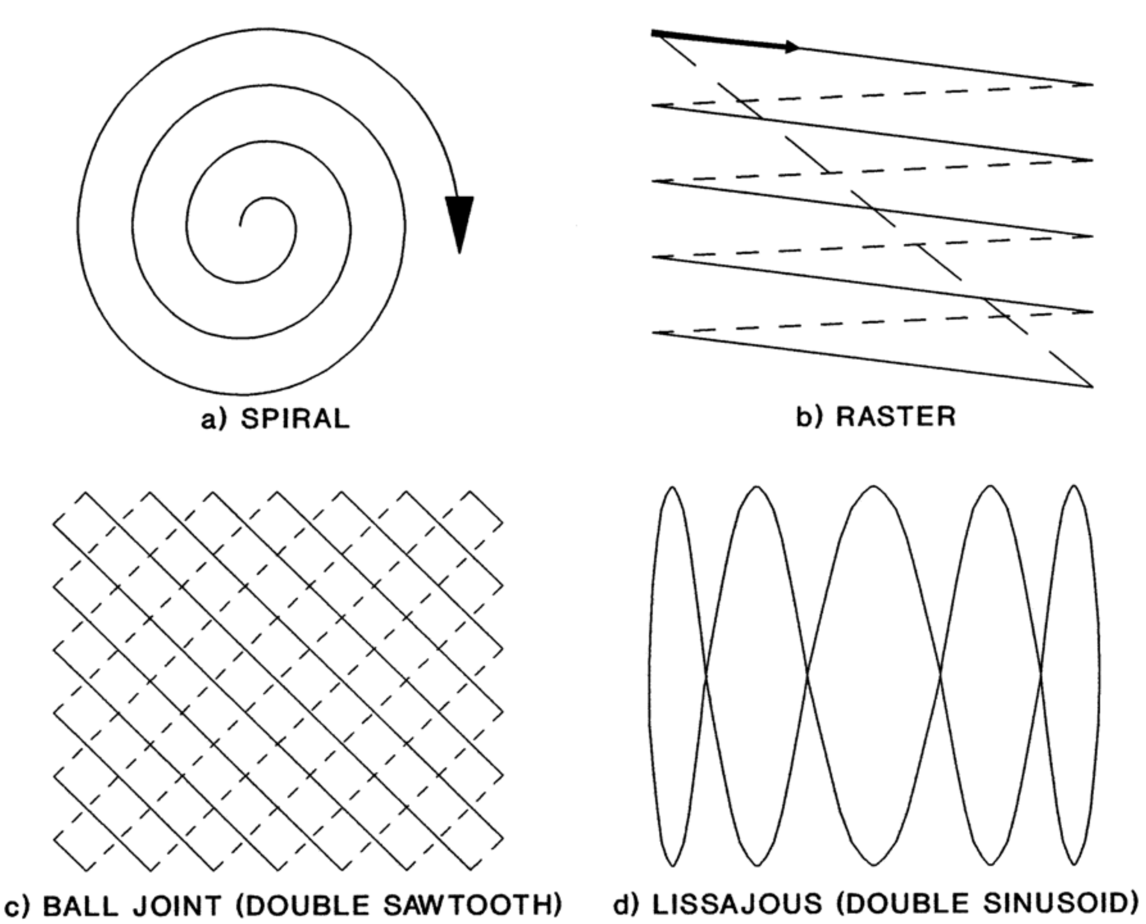
where R is the maximum range to the target (i.e., the range associated with the search volume) and c is the speed of light. With this restriction, the maximum scan rate becomes

$$\omega = \phi\, PRF_{MAX} / N_I = \phi c / 2R\, N_I . \tag{10.48}$$

As before the restriction on scan rate means that search cannot be planned independently of the specific sensor parameters envisioned. Trades will be the name of the game in volume search.

Once scan rates and swath sizes have been determined, the scan pattern remains to be determined. Figure 10-8 shows some of the generic kinds of scan patterns that are used to implement volume searches. The primary distinction between them is how they are implemented. For example, the **spiral scan** is a composite of a simple circular scan (that can be produced by a

Figure 10-8. Commonly used volume search scan patterns.


a) SPIRAL
b) RASTER
c) BALL JOINT (DOUBLE SAWTOOTH)
d) LISSAJOUS (DOUBLE SINUSOID)

rotating antenna optical element) and a slow linear scan. The pattern spirals out and then spirals in. The **raster scan** is produced by combining a fast (horizontal in the figure) linear scan in one dimension and a slow linear scan in the orthogonal direction. Between each element of the linear scan there is a fast "flyback". The **ball joint scan** is a combination of two simple orthogonal sawtooth (constant velocity back and forth) scans with slightly different features. It can be mechanically implemented using a mirror mounted on a rod passing through a ball joint and driven by two orthogonal rotating cams. The **Lissajous scan** is named after the famous patterns observed while testing oscilloscopes. It is formed by combining a vertical sinusoidal scan with a horizontal sinusoidal scan at a slightly different frequency. When both frequencies are integer multiples of a common frequency then a closed form pattern (one example with vertical frequency five times the horizontal frequency is shown in the figure) will result. The raster scan is the most commonly used pattern, but mechanical design concerns have resulted in all of these patterns being used in one real-world system or another.

Incorporating Effects of Non-Unity Detection Probability into Search

The preceding sections have all assumed "cookie cutter" detection. That is, the detection probability is equal to one everywhere within a finite range of the platform, and equal to zero outside that finite range. All calculations are based on a search radius $r = a,$ where a is a characteristic cookie cutter distance), a swath width ($W = 2a$), or a beam width $r = a$ or $d = 2a$). In the real world, the detection probability is never one. As we shall see in the next chapter, the detection probability is usually a strong function of the carrier-to-noise ratio (CNR) of the sensor system. As the CNR decreases, the detection probability drops rapidly.

For example, in the Swerling 2 radar example in the next chapter (Ch.11), the detection probability (for 10^{-12} false alarm probability) versus CNR (in dB) is roughly

CNR(dB)	P_D
27	0.95
24	0.90
21	0.80
19	0.70
16	0.50
14	0.30
12	0.20
10	0.08

In a radar such as this, the CNR is inversely proportional to the fourth power of range. Thus, a 1.5 dB (41%) increase in range will produce an 6 dB decrease in CNR. Thus, if this radar were flying on a sensor at an altitude h that produced a detection probability of 90% in the nadir direction, then as this radar scans from the nadir out to an angle of 45°, the detection probability will fall from 90% to a value of about 65%. Scanning out to an angle of 60° (3 dB increase in range) would cause detection probability to fall below 20%.

This same radar may well have a transmitted radar beam whose transverse intensity profile varies as $sin^2(\theta/\theta_0) / (\theta/\theta_0)^2$ where θ_0 is a characteristic half-angle. The CNR may be assumed to vary as the square of this quantity. The CNR versus angle is

θ/θ_0	*Relative CNR*	*Relative dB*
0	1.000	0
0.5	0.845	-0.7
1.0	0.501	-3.0
1.5	0.196	-7.1
2.0	0.043	-13.7

Assume that the detection probability is 90% at θ_0. At the center of the beam the CNR is 3 dB larger that at θ_0 yielding a detection probability of around 95%. The CNR at $1.5\theta_0$ is 4 dB less than at θ_0 yielding a detection probability that is only about 75%. The CNR at $2.0\theta_0$ is more than 10 dB less than at θ_0 yielding a detection probability slightly less than 30%.

The preceding discussion should immediately raise questions such as "how is the cookie cutter range determined?" and "how is the transverse variation in detection probability accounted for?". These are extraordinarily difficult problems to analyze. However, rough guidelines or heuristics can be established. First, it is necessary to determine what total detection probability is the minimum acceptable. It can never be unity. Typically, values of 0.9 to 0.99 are acceptable. The procedure for determining an acceptable detection probability is given in Chapter 11.

The minimum acceptable value of detection probability is typically assigned to the cookie cutter parameter. With this assignment, everything outside the cookie cutter will have less than the acceptable detection probability, everything inside will have higher detection probability. When used with the search examples from earlier sections in this chapter, this assignment guarantees that the endpoint of the search will be at least the acceptable detection probability. This approach will underestimate the total achievable detection probability because it ignores contributions outside the cookie cutter and slightly reduces (to the value at the edge) contribution from inside the cookie cutter.

The dimension of the swath width is usually sized based on worst case considerations. For example, a radar system will be designed to provide a desired detection probability against the smallest reasonable target at the longest required range under the worst case propagation condition. If any of these conditions becomes worse than the design conditions, then the radar is not expected to "work", i.e., produce the desired detection probability. Given the radar's design conditions and the desired search detection probability (which may not be the same as the radar's design detection probability), it is possible to determine the range at which the search detection probability can be obtained. This is usually used to determine the swath width of a radar sensor. It is important of course to take the geometry into account (such as in the down-looking scan example) and never exceed the radar's detection range.

The choice of an effective sensor beamwidth is also determined using worst case analysis. Under the worst case design conditions, the "size" of the beam that produces the desired search detection probability can be evaluated. This then is assumed as the size of the cookie cutter in the preceding analyses.

References

[1] Washburn, Alan R., Search and Detection 3rd Ed. (Institute for Operations Research and the Management Sciences, Linthicum MD, 1996).

Problems

10-1. An exhaustive search is performed by square tiling of circular detection patches at a constant rate instead of sweeping a swath at constant velocity. The circular patches have radius *a* and are tiled with a center-to-center spacing *2d*. One circular detection patch is covered in each unit τ of elapsed time. Show that the probability of detection as a function of time is approximately

$$P(t) = \eta_{SQ} 4d^2 t / \tau A_T$$

where η_{SQ} is given by Eq.(10.18). How much time does it take to cover the search area? What is the fundamental difference between Eq.(10.25) and the expression above?

10-2. An immobile target is known to be in the vicinity of a reference point with probability density

$$p(r) = \frac{1}{2\pi\sigma^2} e^{-r^2/2\sigma^2}$$

where *r* is the distance from the reference point and σ^2 is the variance. An exhaustive search is performed with swath width *W* and search velocity *V* using a spiral-out search pattern. Assuming that the swath width *W* is small compared to σ, then show that the probability of detection as a function of time is approximately

$$P(t) = 1 - e^{-\gamma t}$$

where

$$\gamma = WV / 2\pi\sigma^2 .$$

What is the mean time required to detect the target given this situation?

10-3. A rectangular volume of 1 degree vertical extent and 2 degree horizontal extent and 15 km range is to be covered by a perfect raster scan. The search sensor has a circular coverage with angular resolution of 0.01 degrees. Derive a generic expression for the maximum frame rate (number of complete searches per second) that this sensor can generate. Evaluate this expression for the specific data provided.

10-4. The search sensor described has a raster scan efficiency η_{SCAN} defined by

$$\eta_{SCAN} = \frac{t_{LineScan}}{t_{LineScan} + t_{Flyback}}$$

where $t_{LineScan}$ is the time required for the scan to sweep from one side of a scan line to the other and $t_{Flyback}$ is the time required for the scan to (more rapidly) sweep back to the original side. How would the expression derived in Problem 10-3 be modified to account for the raster scan efficiency?

CHAPTER 11

DETECTION

What is Detection?

Detection is the set of processes by which a system determines the likely presence of an object of potential interest within the “space” being “observed” by the system. This definition is intentionally vague because there are many different ways implement a detection system using sensors and signatures of all forms. For example, a magnetic influence mine detects a ship target when the nearby presence of a ship causes the magnetic field in the vicinity of the mine to change by more than a predetermined amount. A sonar operator detects a submarine when the sound from the submarine becomes loud enough relative to background noise for the operator to perceive the significant characteristic differences. A radar detects a target when the amplitude of the backscattered electromagnetic radiation produces an electrical signal that exceeds a predetermined threshold voltage. A sniper detects the enemy trooper, when he can distinguish enough visual characteristics to discriminate the enemy from the background. In every instance, there is a signal which is characteristic of “target” and a signal characteristic of “background” or “noise”. Detection occurs when an individual or an electronic system can sufficiently discriminate “target” from “background” or “noise” to declare the target to be present with sufficient certainty that the system should take action (detonate the mine, establish a track on the target, or pull the trigger).

We will differentiate between two different forms of detection, because their applications and analysis seldom overlap. The first of these is the detection of the existence of a basically one-dimensional signal (amplitude versus time) in the presence of noise from any source. This problem is the basis for the field of detection theory. We will investigate this type of detection in considerable more detail in this chapter. It finds use in active sensors and passive sensors, electromagnetic sensors operating from the radio frequency to the ultraviolet, acoustic sensors, and sensors operating on esoteric observables such as field anomalies. As we shall see, a common formalism exists which permits accurate modeling of signal detection phenomena regardless of sensor type.

The second kind of detection is that based on human visual observation of an image. Image signal characteristics other than amplitude play a part in human perception. For example, contrast (brightness difference between target and background) is more important than total brightness. Shape and surface texture may permit detection when contrast is extremely poor. The process of image-based detection is complicated by the presence of clutter (real but uninteresting objects), finite resolutions of the images, and motions of targets, backgrounds, or both. Image-based detection is part of a larger psychophysical problem of image understanding. Because image-based detection differs so strongly from signal-based detection, we will cover it in a separate chapter in this volume (Chapter 14).

Basics of Detection Theory

In the following few pages we will investigate the detection of a radar signal in the presence of radar receiver noise.[1] A radar was chosen to illustrate this process for several reasons. First, a concrete example is more satisfying than a purely hypothetical one. Second, deriving all of the expressions for a radar at this time will spare us from having to do so when the chapters on radars are reached. Thirdly, the rapid pace of development of radar in World War II let hardware design exceed the limits of existing theory. As part of an attempt to keep pace, the field of detection theory was invented and almost universally applied first to radar systems. The process described here closely mirrors the historical development of the field using the examples that drove that development.

The fundamental question facing any signal-detection system is whether or not a target is present at any given time. If we simplify the problem to its most basic form, we find there are two hypotheses:

Hypothesis 0 (H_0):	Target Absent
Hypothesis 1 (H_1):	Target Present

There are no other possibilities. A target cannot be both present and absent. If multiple targets are present, then Hypothesis 1 is true by default. We cannot ask if a particular kind of target is present. This goes beyond mere detection. Thus, a radar detection system is tasked to decide which hypothesis, H_0 or H_1 is true based on processing of the radar return.

The Neyman-Pearson approach to binary hypothesis testing is to maximize the **detection probability**

– **P_D is the probability that H_1 is chosen, given that H_1 is true** –

subject to the constraint of keeping the **false alarm probability**

– **P_F is the probability that H_1 is chosen, given that H_0 is true** –

below some preselected value ε. Neyman-Pearson is not the only approach that could have been taken. For example, (a) one could minimize P_F subject to the constraint of keeping P_D larger than some value ζ or (b) one could attempt to maximize P_D subject to the constraint of keeping the ratio P_F / P_D less than some value η. However, the first alternate approach put less importance on the basic detection function and tend to give less than optimum detection performance. The second alternate approach put less emphasis on false alarms; although detection probability may be enhanced, it is more difficult to control false alarms. In practice, the Neyman-Pearson approach has been deemed to be optimum for detection problems.

The optimum Neyman-Pearson processor for a continuous variable V (such as the voltage output from a radar video detector) is a likelihood ratio (threshold) test:

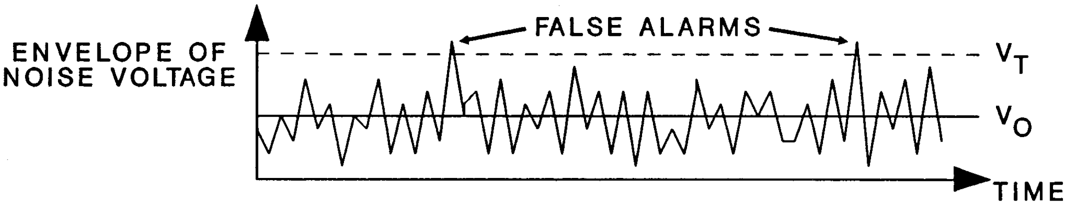
$$V \quad \begin{array}{l} \geq \\ < \end{array} \quad V_T \qquad \begin{array}{ll} \text{Choose } H_1, & (11.1a) \\ \text{Choose } H_0, & (11.1b) \end{array}$$

where V_T is chosen to yield $P_F = \varepsilon$.

What does this test mean for detection? If the signal-plus-noise voltage (i.e., the total output of the detector) exceeds the threshold voltage, the system will declare a target to be present – whether or not a target is really present. If the target is really present, we have a valid detection. If the target is actually absent, we have a false alarm. If the signal-plus-noise voltage fails to exceed the threshold, the system declares that no target is present – whether or not the target is really absent. The question should be asked as to why one can have a threshold crossing in the absence of a target. The answer lies in the character of the system noise.

Let us consider the case where no target is present. The output of the receiver is due entirely to system noise. Such noise is generated internally in every electrical system. There is no way to completely eliminate system noise. One of the characteristics of noise is that it tends fluctuate in amplitude about an average value, as shown in Figure 11-1. However, occasionally, the amplitude will fall well below average, and on other occasions, the amplitude may spike well above average. The higher the amplitude of the spike, the less likely it is to occur. Nevertheless, there is a non-zero probability that a spike of any magnitude can be produced.

Figure 11-1. Typical behavior of system noise. V_0 is the rms noise level; V_T is a threshold.



The fraction of any long time interval that the noise voltage lies above the threshold voltage can be equated to the false alarm probability. The higher the threshold, the smaller the fraction of time that it will be exceeded, and the lower the false alarm probability. For classical radar receivers the system noise has been characterized by the following probability density function

$$p_N(V) = \left(V / V_0^2\right) e^{-V^2/2V_0^2} \tag{11.2}$$

Equation (11.2) results when Gaussian noise is narrow-band filtered. The rms noise V_0 is determined by the system temperature and the filter bandwidth among other things.

By way of a quick review, the probability density function is that function of a random variable whose definite integral over an interval gives the probability that the variable lies within that interval, i.e.

$$\int_a^b dx \; p_x(x) = \text{Probability that a} \geq \text{x} \geq \text{b} . \tag{11.3}$$

Thus we would expect that the false alarm probability can be directly determined by definite integration of the noise probability density function for the interval lying above the threshold. Thus, we find

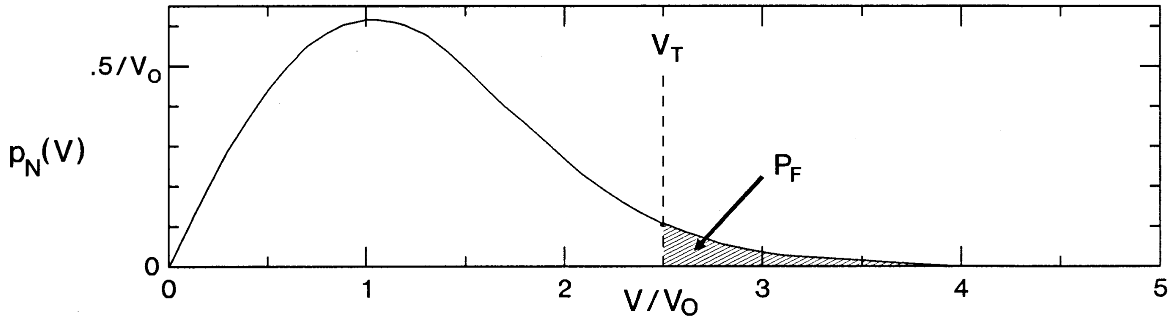
$$P_F = \int_{V_T}^{\infty} dV \; p_N(V) \; . \tag{11.4}$$

Equation (11.4) is a general expression. It will be valid for any and every sensor system provided we know the form of the appropriate $p_N(V)$ function. For the case of radar noise, the integration gives a straightforward closed-form result (this does not happen for every possible noise density)

$$P_F = e^{-V_T^2/2V_0^2} \; . \tag{11.5}$$

These results are shown graphically in Figure 11-2. The total area under the curve $p_N(V)$ is unity. The shaded area (the area lying above the threshold voltage) is equal to the false alarm probability.

Figure 11-2. Graphical representation of false alarm probability for a radar.



If the false alarm probability of a system is very low, then a considerable amount of time may pass between false alarms. In more than a few systems, it is convenient to specify an average time between false alarms, rather than specifying a false alarm probability. For example, requiring that a radar should have less than one false alarm per hour is more meaningful than requiring that radar to have a false alarm probability less than 10^{-11}, even though the two requirements may be equivalent. If we define T_F to be the mean time between false alarms and we define B to be the bandwidth of the radar that determines its electrical noise characteristics (usually the bandwidth of a filter in the IF section of the receiver), then the equivalent false alarm probability can be determined from the relation

$$P_F = 1 / T_F B \tag{11.6}$$

For example, let us assume that a radar is desired to have a mean time between false alarms of one hour and that the radar has an IF limiting bandwidth of 10 MHz, then

P_F = 1/[(3600 seconds) x (10,000,000 Hz)] = 2.777 x 10^{-11}.

If a target is present, then the voltage produced by its signal adds to the system noise voltage. The probability density function is altered. If the target produces a constant signal of voltage *A*, and the radar possesses narrowband Gaussian noise, the probability density for signal-plus-noise becomes

$$p_{S+N}(V) = \left(V / V_0^2\right) e^{-\left(V^2+A^2\right)/2V_0^2} I_0\left(VA / V_0^2\right) \tag{11.7}$$

where $I_0(x)$ is a Bessel function of purely imaginary argument. Equation (11.7) is only valid when the target is present. Thus, the integral of the probability density above threshold should be equal to the probability of detection

$$P_D = \int_{V_T}^{\infty} dV \; p_{S+N}(V) \tag{11.8}$$

Equation (11.8) is valid for any sensor system if the correct function (not necessarily Eq. (11.7)) for $p_{S+N}(V)$ is used. Unfortunately, the integral of Eq. (11.8) using the function in Eq. (11.7), cannot be evaluated in closed form. The integral function defined in Equation (11.9) is called Marcum's Q-function. It is relatively easily evaluated numerically.

$$Q_M(A,V_0;V_T) = \int_{V_T}^{\infty} dV \; \left(V / V_0^2\right) e^{-\left(V^2+A^2\right)/2V_0^2} I_0\left(VA / V_0^2\right) \tag{11.9}$$

Figure 11-3 graphically illustrates the probability of detection and compares it to the false alarm probability.

Figure 11-3. Graphical representation of detection probability of a radar.

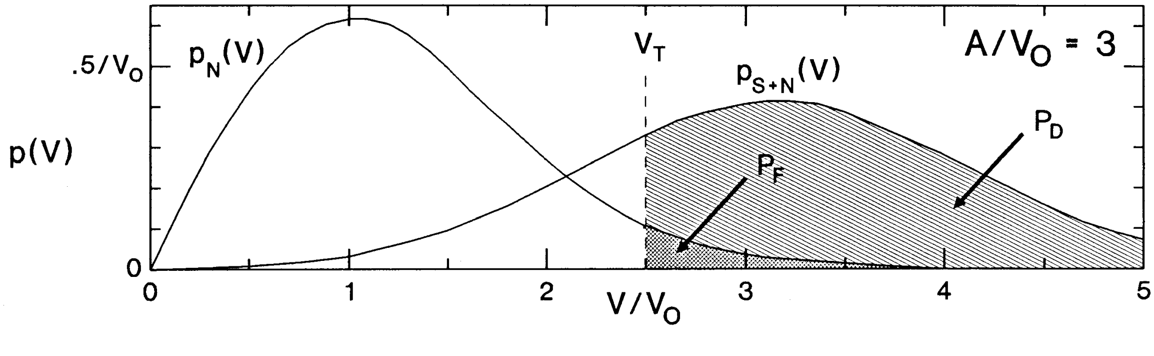


Figure 11-3 gives us the first hint of the contradictory (or perhaps complementary) natures of P_D and P_F. Detection probability can be increased (which is good) by lowering the threshold. However, lowering the threshold means increasing the false alarm probability (which is bad). Before addressing this point further, it is important to introduce some new terminology. One term is familiar to everyone, but understood by few; the other term is unfamiliar to most people.

The output of the receiver is a complex function $\underline{r}$ which has two contributors: a signal component $\underline{y}$ and a receiver noise component $\underline{n}$, [2]

$$\underline{r} = \underline{y} + \underline{n} \tag{11.10}$$

The signal-to-noise ratio *SNR* is a quantity that tells us how much information a signal can convey. It is defined by the following expression

$$SNR = \frac{\left\langle \left|\underline{y}\right|^2 \right\rangle^2}{\mathrm{var}\left(\left|\underline{r}\right|^2\right)} = \frac{\text{(Mean Square Signal Power)}}{\text{(Mean Square Fluctuation Level)}} \tag{11.11}$$

Everyone thinks they know what signal-to-noise ratio means. However, most authors define the *SNR* as the signal power divided by the noise power. The information theory definition recognizes that random fluctuations in signal level can occur in addition to random fluctuations in noise level. Both kinds of fluctuations limit the ability to extract information. The term that most authors call signal-to-noise ratio is actually the carrier-to-noise ratio defined below

$$CNR = \frac{\left\langle \left|\underline{y}\right|^2 \right\rangle}{\left\langle \left|\underline{n}\right|^2 \right\rangle} = \frac{\text{(Mean Signal Power)}}{\text{(Mean Noise Power)}} \tag{11.12}$$

It is the *CNR* that is calculated by the radar equation, not the *SNR* (although few radar texts use this more modern terminology). We are attempting to give the system engineer a new tool that he can use to analyze every kind of sensor from a common perspective. In other sections of this work we will need to distinguish between the two quantities. For that reason we introduce them both at this time. In almost every instance that the reader will find in outside sources, he should feel free to substitute *CNR* for *SNR* and they would be consistent with the present work. The author has made every attempt to be 100% consistent in his use of CNR vs SNR, however, there may be occasional mistakes in not altering expressions obtained from others.

For radar systems it is possible to model the noise $\underline{n}$ as a zero-mean, complex, Gaussian random variable with unit variance.[2] Using this normalization we have

$$\left\langle \left|\underline{r}\right|^2 \right\rangle = \left\langle \left|\underline{y}\right|^2 \right\rangle - 1 \tag{11.13}$$

and

$$\operatorname{var}\left(\left|\underline{r}\right|^2\right) = \operatorname{var}\left(\left|\underline{y}\right|^2\right) + 2\left\langle \left|\underline{y}\right|^2 \right\rangle + 1 \tag{11.14}$$

for the normalized mean and variance of the output of the radar matched filter detector. The factor of unity in Eq. (11.14) corresponds to a target-independent, constant term due to receiver noise. Combining Eqs. (11.11) and (11.14) yields

$$SNR = \frac{\left(\left\langle \left|\underline{y}\right|^2 \right\rangle\right)^2}{\operatorname{var}\left(\left|\underline{y}\right|^2\right) + 2\left\langle \left|\underline{y}\right|^2 \right\rangle + 1} \tag{11.15}$$

Substitution of the definition of CNR into this expression yields

$$SNR = \frac{CNR/2}{1 + CNR\,\dfrac{\operatorname{var}\left(\left|\underline{y}\right|^2\right)}{2\left\langle \left|\underline{y}\right|^2 \right\rangle} + \left(2CNR\right)^{-1}} \tag{11.16}$$

If we consider only the situation where *CNR* ≥ *5*, it is possible to neglect the (*2CNR*)$^{-1}$ term in the denominator of Eq. (11.16). If we further define a quantity

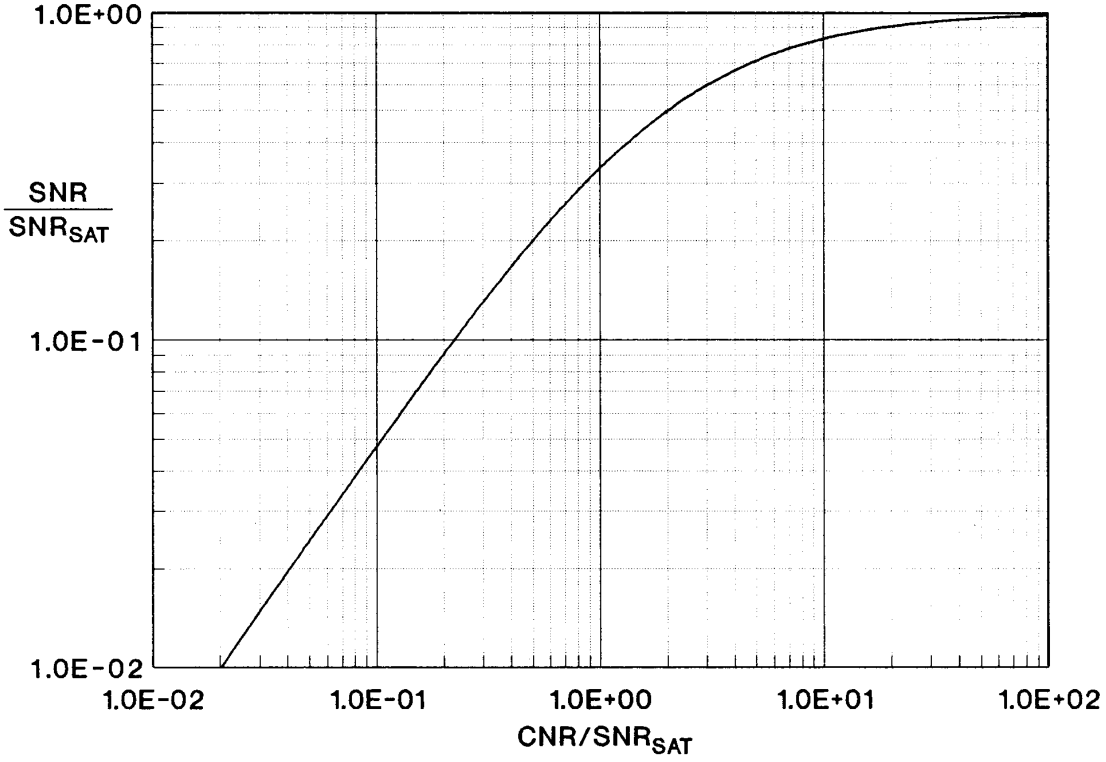
$$SNR_{SAT} = \frac{\left\langle \left|\underline{y}\right|^2 \right\rangle^2}{\operatorname{var}\left(\left|\underline{y}\right|^2\right)} \tag{11.17}$$

Eq. (11.16) can be written in a universal form

$$\frac{SNR}{SNR_{SAT}} \approx \frac{\dfrac{CNR}{2SNR_{SAT}}}{1 + \dfrac{CNR}{2SNR_{SAT}}} \tag{11.18}$$

This is plotted in Figure 11-4. At very large *CNR*, *SNR* asymptotically approaches SNR_{SAT}. At smaller values of *CNR*, *SNR* is approximately *CNR/2*. Thus we see that SNR_{SAT} is the largest *SNR* that can be achieved for any *CNR*. It should be noted that for Swerling II cross section statistics (see section on target fluctuations) $SNR_{SAT} = 1$.

Figure 11-4. Universal curve for normalized signal-to-noise ratio versus normalized carrier-to-noise ratio.



In our radar example above, the carrier-to-noise ratio can be easily calculated to be

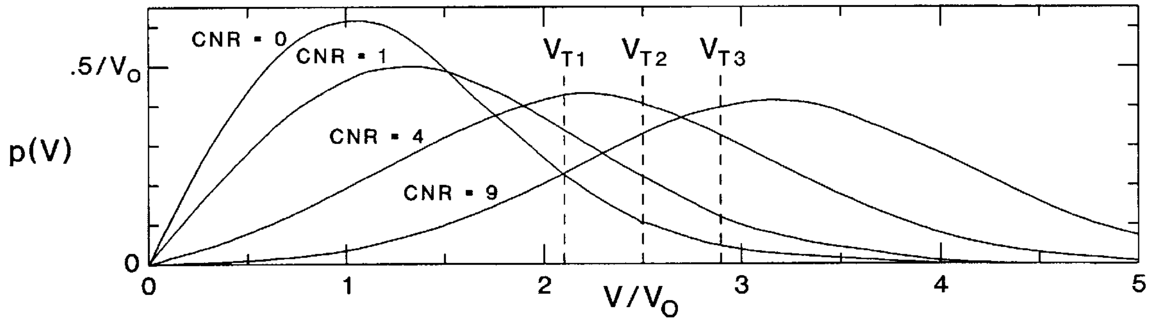
$$CNR = A^2 / V_0^2 . \tag{11.19}$$

The signal-to-noise ratio is given by

$$SNR = A^2 / 2V_0^2 = CNR / 2 . \tag{11.20}$$

A thorough examination of Equations (11.5) and (11.9) indicates a complex interrelationship between detection probability, false alarm probability, and carrier-to-noise ratio. Consider the functions plotted in Figure 11-5. Note: the curve for CNR=0 is the same as the curve for noise alone.

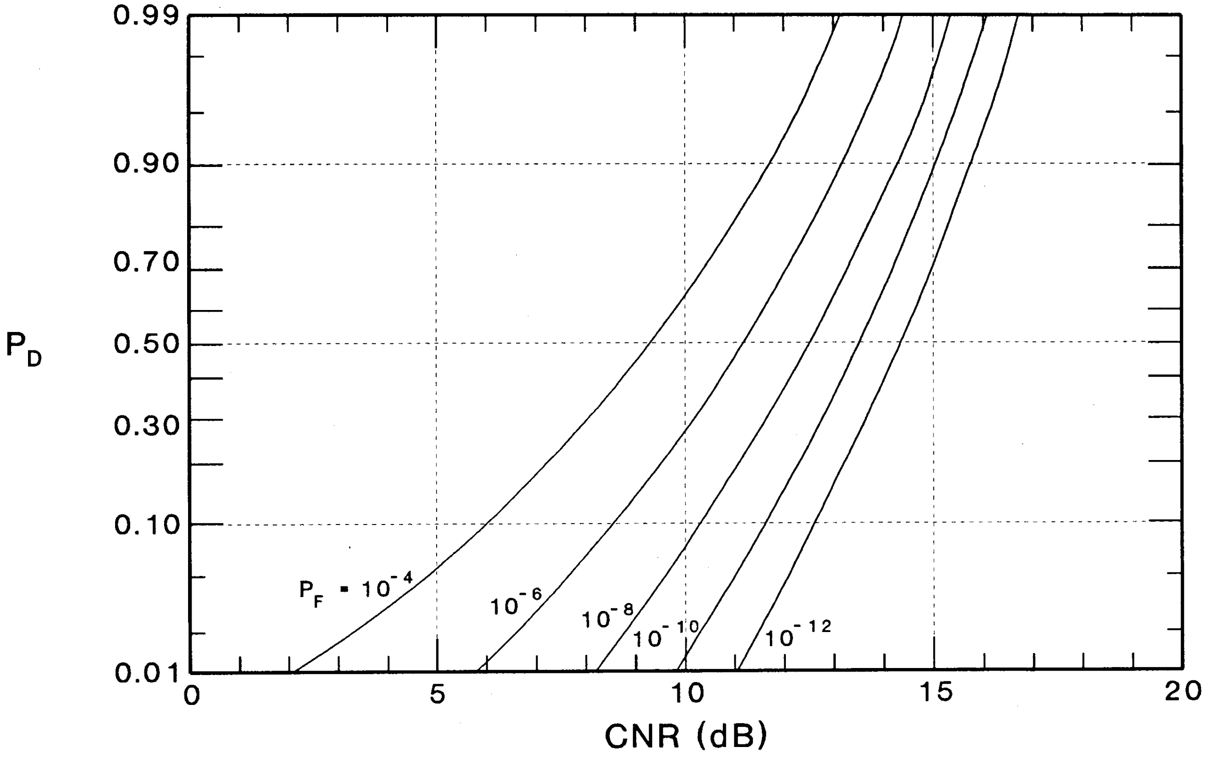
Figure 11-5. Graphical representation of the effects of changing CNR on $p_{S+N}(V)$.



As CNR increases, the peak of the probability density function shifts to higher and higher values. Consequently, for any fixed threshold, the probability of detection increases with increasing CNR. The false alarm probability is unaffected by CNR changes. As the threshold is increased (for fixed CNR), the false alarm probability decreases, but the detection probability also decreases. Conversely, lowering the threshold to increase the probability of detection produces an accompanying increase in the false alarm probability. To obtain an increase in probability of detection and a decrease in false alarm probability, it is necessary to increase the threshold while simultaneously increasing the CNR even faster.

These complex interrelationships between P_D, P_F, and CNR are summarized in the receiver operating characteristic (or ROC). The ROC may take different forms. However, one of the more useful forms is a plot of P_D vs. CNR, parameterized in P_F. For the radar example we have been following, the ROC is shown in Figure 11-6. The ROC is a key tool in the sensor engineer's toolbox. It is used in the following fashion. From higher level system specifications, the designer will have been given acceptable levels of false alarms and required detection probabilities. Given P_D and P_F (or T_F), he then finds the appropriate curve on the chart and determines what value of CNR is required to yield the required P_D & P_F performance. The job of the sensor designer then becomes the design of the sensor to yield that required CNR or better under all operational conditions of interest. For the ROC shown in the figure, if $P_D \geq 0.90$ and $P_F \leq 10^{-8}$ are required to achieve acceptable system performance, then a CNR ≥ 14.2 dB is required. As indicated in the chapter in radiometry and as will be explicitly derived in the chapter on radar, a "radar range equation" can be defined which relates CNR to explicit radar parameters such as power, antenna size, and frequency and to operational parameters such as target range and target radar cross section. Given the required CNR, the radar designer can use the range equation to perform trade studies to determine the optimum choice of radar parameters.

Figure 11-6. Receiver operating characteristic for a radar with a non-fluctuating target.



Impact of Signal Fluctuations on Detection

The ROC generated in the preceding section explicitly assumed a target with a constant signal level *A*. Unfortunately, the returns from real target are seldom constant in amplitude. For a variety of reasons, the target returns usually fluctuate violently, even if the radar illumination is constant. Possible significant contributors to signal level fluctuations include:

- Misalignment of the radar beam with respect to the target (due to beam scanning, pointing jitter, target dynamics, or tracking errors)
- Changes in target orientation
- Constructive and destructive interference between reflections from different parts of the targets
- Multipath interference (between direct and indirect paths between the radar and the target)
- Temporal and/or spatial variations in atmospheric absorption
- Temporal and/or spatial variations in atmospheric scattering (e.g., from dust, clouds, smoke, or fog)
- Presence of precipitation (rain, snow, hail, etc.)
- Atmospheric turbulence
- Ionospheric Scintillation
- Jamming

In general, most fluctuations can be classified as being either target-induced or propagation channel-induced.

The presence of signal fluctuations significantly alters the probability density of the signal-plus-noise. In turn, the receiver operating characteristic will be correspondingly altered. In many instances, the fluctuation statistics can be adequately modeled or measured. If this is possible, then modified receiver operating characteristics can be constructed.

If we consider target-induced fluctuations, then any measured signal strength (*CNR*) can be related to an effective radar cross section (σ). In turn any signal voltage level *A* can be related to a square root of that effective cross section. Fluctuations in *A* can be considered to be equivalent to fluctuations in σ. That is,

$$CNR \propto \sigma \qquad \Rightarrow \qquad A \propto \sigma^{1/2} . \tag{11.21}$$

For convenience let us rewrite Marcum's Q function (the detection probability for non-fluctuating signals – Eq. (10.9)) in the following fashion where we have highlighted the implicit dependence on cross section ($A = \alpha\sigma^{1/2}$)

$$Q_M\left(\sigma, V_0;\ V_T\right) \equiv \int_{V_T}^{\infty} dV\ p_{S+N}\left(V;\alpha\sigma^{1/2},V_T\right) . \tag{11.22}$$

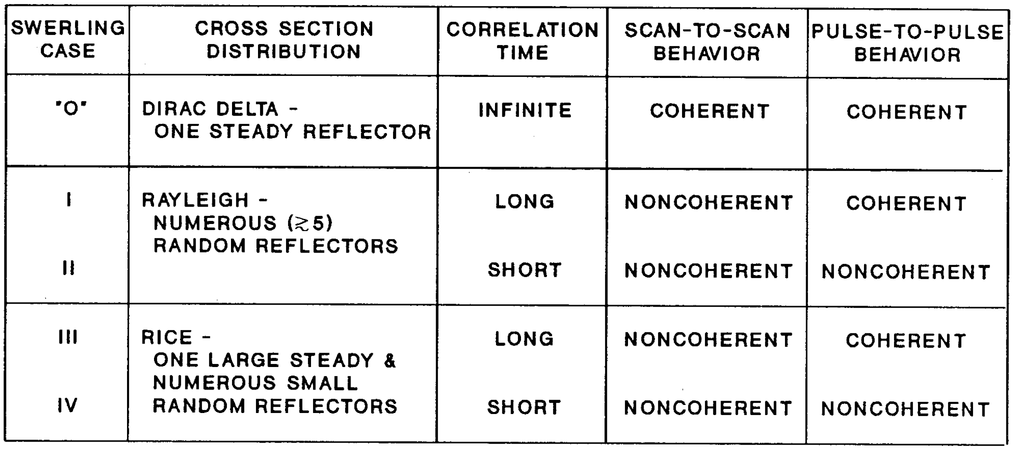
If the probability of seeing any given value of effective cross section is $p(\sigma)$, then if we average Marcum's Q function over this distribution of possible cross section values we should obtain the detection probability averaged over the effective cross section fluctuations

$$P_D = \int_0^\infty d\sigma \; p(\sigma) \, Q_M(\sigma, V_0; V_T) . \tag{11.23}$$

The trick is to obtain the statistics for the cross section fluctuations.

Peter Swerling did just that for several model radar targets.[3] These models are summarized in Table 11-1. His first case consisted of a large number of small reflectors of roughly equal size, spaced randomly over a volume large compared to a wavelength of the radar being considered. Interference will occur between each reflector and every other reflector. The net result of all the constructive and destructive interferences was an effective cross section that was Rayleigh distributed.

Table 11-1. Swerling's models of radar cross section fluctuations.[3]



SWERLING CASE	CROSS SECTION DISTRIBUTION	CORRELATION TIME	SCAN-TO-SCAN BEHAVIOR	PULSE-TO-PULSE BEHAVIOR
"O"	DIRAC DELTA - ONE STEADY REFLECTOR	INFINITE	COHERENT	COHERENT
I	RAYLEIGH - NUMEROUS (≳5) RANDOM REFLECTORS	LONG	NONCOHERENT	COHERENT
II		SHORT	NONCOHERENT	NONCOHERENT
III	RICE - ONE LARGE STEADY & NUMEROUS SMALL RANDOM REFLECTORS	LONG	NONCOHERENT	COHERENT
IV		SHORT	NONCOHERENT	NONCOHERENT

That is, the cross section has the following probability distribution

$$p_{RAYLEIGH}(\sigma) = \frac{1}{\sigma_{AV}} e^{-\sigma / \sigma_{AV}} \tag{11.24}$$

where σ_{AV} is the average (mean) value observed for the effective cross section. In Swerling's later calculations he also assumed different correlation times for the cross section. The correlation time is essentially the time over which the cross section can be assumed to be constant. If one value of cross section is observed at time T, then if a second observation is made a time τ later, the two values will be the same if the time difference τ is small compared to the correlation time. If τ is large compared to the correlation time, then a different value of cross section, randomly selected from the cross section distribution will be observed. Swerling had two time scales of interest: the time between successive scans of the radar (usually seconds) and the time between successive pulses from the radar (usually fractions of a millisecond). For computational purposes, he assumed one collection of random scatterers had a long correlation time (larger than milliseconds but smaller than seconds). Such a model would exhibit a difference cross section every scan of the radar, but would have a constant cross section pulse to pulse within a scan. He called this Case I. He also assumed a second collection of random scatterers would have a short correlation time (smaller than a fraction of a millisecond). Such a model would exhibit a different cross section for every radar pulse. He called the short correlation time model Case II.

Swerling put forward a second model which was composed of one large steady (non-fluctuating) scatterer plus numerous small random reflectors whose integrated return was comparable to that of the large single scatterer. This model produces a Rice distribution for the cross section

$$p_{RICE}\left(\sigma\right) = \frac{4\sigma}{\sigma_{AV}^{2}} e^{-2\sigma/\sigma_{AV}} . \tag{11.25}$$

Once again he proposed two correlation time behaviors and called the respective instances Case III and Case IV. For completeness, this author has called the non-fluctuating target as Swerling Case “0", although this was not Swerling’s usage. The cross section distribution for a non-fluctuating target is a Dirac delta function

$$p_{DIRAC}\left(\sigma\right) = \delta\left(\sigma_{AV}\right) . \tag{11.26}$$

Swerling took his models and basically performed the averaging described above. What he found was that fluctuating targets required higher *CNR* to detect than did non-fluctuating targets. The higher the desired detection probability, the larger the increase in *CNR* required to yield that detection probability. However, the results were virtually independent of false alarm probability. This is summarized graphically in Figure 11-7. Figure 11-7 is used in the following fashion. After determining (from Figure 11-6) the *CNR* required to yield desired detection and false alarm probabilities for a non-fluctuating target, the additional *CNR* required to detect a fluctuating target is determined for the same detection probability from the curve for the appropriate Swerling Case. For example, if $P_D = 0.9$ and $P_F = 10^{-8}$ is desired for Swerling II, then from Figure 11-6 we require *CNR* = 14.2 dB for the non-fluctuating case **plus** 8 dB from Figure 11-7 to accommodate the Swerling II fluctuations for a total of 22.2 dB. If the radar can provide 22.2 dB *CNR* then the desired detection and false alarm probabilities will be achieved.

It often transpires that Swerling II is a good worst case model for fluctuations of real targets. For this reason it is useful to have an explicit receiver operating characteristic for this case. One can be generated in either of two ways. First, the values of Δ*CNR*, the additional *CNR* required to accommodate the Swerling II fluctuations, can be extracted from Figure 11-7 and added to each curve in Figure 11-6. The second way is to explicitly evaluate Equations (11.23) and (11.24). For the case of Rayleigh cross section statistics only, the equations can be solved to yield the simple closed form solution

$$P_D = P_F^{\left(1/(1+CNR)\right)} . \tag{11.27}$$

Either way, the curves of Figure 11-8 result. This is the receiver operating characteristic for Swerling II radar cross section statistics.

Figure 11-7. Additional CNR required to accommodate target fluctuations.

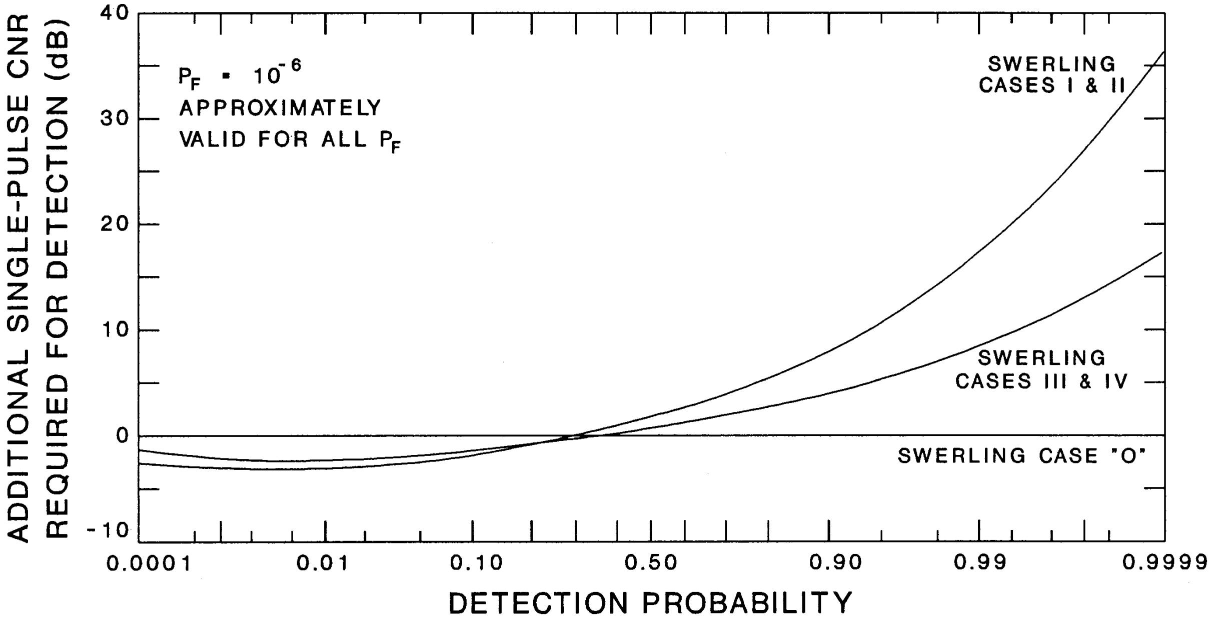
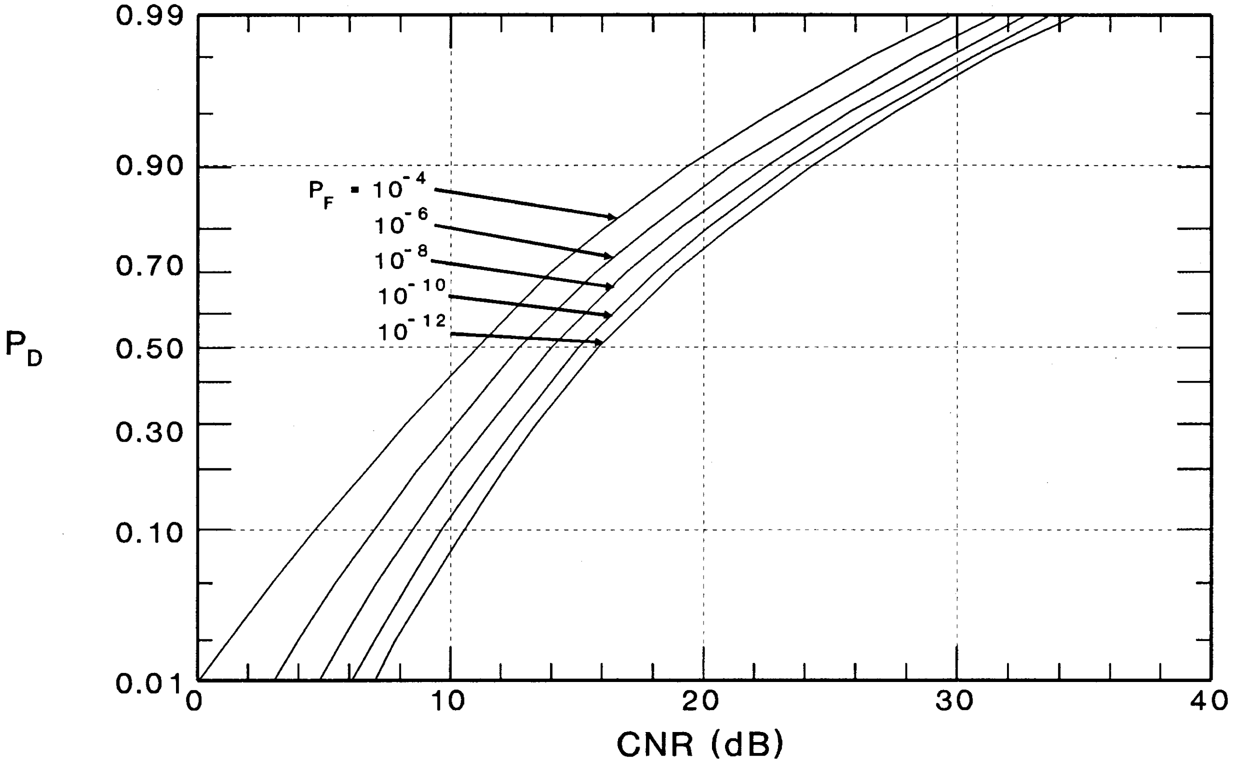


Figure 11-8. Receiver operating characteristic for Swerling II statistics.



Impact of Integration on Detection

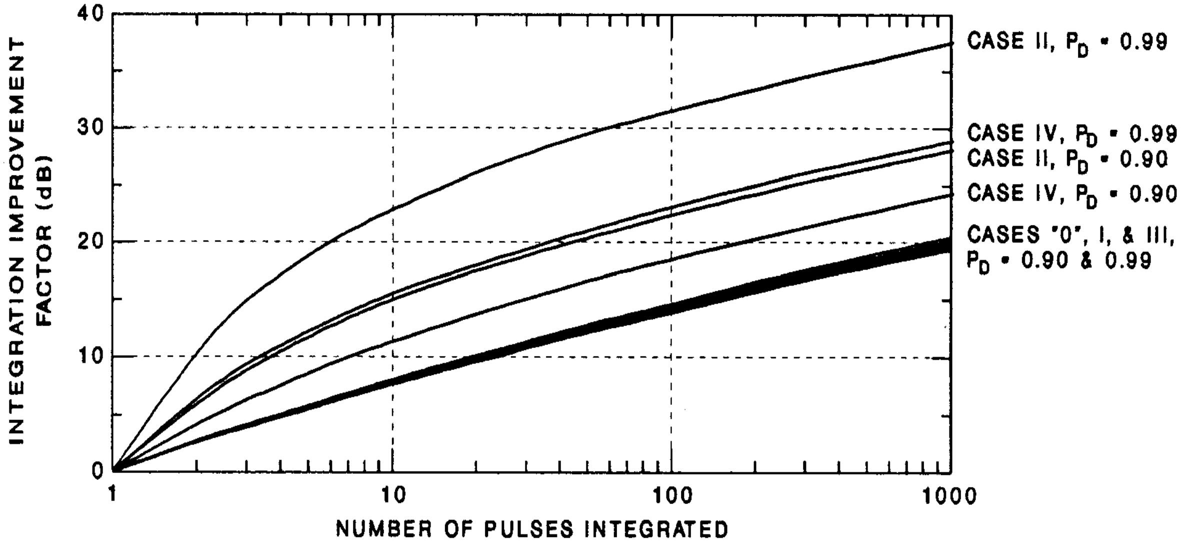
The presence of fluctuation causes significant increases in the required carrier-to-noise ratio to make a detection. The additional 10-20 dB required may prove to be the difference between a practical system and an impractical system. As we will see, increased CNR can only be obtained at the expense of increased transmitter power, increased aperture size, or decreased useful range. Few systems can afford any of these changes. Therefore, it is of critical importance to investigate approaches that may alleviate this potential problem. Pulse integration is arguably the most important.

When the signal fluctuations result from interference processes such as in the Swerling models, the phases of each target return will be random from return to return. The optimum processing strategy is to non-coherently integrate N pulses. Specifically, the returns from N pulses are added and normalized by the number of pulses, and then compared to a threshold γ_N that is optimized for the N-pulse-integrated return.

$$\frac{1}{N}\sum_{m=1}^{N}\left|r_m\right|^2 \begin{array}{ll} \geq \gamma_N & \text{Choose } H_1 \\ < \gamma_N & \text{Choose } H_0 \end{array} \qquad (11.28)$$

There is a computationally intensive process involving moment generating functions that permits evaluation of detection and false alarm probabilities in pulse-integrated systems. The mathematics is beyond the scope of this text. However, a useful summary of results is shown in Figure 11-9.

Figure 11-9. Integration improvement factors for N-pulse noncoherent integration.



The integration improvement factor is the amount by which the single-pulse CNR can be reduced if N-pulse optimum integration is used. In our earlier example with Swerling II statistics, $P_D = 0.9$, and $P_F = 10^{-8}$, a single pulse CNR of roughly 22 dB was required. From Figure 11-9 we observe that for $P_D = 0.9$ and Swerling II, then the integration improvement factor for N=10 pulses is 15 dB. This means that if ten pulses, each yielding a CNR of only 7 dB (= 22 dB - 15 dB), are integrated then the desired P_D and P_F will be obtained.

It is interesting to note that the integration improvement factor can yield a numerical improvement that is larger than the numerical value of the number of pulse that are integrated. This is surprising because in statistics it is often true that standard deviations decrease by the inverse square root of the number of samples obtained (the variances decrease by one over the number of samples). However, consider the exact nature of the integration process. N samples of the signal are averaged and N samples of the noise are averaged. The averaged signal divided by the averaged noise might produce a factor of 1/N. However, because the noise is reduced, the threshold can be reduced while maintaining the same false alarm probability. If noise fluctuations are decreased, the threshold must be reduced if the same rate of exceedance is to be produced. However, reducing the threshold to maintain the desired false alarm probability has the beneficial effect of increasing the detection probability. If a fixed detection probability is acceptable then the CNR can be further decreased. It is the combination of averaging out signal fluctuations, averaging out noise fluctuations, and reducing the threshold, that gives integration improvement factors that may be greater than N.

Pulse integration is not free of costs or penalties. The integrator requires considerable memory and sophisticated arithmetic computation units and is considerably more complicated than a simple threshold comparator. Thus dollar costs will increase. Each measurement consumes the time required to gather N pulses. The time required to search a specified angular field of view will increase proportional to the number of pulses integrated. In summary, integration increases cost and decreases search rate in exchange for permitting reduced power and aperture and/or increased range.

As mentioned above, optimum pulse integration requires a complicated processor. There are two sub-optimal integration approaches that have considerably simpler processors. One-out-of-N pulse detection involves transmitting N pulses and declaring a detection if any of the N pulses exceeds the single-pulse threshold. If $P_F(1)$ is the single-pulse false alarm probability and $P_D(1)$ is the single-pulse detection probability, then the probability of having 1 false alarm in N tries $P_F(N)$ is given by

$$P_F(N) = 1 - \left[1 - P_F(1)\right]^N \tag{11.29}$$

while the probability of having 1 detection in N tries $P_D(N)$ is given by

$$P_D(N) = 1 - \left[1 - P_D(1)\right]^N . \tag{11.30}$$

If $P_F(N)$ and $P_D(N)$ are specified, the values of $P_F(1)$ and $P_D(1)$ required to produce them can be determined from

$$P_F(1) = 1 - \left[1 - P_F(N)\right]^{1/N} \tag{11.31}$$

and

$$P_D(1) = 1 - \left[1 - P_D(N)\right]^{1/N} . \tag{11.32}$$

For Swerling II targets, we know that

$$P_D(1) = P_F(1)^{(1+CNR)^{-1}} \tag{11.33}$$

Thus, we can calculate the CNR required to yield specific $P_D(N)$ and $P_F(N)$ from the relation

$$CNR_{REQUIRED} = \frac{\ln\left[1 - \left[1 - P_F(N)\right]^{1/N}\right]}{\ln\left[1 - \left[1 - P_D(N)\right]^{1/N}\right]} - 1 . \tag{11.34}$$

One-out-of-N pulse detection is simple but does not produce anywhere near optimum performance. A much improved approach that retains most of the simplicity is M-out-of-N pulse detection. In this approach, N pulses are transmitted and the number of threshold crossings is counted. If the number of threshold crossings exceeds a predetermined number, then a detection is declared. If fewer threshold crossings are detected, then no declaration is made; any threshold crossings are assumed to have been due to noise.

If we use the same definitions for $P_D(1)$ and $P_F(1)$ as before, then the probabilities for M-out-of-N detection can be determined from simple probability theory

$$P_D(M \text{ of } N) = \sum_{m=M}^{N} \frac{N!}{m!\,(N-m)!} \left[P_D(1)\right]^m \left[1 - P_D(1)\right]^{N-m} \tag{11.35}$$

$$P_F(M \text{ of } N) = \sum_{m=M}^{N} \frac{N!}{m!\,(N-m)!} \left[P_F(1)\right]^m \left[1 - P_F(1)\right]^{N-m} \tag{11.36}$$

For any specific problem, one might simply evaluate Eqs. (11.35) and (11.36) using a computer for a variety of M and N and $P_x(1)$ to create a table. Once the desired values of P_F(M of N) and P_D(M of N) are established, then the required $P_D(1)$ and $P_F(1)$ are found by interpolation. These are converted to CNR requirements using the single-pulse results of the preceding section.

For any value of N, there is an optimum value of M. Use of non-optimum M values will require significantly larger CNR values to achieve the desired probabilities. Figure 11-10 shows the optimum value of M for each value of N. For large N, the optimum value of M is approximately

N/4. Figure 11-11 compares the performance of M-of-N detection to the optimum pulse integration scheme. M-of-N detection is seen to suffer less than a 2 dB penalty relative to the optimum scheme.

Figure 11-10. Optimum choice of M given a specified value of N.

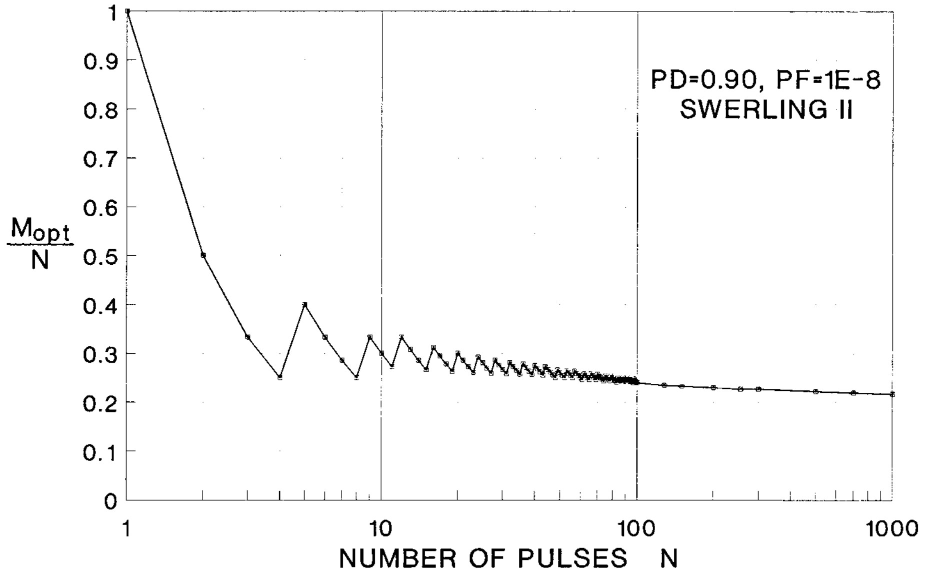
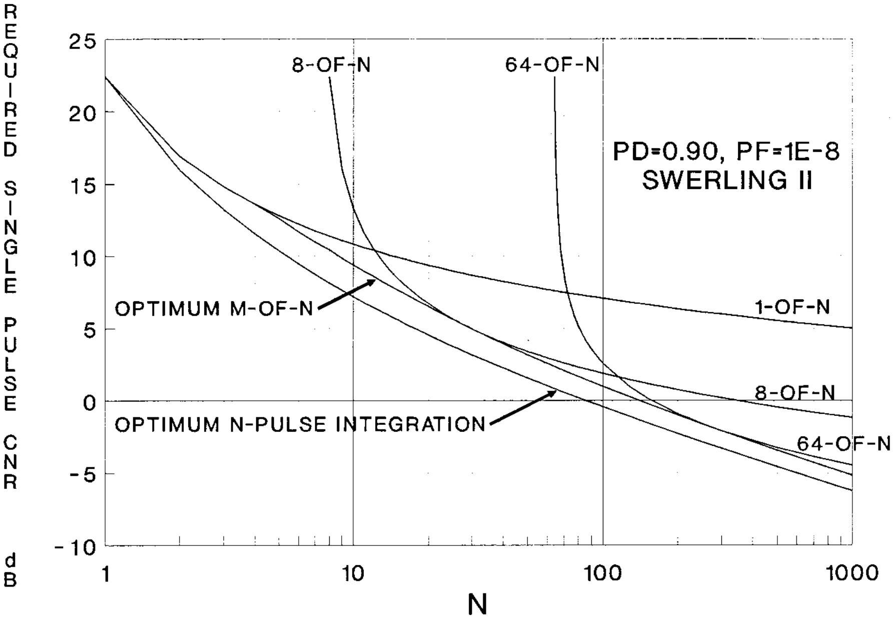


Figure 11-11. Comparison of three specific M-of-N integration schemes with optimum pulse integration and the optimum M-of-N integration scheme.



Example: Turbulence Effects on Laser Radar Detection

In Chapter 4 the effects of atmospheric turbulence on sensor performance were hinted at but could not be quantitatively discussed. Having studied the effect of fluctuations on detection probability we are now in a position to quantitatively discuss some turbulence effects on sensor performance. Turbulent scintillation causes amplitude fluctuations of magnitude

$$I = I_0 e^{2\chi} \tag{11.37}$$

where χ is the log-amplitude. The log-amplitude is described by a normal (Gaussian) distribution with a mean equal to minus the variance, i.e.,

$$p_\chi(\chi) = \frac{1}{\sqrt{2\pi\sigma_\chi^2}} e^{-\frac{\left(\chi+\sigma_\chi^2\right)^2}{2\sigma_\chi^2}} \tag{11.38}$$

where the log-amplitude variance is given by

$$\sigma_\chi^2 = 0.56 k^{7/6} R^{11/6} \int_0^1 d\zeta \, C_n^2(\zeta) \, \zeta^{5/6} (1-\zeta)^{5/6} \tag{11.39}$$

with

$$k = 2\pi / \lambda \tag{11.40}$$

and

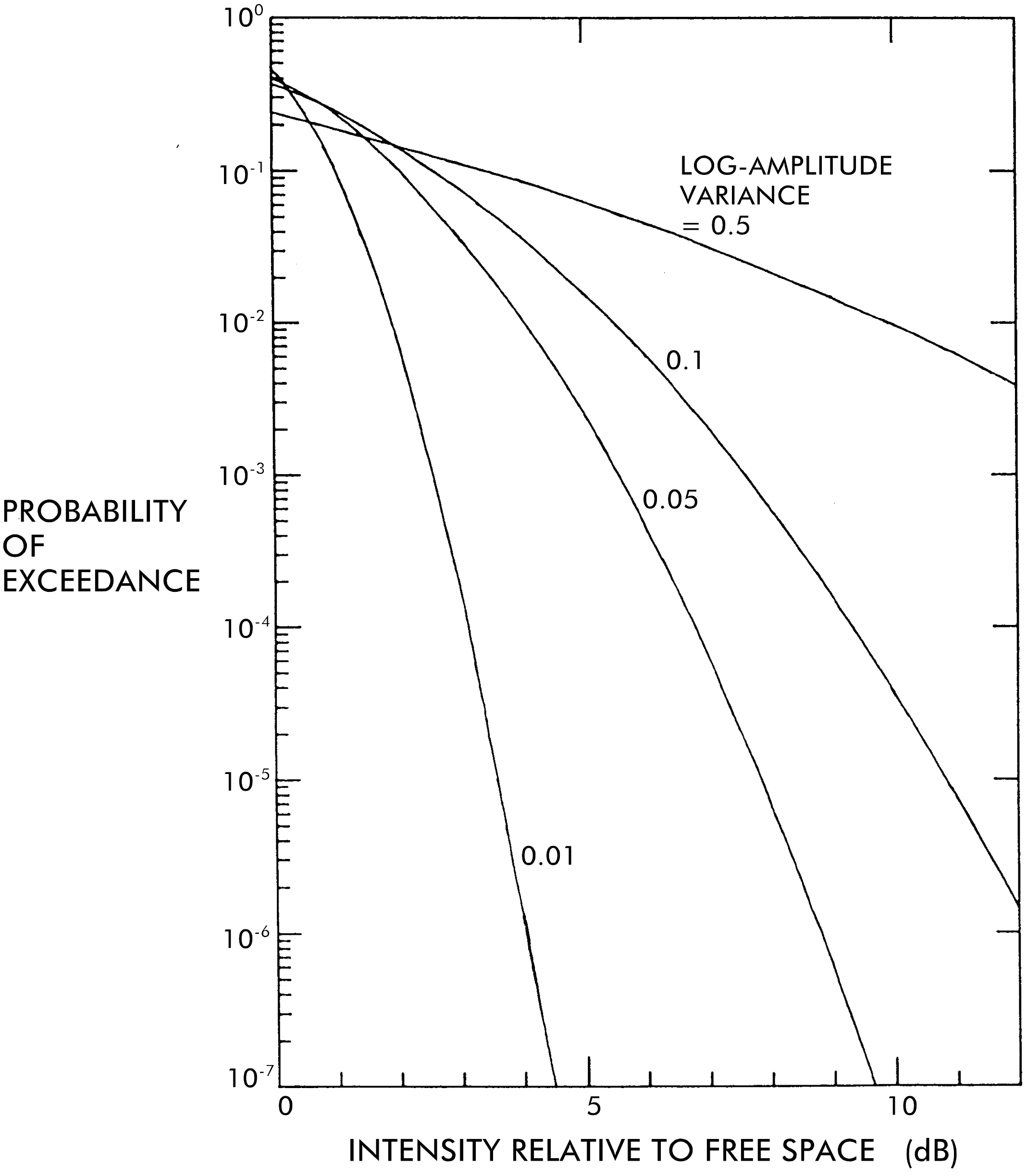
$$\zeta = z / R = \text{Normalized distance along path}. \tag{11.41}$$

Combining Eq. (11.37) with the integral of Eq. (11.38) we can determine the probability that the scintillation can produce very high intensity fluctuations. This is shown in Figure 11-12. It is easy to see that near saturated scintillation ($\sigma_\chi^2 \approx 0.5$) intensity spikes greater than 10x normal can occur as much as 1% of the time. Fades in intensity are even deeper and more frequent. Turbulence fluctuations should be expected to have significant impact on the receiver operating characteristics of systems operating in the visible or infrared. Let us examine turbulence effects on coherent laser radar systems. Except for the turbulence sensitivity, laser radars will behave exactly like the microwave radars we have discussed earlier.

Let us consider the case of a non-fluctuating target (a glint target, a specular reflecting target, or a Swerling 0 target) subject to turbulent scintillation. In a radar, the scintillation can act over both the transmitted and the received path. The intensity statistics are then given by

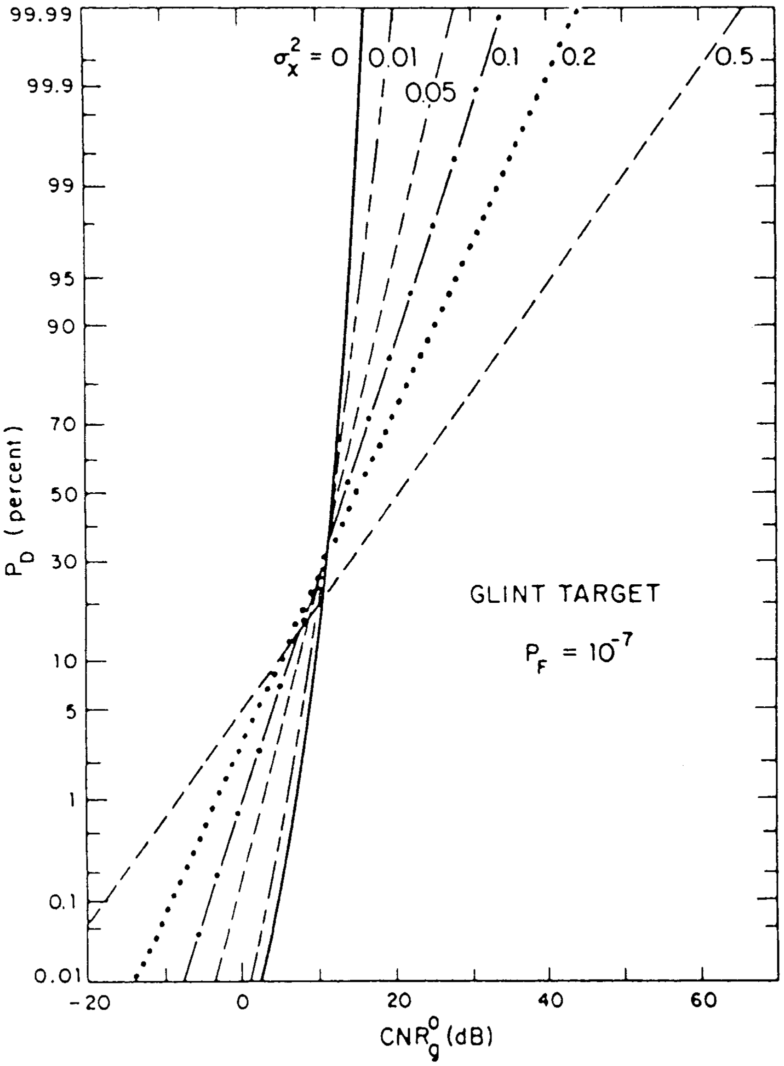
$$I_g = I_0 e^{4\chi} \tag{11.42}$$

Figure 11-12. Percentage of the time that intensity spikes due to turbulent scintillation exceed given limits as a function of turbulence strength.



with χ given by Eq. (11.38). Inserting these results into Eq. (11.23) and numerically integrating the result yields curves of the form of Figure 11-13. A relatively complete set of ROC curves may be found in Reference [4]. Comparing Fig. 11-13 to Figures 11-6 and 11-8 we see that turbulence fluctuations near saturated scintillation requires orders of magnitude more margin for adequate detection than does Rayleigh fading.

Figure 11-13. Parametric behavior of the receiver operating characteristic for $P_F = 10^{-7}$ showing the effect of increasing turbulence strength on normally non-fluctuating targets.



If we wish to investigate the effects of turbulence on a Rayleigh-fading target things get a little more complicated. First, scintillation which occurs on the transmitted path will produce an amplitude modulation at the target. Since Rayleigh-fading targets are not point sources (they are by definition composed of a large number of point scatterers), the transmitter-to-target path scintillation will be partially averaged out by the very process that produces the Rayleigh fading. This aperture averaging alters the probability distribution of fluctuations to

$$I_s = I_0 e^{4\sigma_\chi^2} V e^{2U} \tag{11.43}$$

where V is a unit-mean, exponentially distributed random variable with distribution

$$p_V(V) = e^{-V} \tag{11.44}$$

which accounts for Rayleigh fading and U is a log-normal variable given by

$$p_U(U) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(U+\sigma^2\right)^2}{2\sigma^2}} \tag{11.45}$$

where σ^2 is related to the normal log-amplitude variance by

$$e^{4\sigma^2} - 1 = \varsigma\left(e^{16\sigma_\chi^2} - 1\right) \tag{11.46}$$

with ζ being an aperture-averaging factor given by

$$\varsigma = \begin{cases} \left(d^2 / \lambda L\right) / \left[1 + \left(d^2 / \lambda L\right)\right] & d_T > \lambda L / d \\ 1 / \left[1 + \left(d_T^2 / \lambda L\right)\right] & d_T \leq \lambda L / d \end{cases} \tag{11.47}$$

where d_T is the target dimension. Figure 11-14 shows the effect of aperture averaging on the log-amplitude variance for a range of aperture-averaging factors.

Substituting these results into Eq. (11.23) allows numerical computation of the receiver operating characteristics of a Rayleigh fading targets subjected to additional turbulent scintillation. A sample result for $P_F = 10^{-7}$ is shown in Figure 11-15. A complete set of ROC curves for a variety of false alarm probabilities and turbulence log-amplitude variances may be found in Ref. [4]. As with non-fluctuating targets we see that near saturated scintillation enormous CNR margins are required to give adequate detection performance. It is for this as well as for reasons to be made clear in Chapter 14, that operation anywhere near saturated scintillation is to be avoided if possible..

Figure 11-14. Aperture-averaged log-amplitude variance σ^2 vs. log-amplitude variance σ_χ^2 for resolved targets with various transmitter Fresnel numbers ($d^2/\lambda L$).

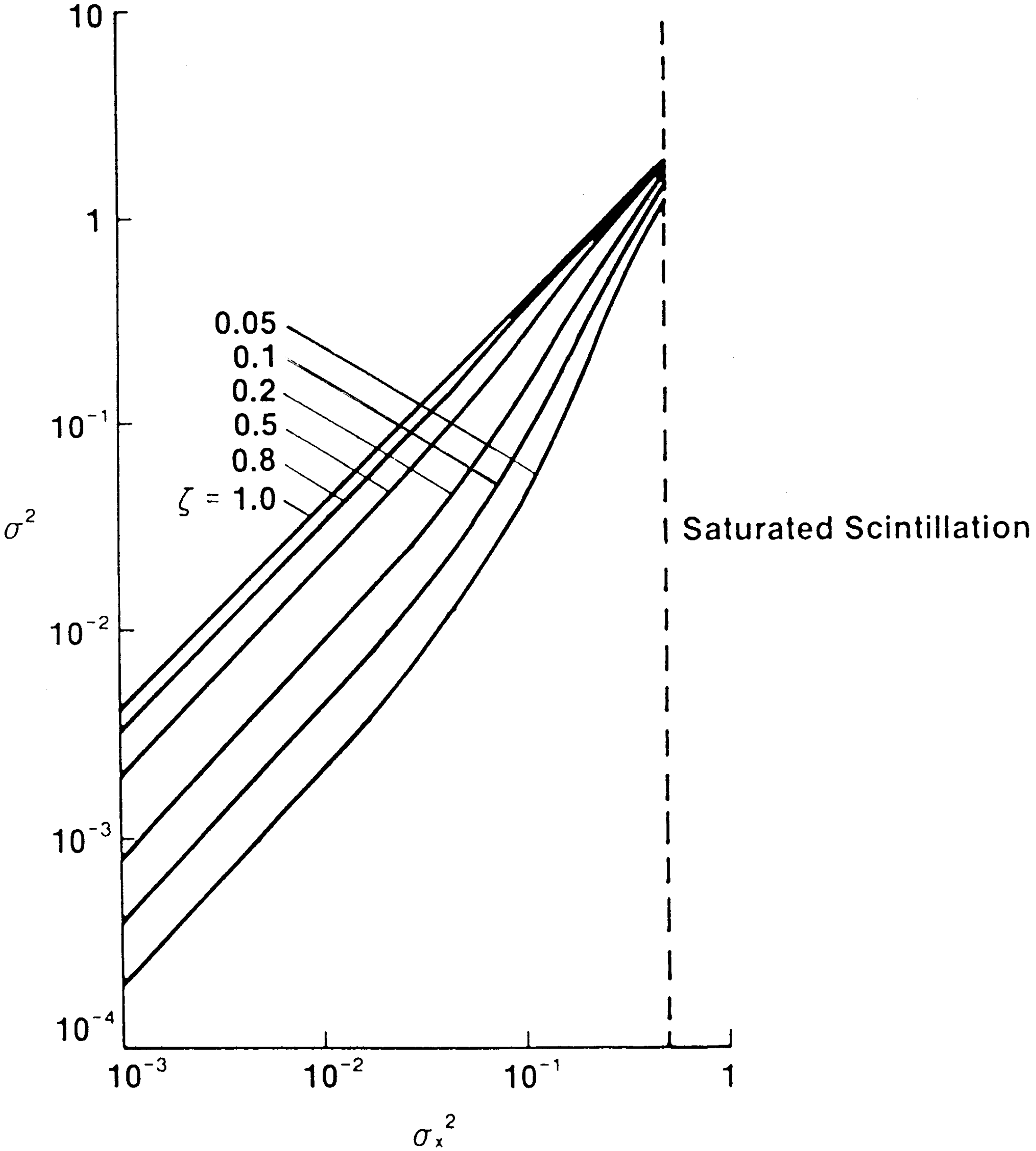
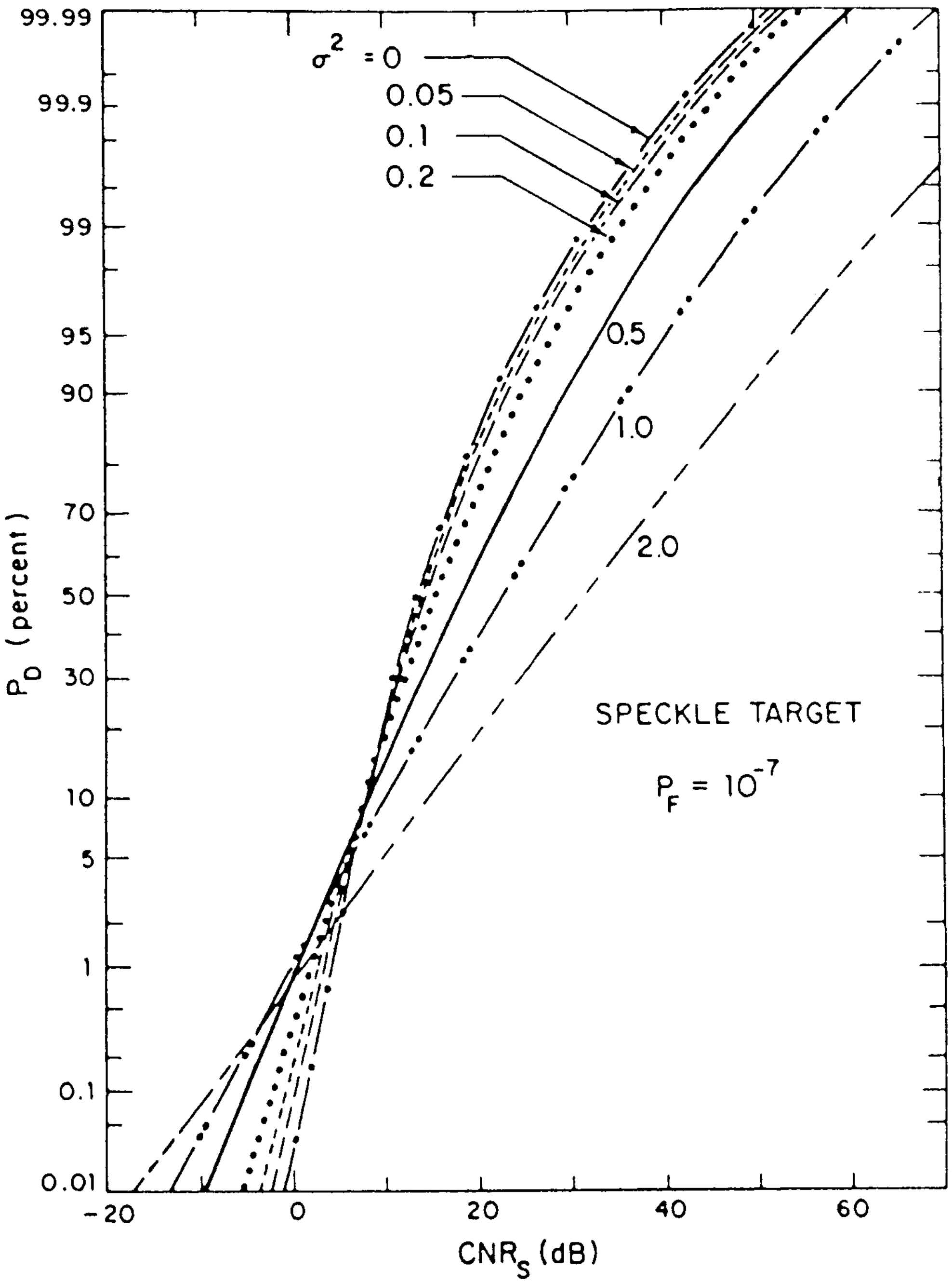


Figure 11-15. Parametric behavior of the receiver operating characteristic for $P_F = 10^{-7}$ showing the effect of increasing turbulence strength on normally Rayleigh-fading targets.



System Level Concerns Involving Detection

The probability of detection and probability of false alarm are system level parameters. That is, they are usually determined by mission performance considerations. There is no perfect choice of P_D and P_F. Acceptable values are determined by considering the penalties that each choice imposes on the system.

A missed detection imposes a penalty on the system. The penalty may take the form of an increased reaction time (because additional looks at the scene are required to make the detection needed to start any reaction process). It may directly result in increased casualties. It might require only a single additional pulse to produce a required valid range estimate (and have almost no impact on the timeline). A missed detection might also result in total annihilation (e.g., if the missed detection were the launch of a massive nuclear strike or the approach of a single biological weapon containing an "Andromeda Strain" agent.[5] Clearly there is a huge range in the potential penalties that missed detections impose.

Similarly, there is a wide range in the penalties imposed by a false alarm. The penalty might be the switching of a system from a search mode into a tracking mode (with consequent loss of search capability until the false alarm fails to recur and track mode is canceled). A false alarm might be nothing more than a minor irritation to an operator. It might appear as "snow" on a display. If the operator must take some action, then the false alarm might become a major annoyance if it happens too frequently. A false alarm might lead to catastrophe (e.g., if the false alarm were the declaration of an incoming nuclear strike coupled to an automatic counterattack response – which would of course trigger the real launch of the opposing side's missiles).

The probabilities of false alarm and detection need to be established by determining the maximum level of penalties that can be tolerated and determining the values of P_D and P_F that correspond to those maximum tolerable penalties. This will result in the lowest acceptable value of P_D and the highest acceptable value of P_F. Combined these will result in the lowest required carrier-to-noise ratio. Since detection probability affects required CNR more strongly than false alarm probability, more attention should be paid to the penalty analysis associated with missed detections than to the analysis associated with false alarms.

As an example of how this might be done, consider a simple laser rangefinder. The device can make one range measurement per second for as long as the operator pushes the fire button. Each range measurement has a resolution of 10 meters and the range is sampled continuously over the span of 0 m to 10,000 m. The measured range is displayed visually to the operator. Over the course of a year, a typical operator of this device will only use the rangefinder about 100 times. We need an estimate of the required P_D and P_F to include in a system specification.

What are the penalties associated with a missed detection? Basically, if the operator fails to see a valid range appear on the display, he pushes the fire button, he must push the button again. Failure to get a valid range (i.e., the system misses a detection) will cost the operator little more than one second of time. In one second, typical objects of interest will have moved no more than a few tens of meters. This is insignificant. On the other hand if the operator misses ten measurements in a row, a long time (10 seconds) is wasted, and the target situation may have changed unacceptably. A few missed detections (perhaps as many as three) in a row is acceptable; ten is not. If an unacceptable number of missed detections happens only once in the career of an operator he will seldom mind – the unacceptable performance is not only rare, but it is much more likely to occur during training than during combat. If we assume the operator has an average career of ten years and uses the rangefinder 100 times per year, we find that one "run" of missed detections is acceptable every 1000 uses. If we say an unacceptable run is three missed pulses then

$$P(3\text{ misses}) = 1 - (1 - P_D)^3 \le 0.001 \qquad \Rightarrow \qquad P_D \approx 0.90$$

Thus we obtain a result that states that $P_D = 0.90$ is almost certainly acceptable. It is not the right answer, but it is not the only right answer, A second individual might have changed the initial assumptions slightly and obtained $P_D = 0.80$ or $P_D = 0.95$.

What are the penalties associated with a false alarm? A false alarm will manifest itself as an incorrect range value selected randomly from all of the possible range values. The user almost always has a visible check on the rough range to the target. From size alone he can estimate that a tank is closer to 1 km range than to 10 km. Thus as many as 80% of all false alarms will be immediately recognized as being errors. The remaining 20% are too close (within 10%) to the true value to be recognized as erroneous. If the rangefinder is being used to direct artillery, indirect fire weapons, then the penalty for error might be missing the target and wasting a few artillery shells. If the rangefinder is being used to determine a fire control solution for a direct fire weapon, there is a good chance that the range error would cause a missed target. This could conceivably result in the target being able to return fire and kill the person with the rangefinder. This is a penalty that the observer would find unacceptable. Thus it is desirable that the probability of a false alarm occurring be no worse than one "undetectable" false alarm in the career of a user. That is, no more than 1 false alarm in 200 range measurements. Since there are 1000 possibilities for a false alarm to occur on every laser pulse (10,000 range span divided by 10 m resolution), the resultant false alarm probability is

$$P_F = 1/(200 \text{ x } 1000) = 5 \text{ x } 10^{-6}$$

Thus, a reasonable set of requirements is $P_D = 0.90$ and $P_F = 5 \text{ x } 10^{-6}$. However, another systems engineer might have come with different numbers. Correctness is entirely in the eye of the beholder.

Clutter and Clutter Rejection

Most sensor systems are sensitive to signals of one form or another that are produced by objects in the scene that are neither targets nor other objects of interest. Because they are real objects, they produce real threshold crossings that are not false alarms. They will recur in every scan. These signals often have properties that make them difficult to distinguish from the objects of interest. Collectively, these signals are called **clutter**, due to the visual effect they have on detecting targets in a radar display.

Clutter signals can be produced by almost real object that is of secondary importance or less. Common sources of clutter include emission or reflection of radiation from:

* Terrain features
* Birds, animals, or insects
* Clouds
* Precipitation
* Moving foliage or water waves
* Junked or destroyed vehicles and other debris

Some of these objects are far more important is some classes of sensors than in others, and vice versa. For example, birds are a serious form of clutter in infrared search and track systems. They look and act just like aircraft targets. Water waves are a serious clutter source in marine radar systems.

Clutter signals exceed the threshold of detection of a system and produce **"clutter false alarms"**. Real false alarms are due to noise and are random occurrences. They do not recur in look after look. Clutter false alarms are due to real objects and do persist over many observations. The primary deleterious effects of clutter false alarms are that they attract the operator's attention and may cause him to fail to detect real target objects the sensor detected. They make tracking more difficult by making the scene-to-scene association of targets more difficult (see Chapter 13). Their similarity to real targets often precludes automation of the detection process. In general, it is entirely possible for clutter to degrade system performance to unacceptable levels.

Sensors with clutter limitations employ a variety of techniques collectively known as **clutter rejection** to ameliorate the effects of clutter. Among the techniques that might prove useful in any given situation include:

* Raising the detection threshold (with consequent loss of real detection probability)
* Constant false alarm rate (CFAR) processing (varying the detection threshold adaptively to maintain a constant rate of threshold crossings – this also degrades detection probability)
* Moving target indication (discrimination based on target motion)
* Spatial filtering of image data (discrimination based on target shape and size)
* Sensitivity time control (emphasizes distant returns versus close-in returns)
* Optimized waveforms

No clutter rejection technique is perfect. Most result in some loss of detection, but less than occurs if clutter is present. Clutter rejection will be revisited in the individual sensor discussions.

References

[1] Skolnik, Merrill I., Introduction to Radar Systems 2nd Ed. (McGraw-Hill, New York NY, 1980) Chap. 2.

[2] Shapiro, J. H., Capron, B. A., and Harney, R. C., "Imaging and target detection with a heterodyne-reception optical radar", *Applied Optics*, 20, #19, 3292-3313 (1 October 1981).

[3] Swerling, Peter, *IRE Transactions on Information Theory*, IT-3, #3, 175-178 (1957).

[4] Capron, B. A., Harney, R. C., and Shapiro, J. H., "Turbulence Effects on the Receiver Operating Characteristics of a Heterodyne-Reception Optical Radar", M. I. T. Lincoln Laboratory Project Report TST-33 (July 1979).

[5] Crichton, Michael, The Andromeda Strain (Alfred A. Knopf, New York NY, 1969).

Problems

11-1. How does signal-to-noise ratio differ from carrier-to-noise ratio?

11-2. Write the integral expressions for false alarm probability and detection probability?

11-3. What is a receiver operating characteristic and how is it utilized?

11-4. Increasing the threshold voltage in a detection system has what effects on detection probability and false alarm probability?

11-5. A system specification requires certain detection and false alarm probabilities. The initial hardware concept (as sized) does not have adequate CNR to produce those detection and false alarm probabilities with single pulses. Without abandoning the system concept entirely, what options are available to you to provide the required performance? Hint: there are at least three totally unrelated approaches that may be taken.

11-6. In the system of problem 11-5, which option (or options) is likely to yield the best (acceptable) solution in the following cases:

a. The CNR shortfall is 3 dB?
b. The CNR shortfall is 10 dB?
c. The CNR shortfall is 30 dB?

11-7. A radar was designed to work with $P_D = 0.99$ and $P_F = 10^{-8}$ against a target with Swerling 2 statistics. Later intelligence exploitation of the target yields the fact that the target statistics are in fact Swerling 4 even though the estimated cross section was correct.. What implications does this discovery have on required CNR? What is the range performance implication of this discovery assuming the target remains the same? Assume that for this radar

$$CNR \propto 1/R^4$$

11-8. A ground surveillance radar for armored vehicles uses 100-pulse integration to provide 22.3 dB integration improvement. Estimate the miss probability of this radar. If CNR = 2.1 dB yields adequate performance after integration improvement, what is the approximate false alarm probability? If the mean time between false alarms is one day, what is the bandwidth of the radar electronics? If we lower the average noise by 10%, what will be the new mean time between false alarms (for the same bandwidth as above)?

11-9. You are designing a mid-range air search track-while-scan (TWS) radar for littoral operations. Intelligence reports that a traditionally belligerent nation has developed an effective attack technique known only as "flock of seagulls". This technique uses a medium-bomber aircraft that launches a mass of winged, unguided rockets that form a swarm around it as it enters the SM-2 engagement zone of our ships. Intelligence suggests that the purpose is to induce cross section variations that are equivalent to those of one large steady return and numerous, small random reflectors. Given a required $P_D = 0.9$ and $P_F = 10^{-6}$, calculate the CNR required for the TWS radar to effectively detect the "flock of seagulls". If 9 pulses is the maximum number available for pulse integration, recalculate the CNR required if pulse integration is used. The hostile bomber alone obeys Swerling 2 statistics. If "flock of seagulls" is truly designed to adversely P_D and P_F, did the hostile nation make a wise choice in deploying "flock of seagulls"?

11-10. What CNR is required by a radar to provide a 90% detection probability with a 10^{-6} false alarm probability for 10 pulse optimum integration assuming the target obeys Swerling II statistics? If the radar bandwidth is 10 MHZ, what is the mean time between false alarms?

11-11. Given a sensor noise probability density $p_N(V)$ and a signal fluctuation probability distribution $p(\sigma)$, how would you calculate a set of receiver operating characteristic curves for that sensor.

11-12. What is M-of-N detection? Why would a system designer consider using it? What penalty relative to a perfect system would a system using M-of-N detection suffer?

11-13. A laser radar must have a detection probability of 0.999 against a Swerling II (speckle) target. The radar has a 20 dB margin to accommodate a certain amount of turbulent scintillation degradation. What is the maximum turbulence log-amplitude variance that will not cause an unacceptable degradation?

11-14. Determine reasonable values of detection probability and false alarm probability for a simple laser rangefinder used as part of a tank gun fire control system. The rangefinder is capable of making 1 Hz range measurements indefinitely.

CHAPTER 12

ESTIMATION

Characteristics of Measurements

Measurements with sensors are made with a purpose. One of these purposes is to estimate the value of some parameter of interest. For example, we may wish to estimate the range to a waypoint in a navigation application or the angular size of a target in a target recognition application. There is a science of measurement. Investigators have defined a number of characteristics of measurements with which a systems engineer must be familiar. In this section we will look at some of those definitions.

Accuracy of any measurement is clearly of interest. The IEEE Dictionary [1] defines **accuracy** as:

> *"the quality of freedom from error or mistake."*

Error is:

> *"any discrepancy between a computed, observed, or measured quantity and the true, specified, or theoretically correct value or condition."*[1]

It is obvious that one wants measurements to be as accurate as possible. This requires understanding and controlling the sources of error. There are different kinds of errors. A **random error** is:

> *"a component of error whose magnitude and direction vary in a random manner in a sequence of measurements made under nominally identical conditions."*[1]

Random errors may arise from the effects of sensor noise, from imperfect implementation of sensor components, or from temporal variation of systematic errors. A **systematic error** is:

> *"the inherent bias (offset) of a measurement process or one of its components"*[1]

while a **bias error** is defined as:

> *"a systematic error whether due to equipment or ambient conditions."* [1]

Bias errors can result from imperfect implementation of sensor components, imperfect alignment or calibration of sensor components, or use of an imperfect "estimator". An **estimator** is a processing approach to making an estimate of some value of a signal. For example, an estimator of time of arrival of a pulse is to set a threshold, measure the times at which the pulse crosses the threshold (both upward and downward crossings), add the results and divide by 2. Some estimators are biased. That is, they respond differently to upward noise fluctuation than to downward noise fluctuations. This leads to a biased estimate with the magnitude of the bias being a function of the relative noise level.

Related to accuracy, but not equivalent to it, are the concepts of precision, repeatability, and uncertainty of measurements. **Precision** has two definitions. One is:

> *"the degree of exactness or discrimination with which quantity is stated."*[1]

The second is:

> *"the quality of coherence or repeatability of measurement data."*[1]

Repeatability is:

> *"the closeness of agreement among repeated measurements of the same variable under the same conditions."*[1]

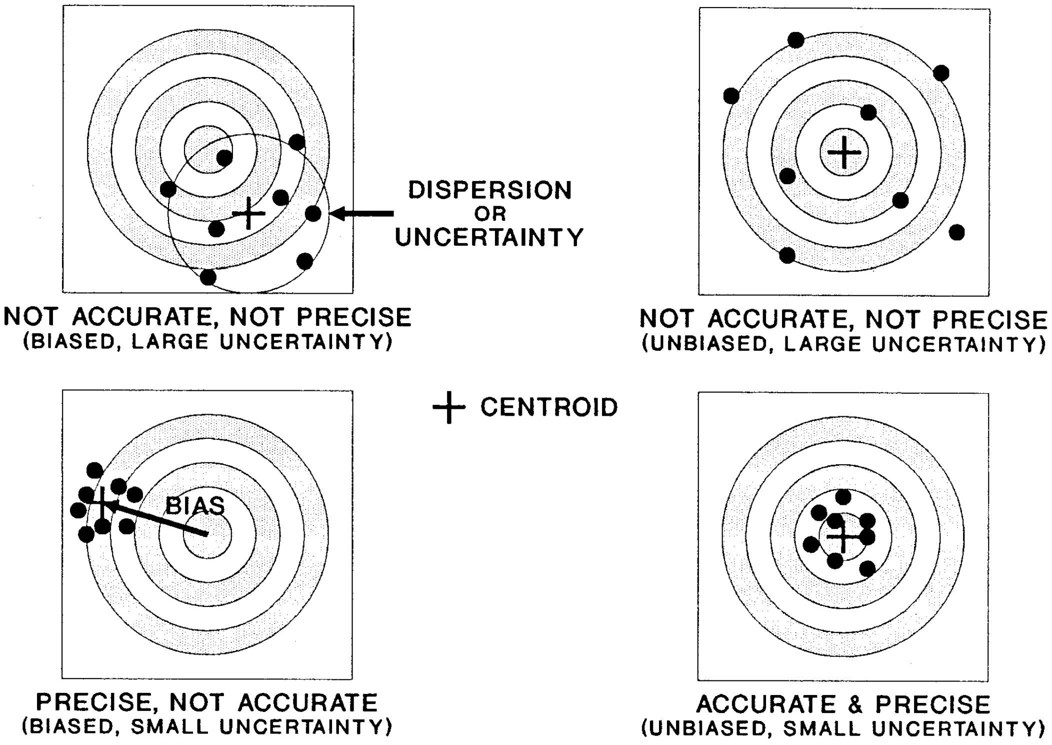
Uncertainty is:

> *"the estimated amount by which the observed or calculated value of a quantity may depart from the true value."*[1]

The uncertainty is usually given as the standard deviation or the probable error of the random distribution associated with the observation or calculation.

The relationship between accuracy and precision is illustrated in Figure 12-1 using the context of bullets fired at a target. The upper left target shows a situation that is neither accurate nor precise. There is a substantial bias and there is a large uncertainty. In the target at the upper right, the bias has been completely eliminated. This might seem accurate, but there is still a large uncertainty, so none of the individual measurements is likely to come close to the true value (accuracy is the freedom from error – of any kind). Thus, the situation is still neither accurate nor precise. The situation at lower left is precise but still not accurate. The uncertainty has been largely eliminated, but a large bias remains. Remember that accuracy is the freedom from all error. You cannot be accurate is the result is biased or has a large uncertainty. The target at lower right shows a situation that is reasonably accurate and precise. Both the bias and the uncertainty are reasonably limited.

Figure 12-1. Relationship between accuracy and precision.



NOT ACCURATE, NOT PRECISE
(BIASED, LARGE UNCERTAINTY)

NOT ACCURATE, NOT PRECISE
(UNBIASED, LARGE UNCERTAINTY)

PRECISE, NOT ACCURATE
(BIASED, SMALL UNCERTAINTY)

ACCURATE & PRECISE
(UNBIASED, SMALL UNCERTAINTY)

Sensor Resolution

Another property of measurements that deserves considerable study is the resolution of the sensor that obtained them. The dictionary defines **resolution** as:

> *"the smallest change in the pulse characteristic, property, or attribute being measured which can be unambiguously be discerned or detected in a pulse measurement process."*[1]

However, the author prefers to use his own definition:

> *"the smallest separation in the variable space between two identical point objects at which the presence of two separate signal contributions can be discerned or detected."*

Resolution should be distinguished from a quantity called resolvability. **Resolvability** is defined by the author as:

> *"the ability of a system to detect or discern the presence of individual signal contributions of two objects of specified (and possibly unequal) physical extent and specified (and possibly unequal) signal strength separated in the variable space by a specified amount."*

It is quite possible that resolvability is the desired system performance value, but resolution is mistakenly specified instead. For example, if the problem is the detection of a small boat alongside a large ship, then specifying resolution may not guarantee that detection is possible. Specifying resolvability using the relative signal strengths and relative physical sizes would guarantee that performance. Resolution is relatively easy to determine. Resolvability involves multiple parameters (not just separation) and is considerably more difficult but not impossible to evaluate.

Angular resolution is qualitatively determined by moving two point source targets apart until the superposition of the two point source response functions exhibits a discernible dip. For a sensor with point source response *G(θ)* and two identical point sources with angular separation *θ*, the ratio of the intensity of the sum response mid-way between the two sources to the sum response at the position of either source defines a **resolution function** *R(θ)*

$$R(\theta) \equiv \frac{I_{MIDPOINT}(\theta)}{I_{SOURCE\ POSITION}(\theta)} = \frac{2G(\theta/2)}{G(0)+G(\theta)} \tag{12.1}$$

The angular separation producing a specified value (determined by the resolution criterion used – see below) of $R \leq 1$ is the sensor's angular resolution. *R(θ)* can be considered to be a "resolution function". It should be noted that this analysis is perfectly general and time *t* could easily be substituted for *θ* (although using units of *λ/D* would no longer be appropriate).

For uniformly illuminated apertures (e.g., as found in most kinds of passive sensor), the point source response function is an Airy function [2]

$$G_{AIRY}(\theta) = \left[\frac{2J_1(\theta)}{\theta}\right]^2 \tag{12.2}$$

where $J_1(\theta)$ is the Bessel function of the first kind of order 1. For Airy function intensity distributions, Rayleigh assumed that two sources would be resolved when the peak of the first Airy function coincided with the first zero of the second Airy function. At this point, for aperture diameter *D* and wavelength λ, the angular separation is

$$\theta = 1.220\,\lambda / D = \alpha \tag{12.3}$$

where α is the angular resolution and the resolution function has a value of

$$R_{AIRY}\left(1.220\,\lambda / D\right) = 0.735 \tag{12.4}$$

This **Rayleigh resolution criterion** corresponds to a 26.5% intensity dip. It is commonly utilized when describing optical sensors.

Sparrow assumed that two sources would be barely resolved when the sum curve was flat at the midpoint. In this case, the angular separation is

$$\theta = 1.007\,\lambda / D \tag{12.5}$$

and the resolution function has a value of

$$R_{AIRY}\left(1.007\,\lambda / D\right) = 1.000 \tag{12.6}$$

The **Sparrow resolution criterion** corresponds to a 0% intensity dip. It is commonly used when describing microwave sensor systems and is often approximated as $\theta = \lambda/D$.

Laser beams typically possess Gaussian intensity distributions. A sensor might have a simple Gaussian point source response:

$$G_{GAUSSIAN}\left(\theta\right) \propto e^{-2\theta^2/\theta_0^2} \tag{12.7}$$

where $\theta_0 = \lambda/\pi w_0$ is the laser beam divergence. A convenient resolution criterion when dealing with Gaussian distributions is to assume that two sources are resolved when the peaks of the two distributions are separated by an amount equal to the full-width-at-half-maximum-intensity of one of the distributions. It is easily shown that this produces

$$\theta = 0.5\left(2\ln 2\right)^{1/2}\theta_0 = 0.588705\,\theta_0 \tag{12.8}$$

and a value

$$R_{GAUSSIAN\ FWHM}\left(\theta\right) = 0.941 \tag{12.9}$$

The **Gaussian FWHM resolution criterion** produces a 5.9% intensity dip. Note: when comparing resolutions between different kinds of sensors, it is important to use the same resolution criterion for all.

In any radar the point target response function must contain contributions from the transmitter beam profile and the receiver response profile. Since the total response profile is the product of two profiles, the total response profile will be narrower than either the transmitter profile or the receiver profile. Thus, for a given aperture, an active sensor will have higher resolution than the comparable passive sensor. Few people, including many sensor designers, are aware of this fact.

In a laser radar, both transmit and receive profiles may be approximated by Gaussian profiles. [3,4] The resulting response profile is given by

$$G_{SQUARED-GAUSSIAN}(\theta) \propto e^{-2\theta^2/\theta_0^2} e^{-2\theta^2/\theta_0^2} = e^{-4\theta^2/\theta_0^2} \qquad (12.10)$$

where $\theta_0 = \lambda/\pi w_0$, and w_0 is the spot size at the Gaussian beam waist. The spot size is usually adjusted such that the aperture diameter is roughly $2\sqrt{2}$ (2 to 3) times the spot size. With this choice, 95% of the beam energy passes through the aperture, sidelobes are small, and the resolution is kept high. Using the new expressions for $G(\theta)$ and w_0, the Gaussian FWHM criterion resolution of the laser radar is found to be

$$\theta = 0.5(\ln 2)^{1/2}\theta_0 = 0.41628\theta_0 \ . \qquad (12.11)$$

The Gaussian FWHM criterion still corresponds to a 5.9% dip. In terms of the aperture-matched spot size we can define the angular resolution of a laser radar as

$$\alpha = \frac{(\ln 2)^{1/2}}{\pi}\frac{\lambda}{w_0} = \frac{(M\ln 2)^{1/2}}{\pi}\frac{\lambda}{D} = 0.26501\sqrt{M}\frac{\lambda}{D} \qquad (12.12)$$

where M is an aperture matching criterion (the Gaussian spot size w_0 must take some fixed relationship to the aperture diameter D. We have defined the matching as $D = (M)^{1/2}w_0$. M typically takes on a value between 4 and 9. Using a conservative value of $M = 8$, we calculate

$$\alpha \approx 0.74956\frac{\lambda}{D} \qquad (12.13)$$

The Rayleigh criterion for the squared-Gaussian profile becomes

$$\alpha = 0.892\,\lambda / D \qquad (12.14)$$

while the Sparrow criterion for the squared-Gaussian profile becomes

$$\alpha = 0.703\lambda / D \, . \tag{12.15}$$

It is relatively easy to evaluate the resolution function for a simple Gaussian profile. If we assume the same aperture matching condition as above ($D = (8)^{1/2}w_0$), then

$$\alpha = 0.994\lambda / D \quad \text{Sparrow} \tag{12.16}$$

$$\alpha = 1.060\lambda / D \quad \text{Gaussian FWHM} \tag{12.17}$$

$$\alpha = 1.261\lambda / D \quad \text{Rayleigh} \tag{12.18}$$

It is equally easy to evaluate the resolution function for a squared-Airy function

$$G_{SQUARED-AIRY}(\theta) = \left[\frac{2J_1(\theta)}{\theta}\right]^4 \tag{12.19}$$

to yield

$$\alpha = 0.707\lambda / D \quad \text{Sparrow} \tag{12.20}$$

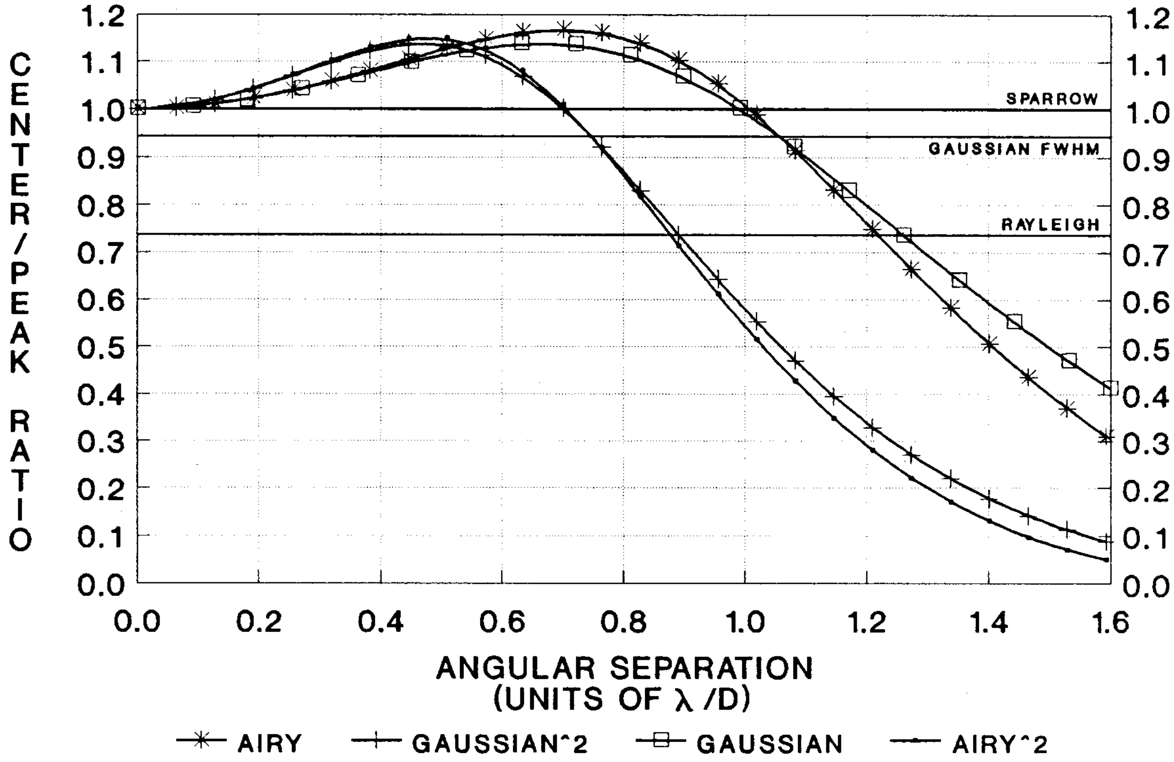
$$\alpha = 0.748\lambda / D \quad \text{Gaussian FWHM} \tag{12.21}$$

$$\alpha = 0.878\lambda / D \quad \text{Rayleigh} \tag{12.22}$$

The preceding results are illustrated in Figure 12-2.

The resolution of a sensor is often estimated from one of the equations above. However, there is little standardization in the definition of resolution. That, in fact, is why the author introduced a well-defined mathematical resolution function. When dealing with sensor resolution, one should always attempt to ascertain what type of point source response has been assumed (Airy, squared-Gaussian, or something else) and what resolution criterion (Sparrow, Rayleigh, Gaussian FWHM, or something else). When writing a specification on a system in which resolution is important, the specification should explicitly state the assumptions to be used. When examining vendor offerings one should bear in mind that vendors will almost always attempt to present their results in the best light. If you ask for the resolution, they will say it is x milliradians. They may be using the Sparrow criterion (an optimistic criterion that leads to small values of resolution) while you may be assuming the Rayleigh criterion (a very pessimistic criterion). The 20% difference between these criteria could easily result in inadequate system performance. If you don't require disclosure of the details, you can be very disappointed with what you actually get.

Figure 12-2. Resolution function dependence on angular separation.

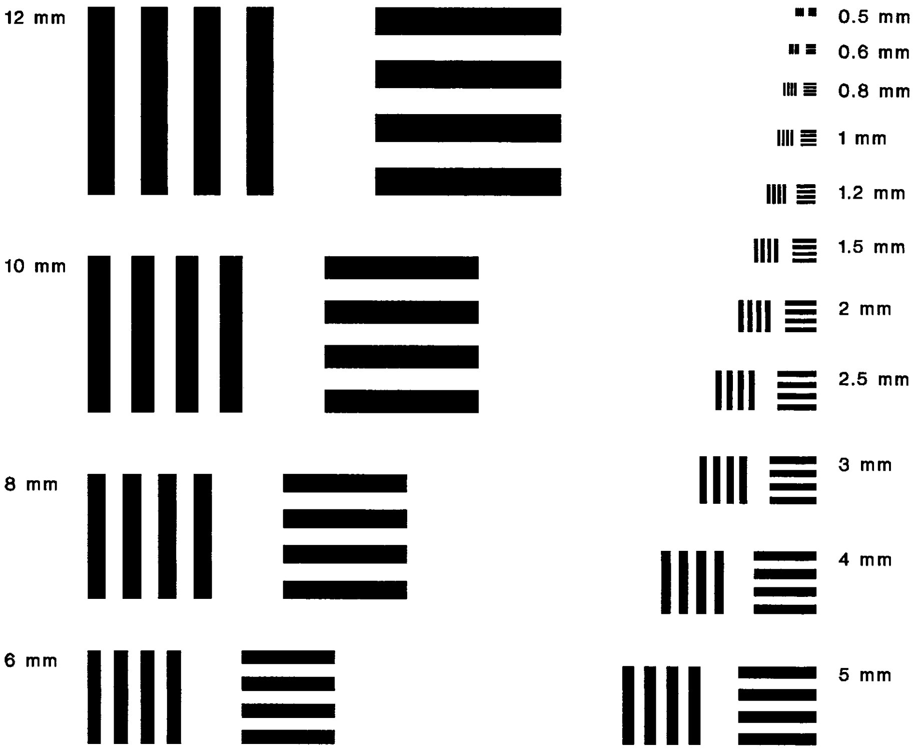


In estimating the resolution of a sensor you begin with the point source response function. In general this is a complicated function with no good analytical fit. However, from Figure 12-2, we see that there is little real difference (a few percent at most) between the results for two radically different active sensor response functions (squared-Airy vs. squared-Gaussian). There is very little difference between the curves for the two passive sensor response functions (Gaussian vs. Airy). There is considerable difference between the active and passive sensor response functions.[4] Thus, if a detailed response function is not available determine first whether the sensor is active or passive and then estimate whether the result is more likely to be approximated by Airy or Gaussian behavior. The resulting choice will give an estimate of limiting resolution that will be reasonably accurate. Next, a resolution criterion should be selected. Your own leanings towards pessimism or optimism should guide this choice. However, this should be tempered by the fact that picking a criterion will drive the physical size of the aperture of the sensor required to meet it. The 20% difference between Sparrow and Rayleigh criteria means a 20% difference in aperture size. This has potentially major impacts on size, weight, and cost of the total system. The author usually prefers the Gaussian FWHM criterion. It is reasonably optimistic but not overly so.

In the preceding paragraph the author referred to a limiting resolution of a system. The physics of diffraction imply that resolution cannot exceed $x\ \lambda/D$ where x is of order unity. However, other design considerations can lead to actual resolutions that are much larger (remember, when dealing with resolution, large is bad) than the limiting values. For example, any defocus of an optical system, any pointing jitter, any turbulence beam spreading, any narrowband filtering of a detector output, any spatial quantization imposed by a detector array, will result in image blurring and poorer resolution. In later chapters we will show how to explicitly handle these effects.

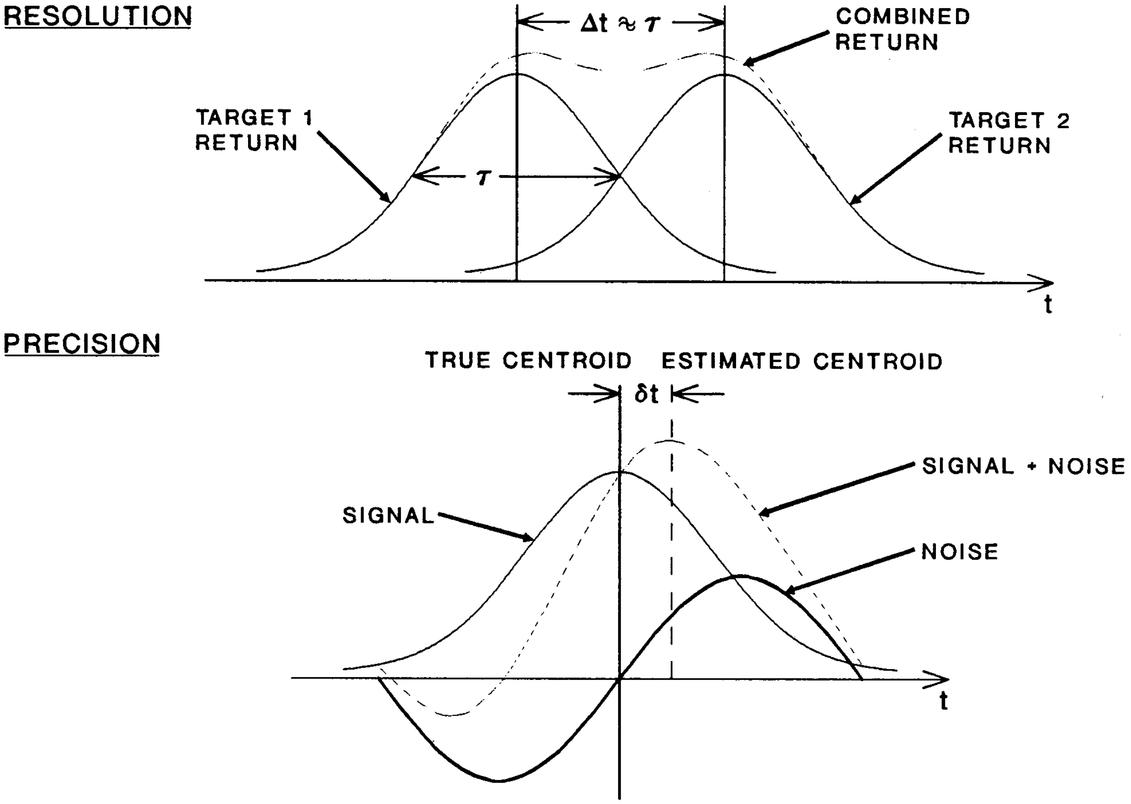
However, in real sensor systems, one need not calculate the resolution, one can measure it. This is done by viewing a resolution chart. A resolution chart is a graphic element that has regions of light and dark of varying sizes, shapes, and separations. It is viewed by the sensor under test and the smallest scale that still yields acceptable "resolution" based on a pre-selected resolution criterion is the effective resolution of the sensor. There are many standardized resolution charts that are available commercially for sensor testing. It is also possible to generate your own. Figure 12-3 shows a resolution chart the author generated for an experiment. It is organized based on four-bar patterns of effectively unit contrast such that horizontal and vertical resolutions can be testing separately. When the full-sized chart was placed at 1 meter from the subject, the resolution could be read directly in mm/mrad from the numbers beside each pattern. Please notice that after printing, scanning, reducing, and printing in this copy of the book, the resolution had degraded until the smallest patterns are no longer resolvable as four bars. The smallest pattern that can be discerned as four bars defines the effective resolution.

Figure 12-3. A simple four-bar resolution chart generated by the author.



Having discussed both precision and resolution, and noting that many individuals use resolution when they mean precision, it is worthwhile to directly compare the two quantities. This is done graphically in Figure 12-4. Remember that resolution involves two objects. When the objects are separated by sufficient distance, their superimposed responses will exhibit a significant dip. For Gaussian responses, a separation equal to the full-width-at-half-maximum gives a clearly discernible dip. Precision involves only one signal and describes our ability to precise locate the response in time due to noise superimposed on the signal. The figure shows a Gaussian signal with added sinusoidal noise. With the chosen relative phase for the noise note that the centroid of the sum is displaced by an amount δT from the true centroid. If the noise were to be selected to have a phase 180 degrees different from that depicted, the sum would jump to the other side of the true value. When all possible phases of the noise are averaged, a root-mean-square deviation that is not zero will exist. This is the precision of the measurement – the fundamental limit on the accuracy with which any set of measurements can be made. There is no fundamental physical connection between resolution and precision. However, as we shall see later, both resolution and precision depend on the characteristic time scale of the signal and will have a definite mathematical relation to each other.

Figure 12-4. Graphical representation of the difference between resolution and precision.



Parameter Estimation

Estimating the value of a target parameter (such as range or velocity) from a signal corrupted by noise is the second "elemental" sensor function that finds frequent utilization in higher-level sensor functions. In most cases, parameter estimation devolves to the process of estimating the location (value in space or time) of the center of an amplitude distribution related in some well-defined way to the target parameter of interest. Often the task is equivalent to estimating the time at which the center of a pulse occurs (most estimation problems can be transformed into such a pulse timing problem).

The simplest estimator of pulse center is to measure the time at which the peak of the pulse occurs and declare this time to represent the pulse center. Unfortunately, this estimator is prone to producing significant errors at low carrier-to-noise ratios and may occasionally generate anomalous estimates (large errors due to random noise spikes) even at moderate carrier-to-noise ratios.

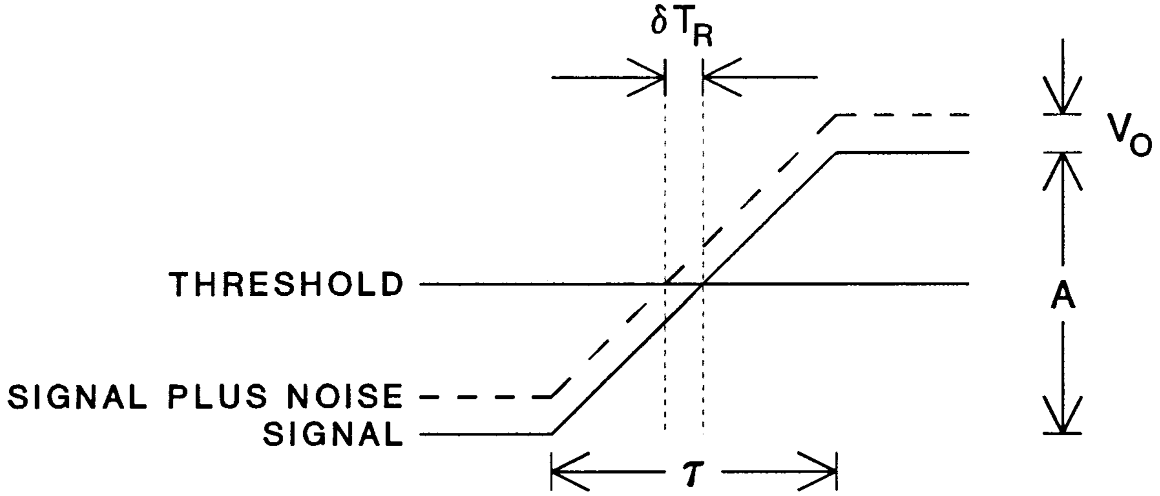
A more accurate estimator is determination of the centroid of the pulse. The centroid could be calculated digitally after sampling the pulse at high speed, or it could be determined in an analog fashion by dividing the pulse into two equal copies, delaying one copy with respect to the other integrating the undelayed copy, setting a threshold at half the amplitude of the integrated pulse, integrating the delayed copy and determining the time at which the integrated delayed waveform crosses the threshold.

A simple yet surprisingly accurate estimator of the pulse center is to threshold the pulse, determine the times at which the rising and falling threshold crossings occur, and declaring the time midway between the threshold crossings as the pulse center. This is the estimator we will investigate in more detail.

Other estimators can be constructed. However, all of them will produce estimates that are affected by noise on the pulse. Barring the production of anomalies by some estimators, the primary effect of noise on all estimators is to reduce the precision of their estimates. The behavior of the CNR-dependence of the precision will be comparable to that of the rising-falling threshold crossing estimator to be analyzed below.

Noise affects the ability to estimate the time of a threshold crossing. Consider a voltage pulse of rise time τ and peak amplitude A corrupted by noise of root-mean-square (rms) voltage V_0. Consider first the rising portion of the pulse as illustrated in Figure 12-5. Please note that the rise time (from minimum to maximum) is essentially equal to the pulse duration if rise and fall times are equal. If the bandwidth of the noise B is roughly equal to the bandwidth of the signal ($= 1/\tau$) then noise can be considered to roughly constant over the first half of the signal and roughly constant (but different) over the second half. This is an acceptable assumption because the random phasing of changes in noise amplitude is accommodated by our use of only rms values of noise in the equations. The nominally constant noise voltage will add to the leading half of the signal voltage. If an arbitrary threshold has been set somewhere between the minimum and maximum signal values, then

Figure 12-5. Geometric model for analysis of precision in estimation processes.



the noise causes a time shift in the threshold crossing. Since the noise voltage has been assumed to be an rms value, the change in threshold crossing time will also be an rms quantity. From the geometry of Figure 12-5 we can immediately calculate the rms time shift to be

$$\delta T_R \approx \frac{\tau V_0}{A} \tag{12.23}$$

Now consider the falling threshold crossing. Obviously, the mathematical result for the rms time shift will be the same. However, because we have assumed a particular phase for the noise that lead to a constant value over the rising half of the pulse, the time shift for the falling half of the pulse will be uncorrelated with the time shift for the rising half. The rms error in estimating the center of the pulse by adding the times and dividing by two will therefore depend on the root-sum-square (rss) of the two threshold crossing time shifts.

$$\delta T = \frac{\left(\delta T_R^2 + \delta T_F^2\right)^{1/2}}{2} \approx \frac{\delta T_R}{\sqrt{2}} \approx \frac{\tau V_0}{\sqrt{2}\,A} \tag{12.24}$$

If we recall the definition for the carrier-to-noise ratio of a simple radar (from the previous chapter)

$$CNR = A^2 / V_0^2 \tag{12.25}$$

then the rms centroid error can be written as

$$\delta T \approx \frac{\tau}{\sqrt{2CNR}} \approx \frac{1}{B\sqrt{2CNR}} \approx \frac{\Delta T}{\sqrt{2CNR}} . \tag{12.26}$$

The last equality in Eq. (12.26) uses the basic equivalence of the pulse duration τ and the resolution ΔT using a small-dip resolution criterion such as the Gaussian FWHM criterion. There is no other fundamental connection between resolution and precision.

Using Equation (12.26) as the basis for an analogy, we may calculate the precision with which a number of other parameters can be estimated. The results are compiled in Table 12-1. In almost every instance we find that mathematically (if not physically) precision (δx) can be related to resolution (Δx) by the square root of *CNR*.

$$\delta x \approx \frac{\Delta x}{\sqrt{2CNR}} . \tag{12.27}$$

This is an extremely useful result, because it is one of the few equations that links system level performance parameters (δx and Δx) to a component level performance parameter (*CNR*).

Table 12-1. Precision (δx) versus resolution (Δx) for a number of commonly used parameters.

Pulse Amplitude (*A*)	$\delta A \approx \frac{\overline{A}}{\sqrt{CNR}}$
Target Cross Section (σ)	$\delta\sigma \approx \frac{\overline{\sigma}}{CNR}$
Target Range (*R*)	$\delta R \approx \frac{c\tau}{2\sqrt{2CNR}} \approx \frac{c}{2B\sqrt{2CNR}} \approx \frac{\Delta R}{\sqrt{2CNR}}$
Target Angle (θ)	$\delta\theta \approx \frac{\lambda}{D\sqrt{2CNR}} \approx \frac{\Delta\theta}{\sqrt{2CNR}}$
Target Velocity (*v*)	$\delta v \approx \frac{\lambda}{2\tau\sqrt{2CNR}} \approx \frac{\lambda B}{2\sqrt{2CNR}} \approx \frac{\Delta v}{\sqrt{2CNR}}$
Target Acceleration (*a*)	$\delta a \approx \frac{\lambda}{2\tau^2\sqrt{CNR}} \approx \frac{\lambda B^2}{2\sqrt{CNR}} \approx \frac{\Delta a}{\sqrt{2CNR}}$

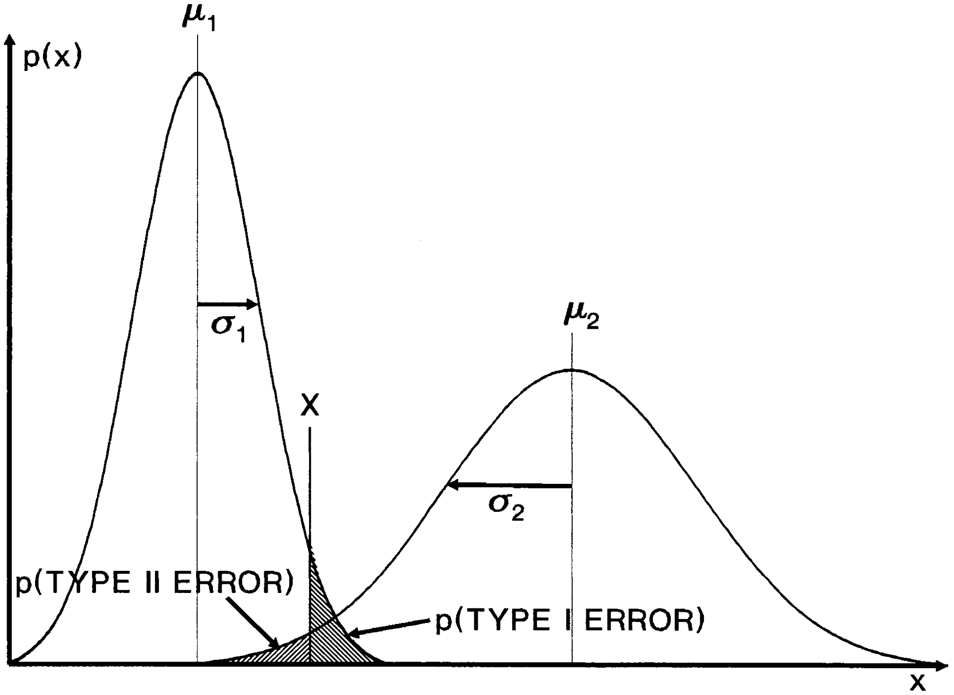
Discrimination

The term **discrimination** has many meanings. One of the common uses is the determination of whether an object is a true target or a decoy based on estimates of some of the object's parameters. For example, decoys tend to decelerate faster than true targets. Thus, an estimate of acceleration might be used to declare the object as target or as decoy. The problem with this simple approach is that there is seldom a clear-cut dividing line between the properties of decoys and the properties of targets. Some targets might decelerate faster than some decoys do. Furthermore, we cannot measure the deceleration with perfect precision. Measurement precision will give a spread to the values in a class even if there were normally not such a spread in the absence of measurement "noise". Any overlap between values that the two classes may possess makes the assignment to a specific class subject to errors. Let us consider two classes each having a distribution of allowable values of some parameter (such as deceleration) as shown in Figure 12-6. Assume that the distributions of allowable parameter values are Gaussian (the normal probability distribution). That is, the probability of having value *y* is given by

$$p(y) = \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-(y-\mu_n)^2/2\sigma_n^2} \tag{12.28}$$

μ_n is the mean of the distribution for class *n* and σ_n^2 is the variance of the distribution of that class.

Figure 12-6. Errors in discrimination between two object classes.



Two kinds of errors can be made in assigning a target to one of two class. Objects in Class 1 can be mistakenly assigned to Class 2 (called a Type I error) and objects in Class 2 can be mistakenly assigned to Class 1 (called a Type II error). If a discriminant “function” is picked such that the class is assigned depending on the value of that function relative to the estimated target parameter, we have the basis for an automatic discrimination system. In the case of our one-dimensional example above, the discriminant function is a threshold value y = X. If the estimated parameter value is greater than X, the object is assigned to Class 2. If the estimated parameter value is less than X, the object is assigned to Class 1. For this discriminant function, the probability of making a Type I error is given by the integral

$$p_{TYPE\ I}\left(X\right)=\frac{1}{\sqrt{2\pi\sigma_1^2}}\int_X^{\infty}dy\ e^{-\left(y-\mu_1\right)^2/2\sigma_1^2} \tag{12.29}$$

while the probability of making a Type II error is given by the integral

$$p_{TYPE\ II}\left(X\right)=\frac{1}{\sqrt{2\pi\sigma_2^2}}\int_{-\infty}^{X}dy\ e^{-\left(y-\mu_2\right)^2/2\sigma_2^2}\ . \tag{12.30}$$

These equations can be cast in slightly simpler form by transforming the integration variable from y to $z = y/\sigma_n$

$$p_I\left(X\right)=\frac{1}{\sqrt{2\pi}}\int_{\frac{X-\mu_1}{\sigma_1}}^{\infty}dz\ e^{-z^2/2}=Q\left(\frac{X-\mu_1}{\sigma_1}\right) \tag{12.31}$$

and

$$p_{II}\left(X\right)=\frac{1}{\sqrt{2\pi}}\int_{\frac{\mu_2-X}{\sigma_2}}^{\infty}dz\ e^{-z^2/2}=Q\left(\frac{\mu_2-X}{\sigma_2}\right). \tag{12.32}$$

where we have defined the integral over the tail of the normal probability distribution as $Q(x)$

$$Q\left(X\right)=\frac{1}{\sqrt{2\pi}}\int_X^{\infty}dz\ e^{-z^2/2} \tag{12.33}$$

Note that we have also used the symmetry of the integral about the mean value μ_2 to make the forms of the integral equal in Equations (12.31) and (12.32).

If Type I and Type II errors are equally bad then the optimum solution is to choose X such that $p_I(X)$ and $p_{II}(X)$ are equal. This minimizes the total error probability. In this instance we have

$$\frac{X-\mu_1}{\sigma_1}=\frac{\mu_2-X}{\sigma_2} \qquad \rightarrow \qquad X_{50}=\frac{\mu_2\sigma_1+\mu_1\sigma_2}{\sigma_1+\sigma_2}. \tag{12.34}$$

There are instances in which Type I and Type II errors are not equally bad and there are other instances in which they are equally bad. For example, it is much more serious to classify an incoming nuclear reentry vehicle as a decoy (and ignore it) than it is to classify a decoy as a reentry vehicle (and attack it) if there is an ample supply of interceptors available. However, if there is only one interceptor available per incoming reentry vehicle, then either type of error becomes equally bad (whether you ignore the real warhead or attack the decoy and preclude attacking a real warhead you still have a warhead that is not attacked).

If we substitute the expression for X_{50} into the probability expressions we find that

$$p_I(X_{50})=p_{II}(X_{50})=Q\left(\frac{\dfrac{\mu_2\sigma_1+\mu_1\sigma_2}{\sigma_1+\sigma_2}-\mu_1}{\sigma_1}\right)=Q\left(\frac{|\mu_2-\mu_1|}{\sigma_1+\sigma_2}\right) \tag{12.35}$$
$$\equiv Q(K)$$

where we have defined a factor K such that

$$K=\frac{|\mu_2-\mu_1|}{\sigma_1+\sigma_2}. \tag{12.36}$$

Since this factor relates the separation of the means in terms of the standard deviations, it is clearly related to discriminability. The discriminability factor (K) used here differs slightly from that used in other less general work by roughly a factor of 2. This other work [6] defines K using only one of the standard deviations (or assumes the two standard deviations are equal). In that work K' (the prime is added to distinguish this result from that of Eq. (12.36) and is absent in the original) is given by

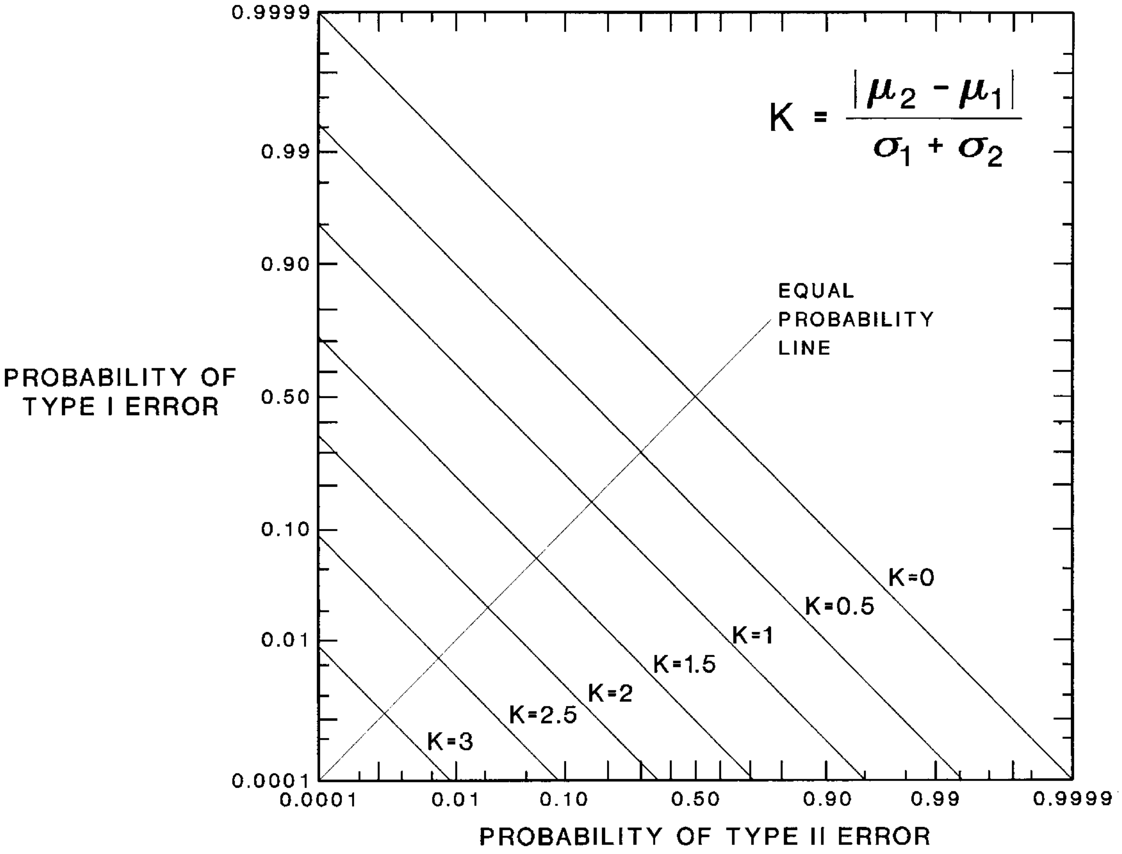
$$K'=\frac{|\mu_2-\mu_1|}{\sigma}. \tag{12.37}$$

If the standard deviations are roughly equal, the data in Figure 12-7 may be used by assuming

$$K'=K/2. \tag{12.38}$$

For example, the K = 1.5 line corresponds to K' = 3.0 in the earlier work.

Figure 12-7. Relation of Type I and Type II errors to the discriminability factor K.



In Figure 12-7 we have plotted the values of $p_I(X)$ and $p_{II}(X)$ for several half-integer values of K. The lines of -45° slope show the probability variations for specific values of K. These show how Type I errors may be traded against Type II errors and vice versa. The line of +45° slope is the line for equal probability, i. e., $p = Q(K)$. A value of $K \geq 1.5$ will keep the probabilities of both kinds of error less than 10% while a value of $K \geq 2.5$ will keep the probabilities of both kinds of error below 1%.

The analysis above is one-dimensional, that is, a single parameter is used to discriminate between the classes. This analysis can be extended to discrimination based on multiple parameters, although we will not do so here. In some instances, an appropriate linear combination of the parameters can be chosen that reduces the multi-dimensional discrimination problem back to a one-dimensional problem and the preceding analysis can be used. When the number of parameters being used for discrimination and/or the number of classes becomes large, the process becomes what is commonly called **statistical pattern recognition**.

Pattern Recognition

The principal components of the statistical pattern recognition process are summarized in Figure 12-8. Sensor data are processed through a variety of operations to extract key elements of information called **features**. Parameter estimation plays a role in establishing values for many features or the components from which a feature is calculated. Features may be related to physical characteristics – brightness, color, size, weight, etc.; waveform characteristics – pulse duration, pulse repetition frequency, modulation type, jitter in pulse duration, jitter in pulse repetition rate, etc.; or image characteristics – area, circumference, aspect ratio, texture, moments of the intensity distribution, etc. The collection of values corresponding to all (total number of features = N) of the identified features forms a **feature vector** in an N-dimensional **feature space**. When numerous known examples from the various target classes (total number of classes = M) are observed and processed, the feature vectors from all examples of a target class tend to cluster in a region of feature space that is different from the regions where examples from other target (or decoy or clutter) classes tend to cluster. Through the process of **training** the classifier (through repeated observations of known targets, decoys, and clutter objects), the feature space can be divided into at least M regions separated by N-dimensional surfaces called **decision surfaces**. Ideally all objects belonging to one class would fall into one and only one of the M regions and all objects belonging to other classes would fall into other regions. An unknown object would then be assigned a class designation based on which region of feature space into which its feature vector fell. This clear-cut situation is illustrated in the left-hand part of Figure 12-9. Unfortunately the real world is seldom so simple. Objects belonging to one class do tend to cluster in feature space, but the N-dimensional probability distribution will have some overlap with the probability distributions corresponding to other objects. The similarity to the two-class discrimination problem presented in the preceding section should be obvious. In the training mode each decision surface is adjusted to minimize the total number of errors of Type I and Type II for every pair of target classes. In a multiple class problem it is assumed that no one class is more important than any other class. The decision surfaces may be piecewise planar or quadratic (paraboloidal) in nature. If piecewise planar decision surfaces are assumed, then pair-wise between classes, the problem usually reduces to the simple discrimination problem. The analysis for quadratic decision surfaces is somewhat more complicated and is reserved for texts on pattern recognition.[7]-[9]

Figure 12-8. The operation of a statistical pattern recognizer. The classification step has a training mode in which known inputs are processed and decision surfaces established and an assignment mode in which unknown objects are classified based on their feature vectors and the decision surfaces.

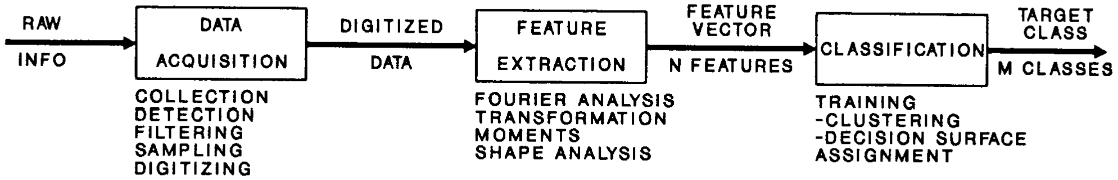
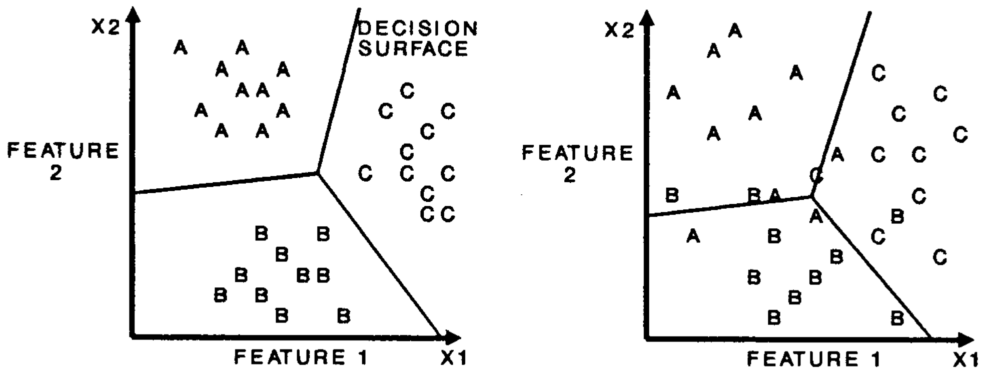


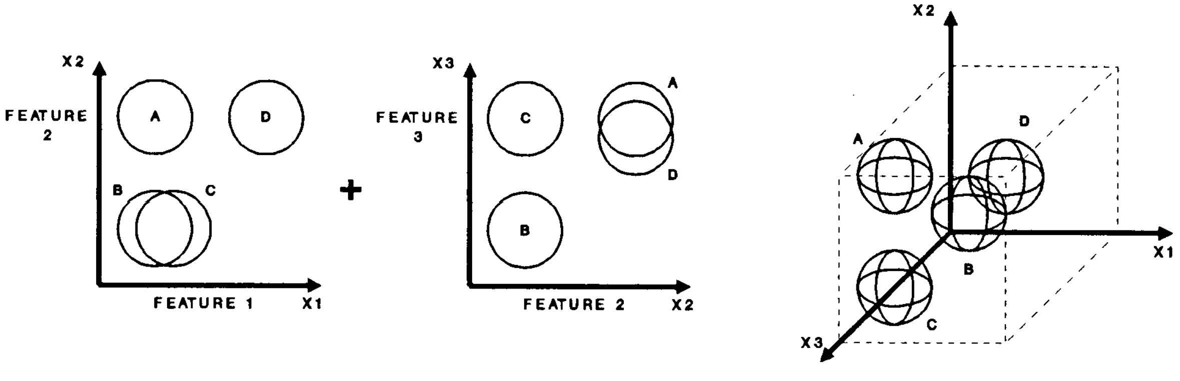
Figure 12-9. Ideal and realistic clustering of objects in feature space with decision surfaces located as indicated.



It might seem to be a given that it would be easy to define enough features to permit target classification with minimal errors. As hinted at in Figure 12-10, there is a fundamental belief that if *N-1* features do not provide the ability to truly discriminate between classes, then adding the N^{th} feature may finally permit the necessary discrimination. This has fueled the development of larger and larger classifiers capable of handling expanded feature spaces. It has fanned the creative spark in defining ever more clever features to be extracted. It has also led to the expenditure of vast sums to assemble multisensor systems capable of providing still more features to incorporate into the feature space.

Unfortunately reality has not always followed prediction.[10] Although improvements in automatic target recognition have occurred, few systems have performance adequate for incorporation into deployed systems. There are a number of reasons for this. First, not all features provide the high degree of clustering desired. The features may lack discriminability. It is possible to define mathematical features (for example the eleventh moment of an intensity distribution) that may have no connection to any physically observable property. It may have more correlation with sensor noise

Figure 12-10. How adding an additional feature may permit discrimination between classes.



than with some fundamental target property. Other features are overly sensitive to environmental or operational changes. For example, the location and intensity of the brightest spot in a thermal image may depend on whether the sun is shining, the direction of the solar illumination, whether the engine is running, or how many rounds its main weapon has fired. Such sensitivity significantly broadens a potentially useful feature. Lastly, there may not be sufficient information present in the data to permit adequate discrimination. As is discussed in the next chapter, there appears to be a correlation between extractable information content in "images" and the ability of humans to perform perceptual functions such as target classification. If the signal used as the input to the classifier does not contain enough information to permit classification to be accurately performed, then no amount of processing or no number of additional features will ever create the missing information. To date, very little study of the information content provided to target recognizers has been performed to determine how well they should perform given the inputs they are being provided.

References

[1] Institute of Electrical and Electronics Engineers, IEEE Standard Dictionary of Electrical and Electronics Terms 4th Ed., ANSI/IEEE STD 100-1988 (IEEE, New York NY, 1988).

[2] Born, Max and Wolf, Emil, Principles of Optics 3rd Ed. (Cambridge University Press, Cambridge UK, 1980), pp.395-398.

[3] J. H. Shapiro, B. A. Capron, and R. C. Harney, "Imaging and target detection with a heterodyne-reception optical radar", *Applied Optics*, 20, #19, pp. 3292-3313, (1 Oct. 1981).

[4] Shapiro, J. H., Dardzinski, V. E. and Tung, E. W., “Coherent laser radar antenna patterns and mixing efficiencies”, Appendix A in “Coherent laser radar remote sensing”, M. I. T. Lincoln Laboratory Report ESD-TR-83-238 (September 1983).

[5] Taub, Herbert and Schilling, Donald L., Principles of Communication Systems 2nd Ed. (McGraw-Hill, New York NY, 1986).

[6] Anonymous, “Discrimination Introduction”, Nichols Research Corporation Briefing Book (14 August 1984).

[7] Bow, Sing-Tze, Pattern Recognition (Marcel Dekker, New York NY 1984).

[8] Watanabe, Satosi, Pattern Recognition: Human and Mechanical (John Wiley & Sons, New York NY, 1985).

[9] Young, Tzay Y. and Fu, King-Sun, Handbook of Pattern Recognition and Image Processing (Academic Press, San Diego CA, 1986).

[10] Harney, Robert C., “Practical Issues in Multisensor Target Recognition” in Sensor Fusion III, R. C. Harney (Ed.), *Proceedings of the SPIE*, 1306, 105-114 (1990).

Problems

12-1. Compare accuracy, precision, and resolution.

12-2. You are the marketing representative for a radar manufacturer. You wish to express the angular resolution of your radar in the best possible light. What resolution criterion are you likely to use to calculate the angular resolution? Give both the criterion name and the explicit expression for the resolution.

12-3. A sensor has a point source response that is triangular. That is,

$$G(x) = 1-(|x|/x_0) \qquad \text{for } |x| \le x_0$$
$$= 0 \qquad \text{for } |x| > x_0$$

Evaluate the resolution function for this point spread function and develop an expression for the Rayleigh criterion of this point spread function in terms of x_0.

12-4. If an angular resolution criterion is assumed to have the form α = x.xx λ/D, what are reasonable values of x.xx (pick a specific criterion) for:

1) a thermal imager
2) a microwave radar
3) a laser radar
4) a television imager
5) a passive sonar
6) a RF interferometer
7) a laser rangefinder?

EXTRA CREDIT

8) an active sonar
9) image-intensified optics
10) a synthetic aperture radar

12-5. How is the precision of a measurement related to the resolution or the characteristic value of the parameter being measured?

12-6. A spectrum analyzer designed to measure the frequency of a radio emission has a frequency resolution $\Delta\nu$. What expression describes the frequency measurement precision $\delta\nu$ of this spectrum analyzer?

12-7. An application requires angular resolution better than 0.8 λ/D and an even finer precision of 0.01 λ/D. Which sensor defined by what resolution criteria can achieve this: i) a laser radar with Sparrow criterion, ii) a passive infrared detection system with Sparrow criterion, or iii) a laser radar with Rayleigh criterion. What is the relation between resolution and precision? What must be done to the sensor chosen above to satisfy both criteria?

12-8. A new antiship missile defense system is required to track missiles with an active 300 GHz seeker head. The system requires an angular resolution of less than 0.005 radians. If two incoming missiles are side-by-side and separated by 35 meters at what range will they become resolved. Utilizing the Rayleigh criterion determine the effective antenna size. Additionally, the system requires an angular precision of 0.0005 radians. Estimate the CNR.

12-9. A millimeter-wave and a microwave radar are candidates for the tracking radar of a new class of ship. The millimeter-wave radar operates at 220 GHz and has an antenna size of 55 cm. The microwave radar operates at 17 GHz and has an antenna size of 2.3 m. What is the maximum range at which each candidate can discriminate between two targets 20 m apart. Assume a Rayleigh criterion for resolution and that atmospheric effects and radar power limitations are negligible. If the millimeter-wave radar has 10 dB less CNR at its maximum range for resolution of 20 m separation targets than the microwave radar at its maximum resolution range , what is the relative precision of the radars at those ranges.

12-10. What effect do the following have on sensor angular resolution? on sensor angular precision? Assume only the parameter listed is allowed to change.

a) increasing carrier frequency
b) increasing CNR
c) increasing signal bandwidth (with no increase in signal power)

Calculate the aperture required for a 10.6 μm laser radar to achieve 0.00001 radians resolution assuming the Gaussian FWHM criterion.

12-11. Two objects are to be discriminated on the basis of temperature. Object 1 is described by a distribution with μ_1 = 270 K and σ_1 = 5 K. Object 2 is described by a distribution with μ_2 = 300 K and σ_2 = 10 K. What is the best error performance that can be expected?

12-12. A pattern recognizer has a feature space in which the region for one target class (Class A) is bordered by regions for five other classes. Assuming the features are completely independent and each pair of features has 5% Type I and 5% Type II errors, what is the probability of correct classification and the probability of false alarms for objects of Class A given that Class A objects are 5 times as numerous as the objects of each of the other five classes. The probability of correct classification is the number of A objects correctly classified as A objects divided by the total number of A objects. The probability of false alarm is the number of objects falsely classified as class A divided by the total number of objects classified as A objects.

CHAPTER 13

MODULATION AND DEMODULATION

Modulation and Demodulation

Information is often carried by modulations (distinct variations of some wave characteristic such as frequency or amplitude) imposed on sinusoidal waves of otherwise constant character. The waves upon which the modulation is imposed are called **carrier** waves. Without modulation the carrier is usually constant and non-varying. The signal is contained within the modulation imposed upon the carrier. It is this association with constant non-varying waves from which the term "carrier-to-noise ratio" used in Chapter 11 derives. The term **modulation** refers to both the process of imposing variations (modulation) on a carrier and to the resulting variation themselves.[1], [2] **Demodulation** is the process of extracting the variations from the received radiation and converting them into signals useable by the sensor system. **Modulators** impose modulation on carriers; **demodulators** extract modulation from modulated carriers.

Modulation and demodulation are seldom end functions of sensors but one or the other or both are often critical intermediates in the accomplishment of desired end functions such as parameter estimation or tracking. In some instances the sensor will perform both modulation and demodulation. For example, a radar may transmit a frequency-modulated waveform and receive a modified frequency-modulated waveform in the pursuit of estimating the range and velocity to a target. In others, one sensor will establish a modulation for another sensor to utilize, such as a coded laser designator signal for a laser guided weapon. In others the interaction of a sensor signal with the target may produce a modulation that is useful for another function, as in a conical scan tracking system.

Many types of modulation can be imposed on either acoustic or electromagnetic waves. These include: **amplitude modulation**, **frequency modulation**, and **phase modulation**. A fourth kind of modulation, **polarization modulation** (switching back and forth between two orthogonal states of polarization) can be imposed only on electromagnetic waves. Amplitude, frequency, and phase modulation are shown graphically in Figure 13-1. Amplitude, frequency, phase, or polarization modulation may be pure **analog** techniques, in which the signal is continuously variable in both time and "amplitude", or they may be implemented as **discrete** techniques in which the signal is **sampled** in time and possibly the "amplitude" is **quantized** into discrete allowable values. Discrete techniques must be sampled, but do not have to be quantized. A continuum of outputs is possible. Techniques in which both sampling and quantization are used are called **digital** techniques. The discrete levels may or may not be encoded into binary or other formats for processing by electronic computers. Figure 13-2 compares an analog signal with one of many digital representations of that signal. The

Figure 13-1. Some signal modulation schemes.

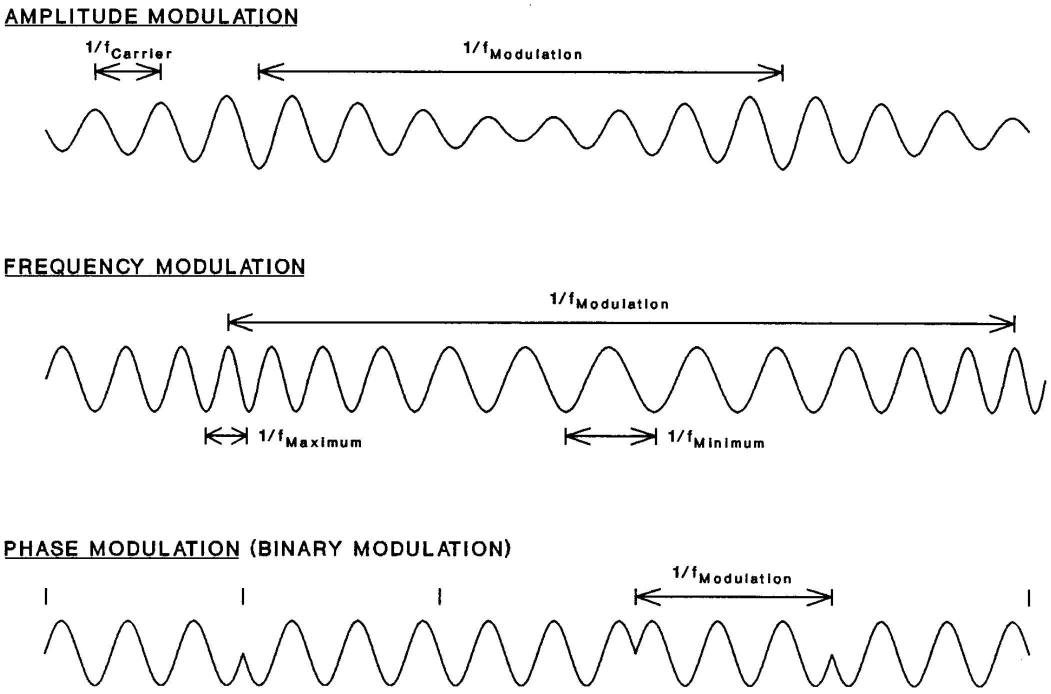
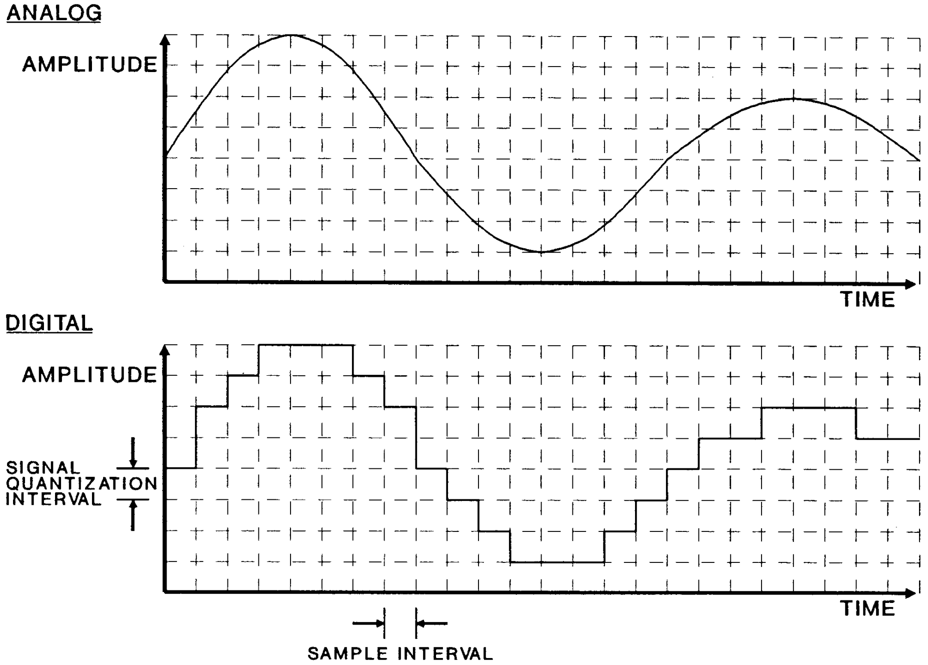


Figure 13-2. Comparison of "Analog" versus "Digital" signals.



key difference between analog and digital lies in the continuous versus the doubly discrete character of the signals being handled.

Amplitude modulation, phase modulation, and frequency modulation are subtly related. All three involve time-dependent variations of the signal. We can represent a general signal as

$$S(t) = A(t)\cos[\Phi(t)] = A(t)\cos[\omega_C t + \phi(t)] \tag{13.1}$$

where $A(t)$ represents any amplitude variation, $\Phi(t)$ is the total phase of the signal, ω_C is the carrier frequency of the signal, and $\phi(t)$ is variable part of the phase. The instantaneous frequency of this signal may be defined as

$$\omega(t) = \frac{\partial \Phi}{\partial t} = \omega_C + \frac{\partial \phi}{\partial t}. \tag{13.2}$$

Let us now attempt to modulate this signal by a modulation function $m(t)$. If

$$A(t) = k\, m(t), \tag{13.3}$$

where k is a scaling factor, then we have **amplitude modulation**. Alternatively, if we allow the variable phase to follow the modulation function, i.e.,

$$\phi(t) = \phi_0 + k\, m(t), \tag{13.4}$$

where ϕ_0 is an arbitrary initial phase and k is a scaling factor, then we have **phase modulation**. The instantaneous frequency is

$$\omega(t) = \omega_C + k\frac{\partial m(t)}{\partial t} \tag{13.5}$$

Consequently, phase modulation implies frequency variations. For this reason, phase modulation is sometimes mis-identified as frequency modulation. If we integrate the modulation function over time to produce an altered modulation function $m'(t)$

$$m'(t) = \int_{-\infty}^{t} d\tau\, m(\tau) \tag{13.6}$$

and let the variable phase follow this integrated modulation,

$$\phi(t) = k\, m'(t) \tag{13.7}$$

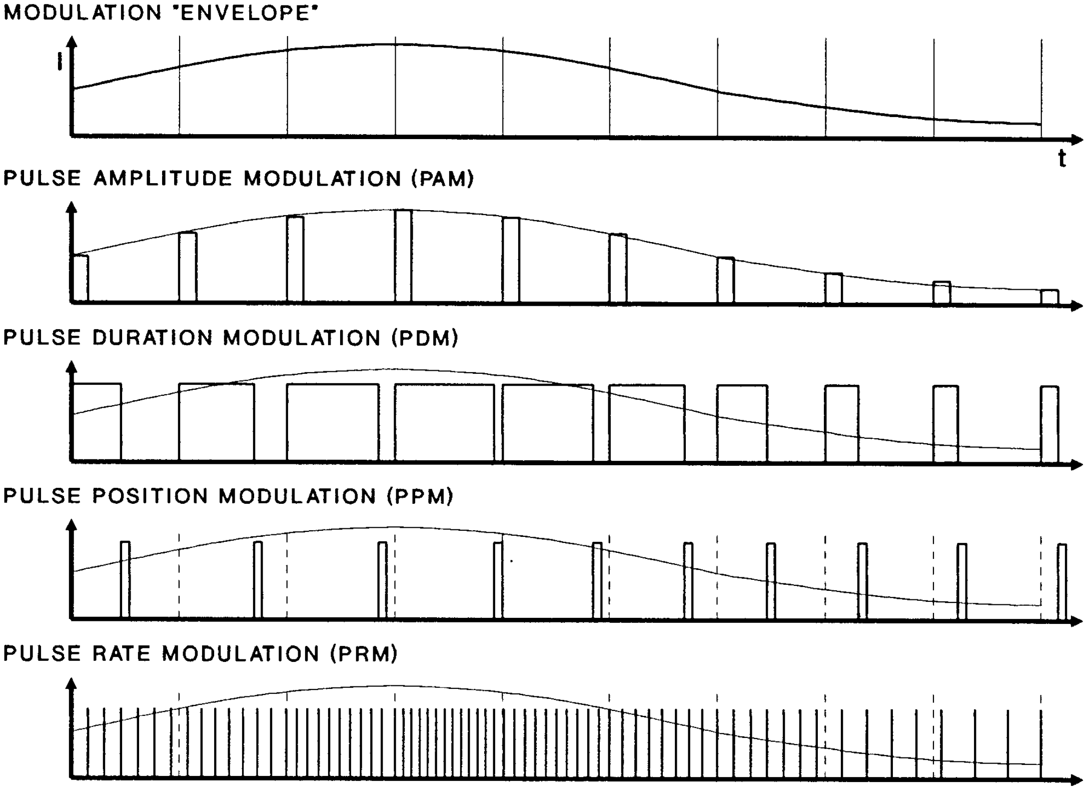
then the instantaneous frequency is found to be

$$\omega(t) = \omega_C + k\frac{\partial m'(t)}{\partial t} = \omega_C + k\,m(t) \tag{13.8}$$

The frequency varies directly with the modulation function. This is true **frequency modulation**. The difference between frequency and phase modulation is essentially a single integration of the modulation function.

Another class of modulation, called **pulse modulation**, is applicable in sampled waveform scenarios. The information in a sampled signal can be represented (or encoded) by varying the amplitude, duration, position (relative to a time reference), or repetition rate of the transmitted pulses. The way in which information is encoded by several types of pulse modulation are illustrated in Figure 13-3. The most common forms of pulse modulation are essentially forms of amplitude modulation. However, frequency or polarization modulation could also be used to encode the signal amplitude modulation.

Figure 13-3. Several pulse modulation schemes.


MODULATION "ENVELOPE"
I
t
PULSE AMPLITUDE MODULATION (PAM)
PULSE DURATION MODULATION (PDM)
PULSE POSITION MODULATION (PPM)
PULSE RATE MODULATION (PRM)

Related to but not a true form of pulse modulation is **pulse code modulation (PCM)**. PCM is a truly digital modulation technique. In PCM the signal is sampled, the sampled signal is quantized into two more discrete levels, and the quantized level is transmitted as a unique string of bits, the bits having been encoded into pulses. Pulse amplitude, position, duration, pulse rate, frequency, polarization, phase, and polarity among others are capable of being used to encode the bits.

In the following sections each of the major modulation schemes will be examined in more detail. The nature of the modulation is discussed as are any peculiarities or distinctive characteristics it may possess. Modulation and demodulation hardware is also discussed.

Amplitude Modulation (AM)

Amplitude modulation (AM) is obtained by multiplying a low frequency modulation onto an otherwise constant carrier wave. The modulation can be achieved by varying the gain or loss of a component in the electronics. An amplitude modulated signal can be expressed as

$$E(t) = A(t)E_0 \cos(2\pi f_C t) \tag{13.9}$$

where $A(t)$ is the modulation waveform and f_C is the carrier frequency. We assume that $0 \le A(t) \le 1$. If the modulation is a simple sinusoid, then the modulation factor can be written as

$$A(t) = 0.5\big(1 + \alpha \cos(2\pi f_M t)\big) \tag{13.10}$$

where $0 \le \alpha \le 1$, is the depth of modulation and f_M is the modulation frequency. Expansion of the terms in Eq. (13.10) yields

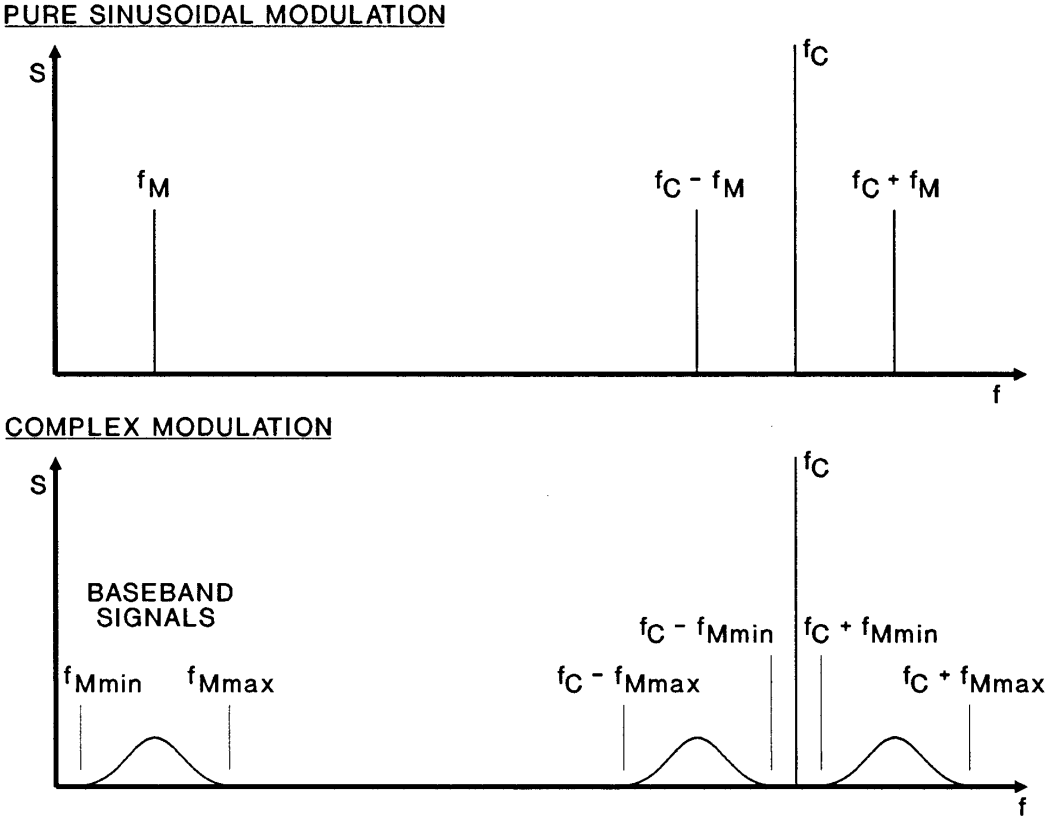
$$\begin{aligned} E(t) &= 0.5\big(1 + \alpha \cos(2\pi f_M t)\big)E_0 \cos(2\pi f_C t) \\ &= 0.5E_0\big[\cos(2\pi f_C t) + \alpha \cos(2\pi f_M t)\cos(2\pi f_C t)\big] \\ &= 0.5E_0\begin{bmatrix} \cos(2\pi f_C t) + \\ 0.5\alpha\big[\cos(2\pi(f_C - f_M)t) + \cos(2\pi(f_C + f_M)t)\big] \end{bmatrix} \end{aligned} \tag{13.11}$$

Thus, as shown in the upper trace of Figure 13-4, the spectrum of the amplitude-modulated signal consists of the carrier and two equal side-bands shifted from the carrier by $\pm f_M$. The amplitude of the sidebands is proportional to the depth of modulation but is never greater that half the carrier amplitude. The spectrum of the original baseband signal (the modulation) is shown superimposed on the AM signal at the lefthand edge of the trace. The lower trace in Figure 13-4 shows the baseband and AM signal spectra in the event of more complex modulation.

The total bandwidth of an AM signal is thus easily seen to be twice the maximum modulation frequency in the signal

$$B = 2 f_{M\,\max} \,. \tag{13.12}$$

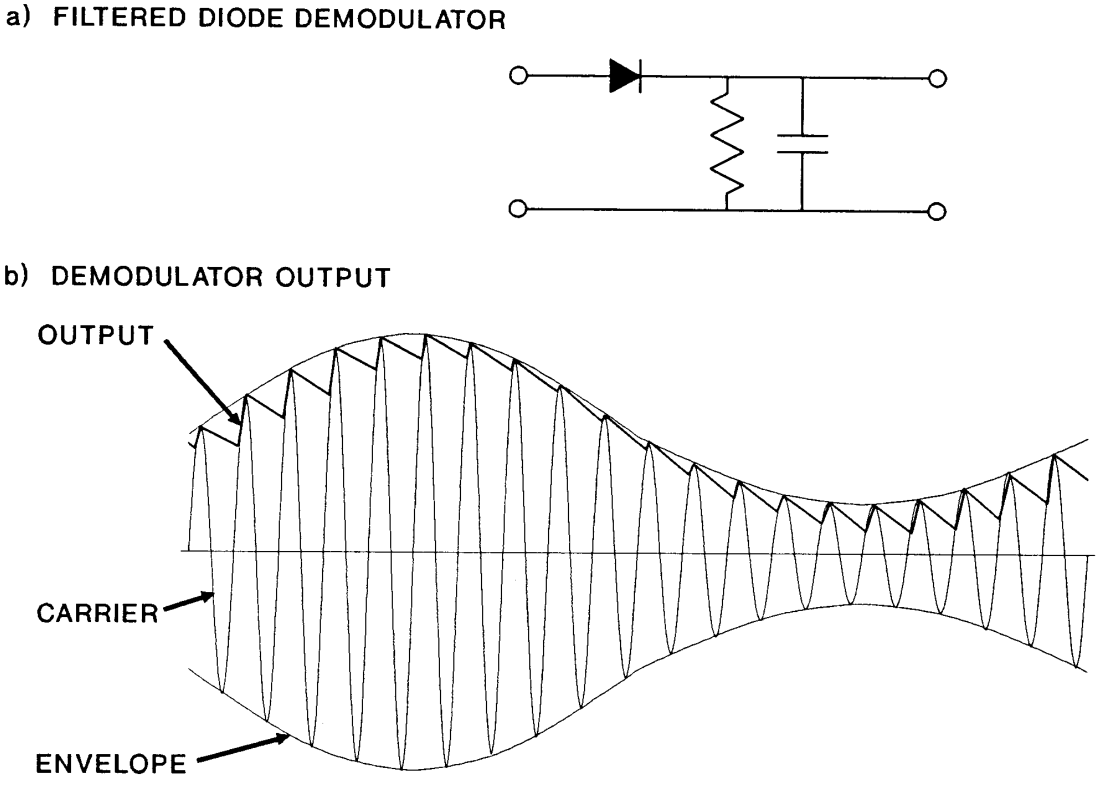
Figure 13-4. Spectra of baseband and amplitude-modulated signals for pure sinusoidal modulation and for more complex modulation.



Amplitude modulation can be detected (demodulated) by mixing the modulated signal with a local oscillator that is identical to the unmodulated carrier. This process is called superheterodyne reception. The "dc" or baseband component of the mixer output voltage will reproduce the modulation. The higher frequency components which are present in the mixer output are usually easily filtered to give only the modulation waveform $A(t)$ as the final output. A **filtered diode** is also capable of serving as a demodulator. Such an arrangement is shown in Figure 13-5. The modulated carrier is passed through a simple diode (half-wave rectifier) and then filtered by an RC filter. The capacitor charges up to the signal voltage during the positive half-cycle, and decays during the negative half-cycle. The RC time constant is set to provide an intercycle decay at least as large as the maximum rate of change of the AM envelope function. The sawtooth voltage across the capacitor is a reasonable representation of the original modulation waveform.

There are some applications where the presence of the carrier in the transmitted signal is not desired. By using a **balanced modulator**, a signal with both sidebands present, and little or no carrier can be produced. This is called **double-sideband suppressed-carrier (DSB-SC)** operation.

Figure 13-5. The RC-filtered diode demodulator and an illustration of its principle of operation.

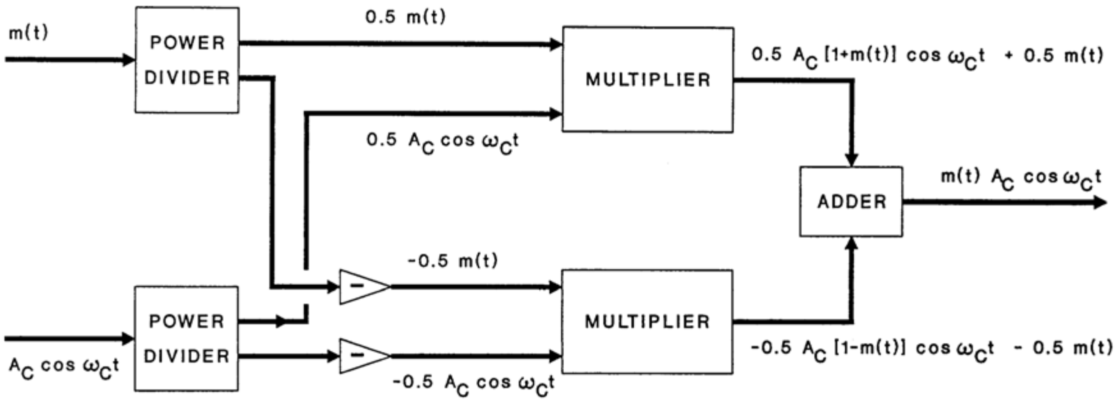


a) FILTERED DIODE DEMODULATOR

b) DEMODULATOR OUTPUT

The balanced modulator is shown in Figure 13-6. By adding a standard modulated signal to the modulated signal produced by first inverting the carrier and the modulation, the carrier can be canceled out without affecting the sidebands.

Figure 13-6. The balanced modulator.



m(t)
POWER DIVIDER
0.5 m(t)
MULTIPLIER
0.5 A_C [1+m(t)] cos $\omega_C t$ + 0.5 m(t)
0.5 A_C cos $\omega_C t$
ADDER
m(t) A_C cos $\omega_C t$
-0.5 m(t)
−
POWER DIVIDER
MULTIPLIER
-0.5 A_C [1-m(t)] cos $\omega_C t$ - 0.5 m(t)
A_C cos $\omega_C t$
−
-0.5 A_C cos $\omega_C t$

Communication applications often prefer to cancel both the carrier and one of the sidebands. This **single-sideband (SSB)** mode of operation frees up frequency bandwidth that can be used for additional communication channels. Single sideband is usually produced by bandpass filtering to eliminate both the carrier and the unwanted sideband. This requires a bandpass filter with very sharp frequency limits and strong out-of-band rejection. It may be impractical to generate a desired SSB signal at a specific frequency because of limitations imposed by filter technology.

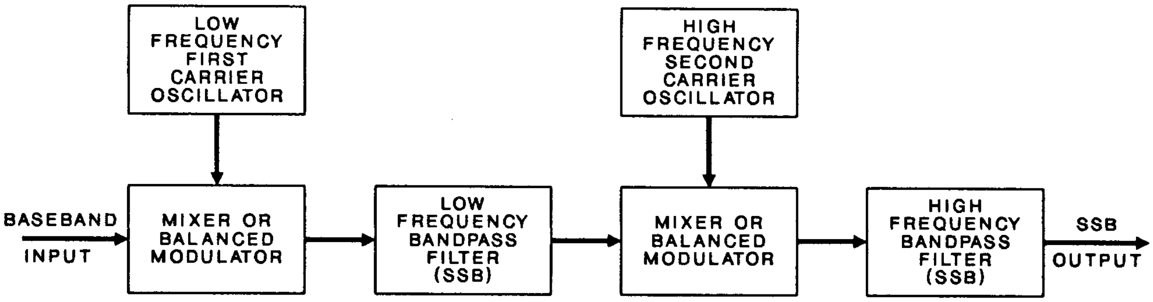
NOTE: it is a general rule that the smaller the ratio of the width of a filter passband is to the center frequency of that pass band, the more difficult (and usually more expensive) it is to make that filter. This true regardless of frequency, whether radio frequency, microwave, infrared, or visible radiation. Filters with ratios of 1% to 10% are relatively easy to obtain. Filters with ratios of 0.1% are more difficult and filters with ratios < 0.01% are exceedingly difficult (but seldom impossible).

For example, filters may be available that can separate a 10 kHz baseband modulation from a 500 kHz carrier. However, filters may not be available that can separate the 10 kHz modulation from a 10 MHz carrier. However, it is likely that filters are available that can separate a 500 kHz modulation from a 10 MHz carrier. SSB generation can then be implemented in two steps: modulation of a 500 kHz carrier followed by filtering plus subsequent modulation of a 10 MHz carrier followed by filtering. This is shown in Figure 13-7. Using up to three stages of modulation plus filtering SSB signals can be easily produced in almost frequency range.

The baseband signal (frequency ω_M) may be recovered from SSB modulation (either $\omega_C + \omega_M$ or $\omega_C - \omega_M$) by mixing the SSB signal with the carrier signal (frequency ω_C). The output of the mixer contains a second harmonic term (either $2\ \omega_C + \omega_M$ or $2\ \omega_C - \omega_M$) and the desired baseband term (either $+\omega_M$ or $-\omega_M$). The baseband term is then isolated by filtering.

In general, AM is extremely simple to produce and simple to demodulate. However, noise on an amplitude modulated signal is indistinguishable from signal modulation. Amplitude fluctuations caused by variations in the propagation channel (such as atmospheric turbulence or multipath interference) likewise are indistinguishable from signal modulations. For this reason, despite its simplicity, AM is usually used only when high signal levels and insignificant fluctuations are expected.

Figure 13-7. Multistage generation of a single-sideband signal.


LOW FREQUENCY FIRST CARRIER OSCILLATOR
HIGH FREQUENCY SECOND CARRIER OSCILLATOR
BASEBAND INPUT
MIXER OR BALANCED MODULATOR
LOW FREQUENCY BANDPASS FILTER (SSB)
MIXER OR BALANCED MODULATOR
HIGH FREQUENCY BANDPASS FILTER (SSB)
SSB OUTPUT

Frequency Modulation (FM)

Frequency modulation (FM) involves imposing small changes in the instantaneous frequency of a carrier wave. This can be accomplished by varying the frequency of an oscillator to produce the original frequency modulation and then mixing this signal with an unmodulated carrier or by directly frequency modulating the carrier using a frequency (or phase) modulator. Since the instantaneous frequency is modulated, there is a maximum frequency $f_{Maximum}$ and a minimum frequency $f_{Minimum}$. The modulation frequency $f_{Modulation}$ is essentially the inverse of the time required to shift from maximum to minimum frequency. An important quantity in discussing FM systems is the **depth of modulation** β of the signal. This is defined by the relationship

$$\beta = \frac{\Delta f}{f_{Modulation}} \equiv \frac{\Delta f}{f_M} = \frac{\left| f_{Maximum} - f_{Minimum} \right|}{f_{Modulation}} \tag{13.13}$$

If β is large enough, the carrier may not be readily distinguishable. It can be found by the relation

$$f_{Carrier} \equiv f_C = \frac{f_{Maximum} + f_{Minimum}}{2} . \tag{13.14}$$

Consider the following simple sinusoidal modulation function

$$m(t) = \beta f_M \cos(2\pi f_M t) \tag{13.15}$$

where f_M is the modulation frequency. If we consider next a simple sinusoidally frequency-modulated signal. The instantaneous frequency is given by

$$f(t) = f_C + \beta f_M \cos(2\pi f_M t) \quad . \tag{13.16}$$

In frequency modulation the phase function is proportional to the time integral of the modulation function

$$\begin{aligned} \phi(t) &= 2\pi \int_{-\infty}^{t} d\tau \left[f_C + \beta f_M \cos(2\pi f_M \tau) \right] \\ &= 2\pi f_C t + 2\pi \beta \sin(2\pi f_M t) \end{aligned} . \tag{13.17}$$

Thus, the signal field can be described by the expression

$$E(t) = E_0 \cos\left(2\pi f_C t + \beta \sin(2\pi f_M t)\right) . \tag{13.18}$$

This signal field can in turn be expressed as

$$E(t) = E_0 \begin{bmatrix} \cos(2\pi f_C t)\cos(\beta\sin(2\pi f_M t)) \\ -\sin(2\pi f_C t)\sin(\beta\sin(2\pi f_M t)) \end{bmatrix}. \tag{13.19}$$

The two terms involving β can be expanded in a Fourier series in which f_M is the fundamental frequency and recombined with the carrier frequency to give

$$\begin{aligned} E(t) = E_0\Big[& J_0(\beta)\cos(2\pi f_C t) \\ & - J_1(\beta)\big[\cos(2\pi(f_C - f_M)t) - \cos(2\pi(f_C + f_M)t)\big] \\ & + J_2(\beta)\big[\cos(2\pi(f_C - 2f_M)t) + \cos(2\pi(f_C + 2f_M)t)\big] \\ & - J_3(\beta)\big[\cos(2\pi(f_C - 3f_M)t) - \cos(2\pi(f_C + 3f_M)t)\big] \\ & + \cdots \end{aligned} \tag{13.20}$$

where the $J_n(\beta)$ are Bessel functions. The frequency modulated signal consists of a carrier of relative amplitude $J_0(\beta)$ and a set of symmetrically placed sidebands at frequency separations of nf_M, etc. If β becomes large, then almost all of the energy is contained in the sidebands, and little is contained in the carrier. The effective bandwidth B of the FM signal can be shown to be

$$B = 2(\beta + 1)f_M = 2(\Delta f + f_M) \tag{13.21}$$

a relation known as **Carson's Rule**.

Frequency modulation can be produced in a variety of ways. One of the more common techniques is Edwin Armstrong's (Armstrong was one of the founders of modern radio along with Marconi, de Forest, and Sarnoff). This technique combines an integrator with a phase modulator (see the next section) to produce a frequency modulator. Armstrong's system is shown schematically in Figure 13-8. The mathematics of the signal generation can be found in the next section on phase modulation.

The basic concept of FM demodulation is shown in Figure 13-9. The FM signal is input to a frequency selective network. A frequency selective network is any electronic circuit or element that gives a response (output) that varies with frequency. Almost every real circuit acts as a frequency selective network to some extent. The frequency selective network converts the frequency modulation into an amplitude modulation. The amplitude modulation is then demodulated using any standard AM demodulator such as the diode demodulator.

Figure 13-8. Armstrong frequency modulation technique.

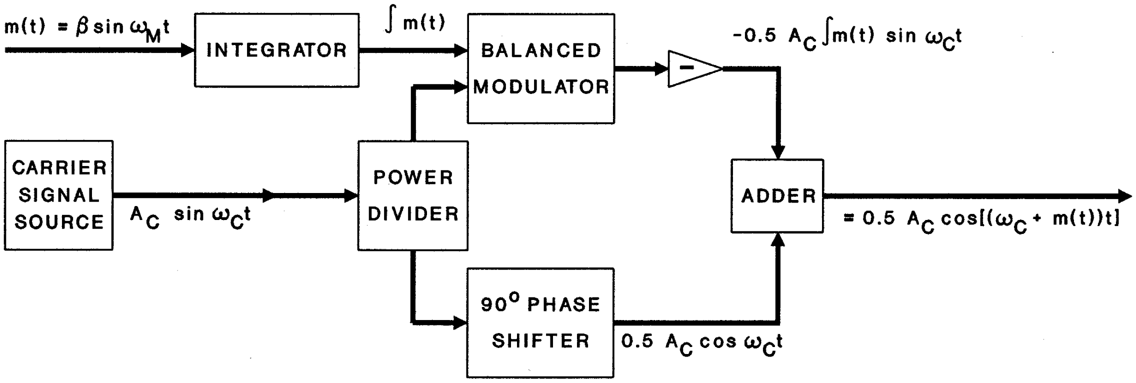
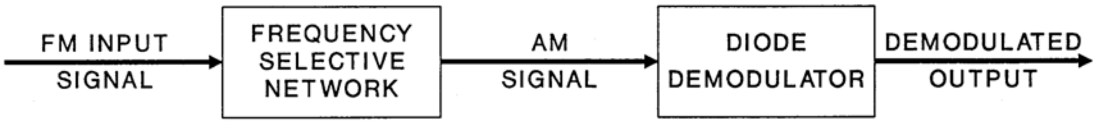
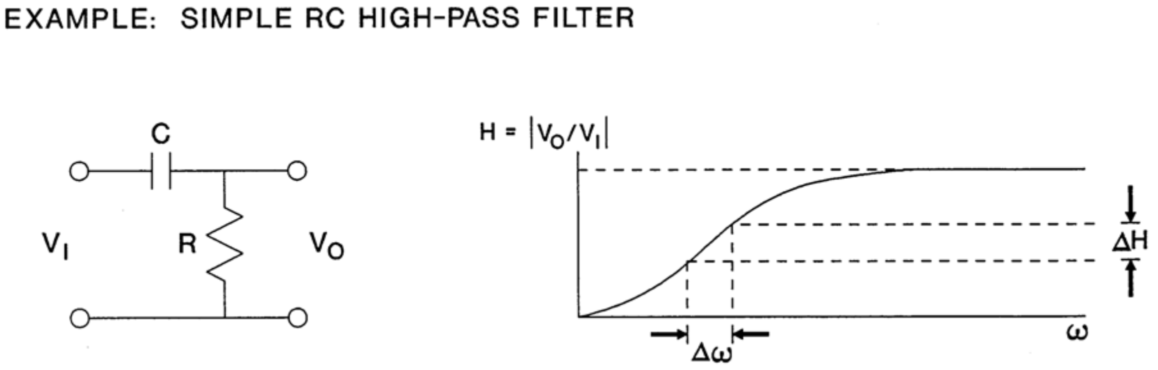


Figure 13-9. Basic principle of FM demodulation.



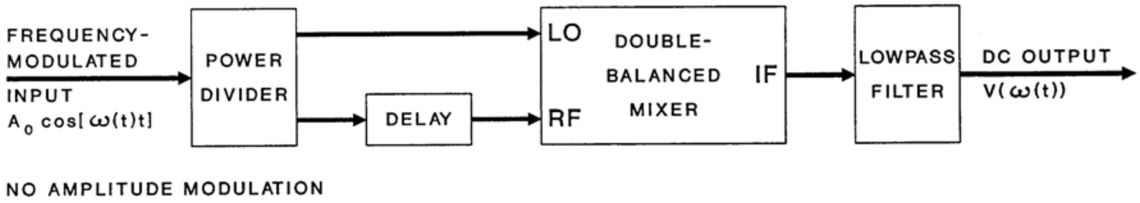


A common form of FM signal demodulator is the limiter-plus-frequency discriminator (commonly called a limiter-discriminator and shown schematically in Figure 13-10). The limiter truncates the peaks of the analog signal to a constant level (eliminating the effects of any amplitude modulation). The frequency discriminator is a mixer-based component whose output voltage is proportional to the input frequency.[3] It converts an FM-modulated amplitude-limited carrier directly into an AM voltage signal that is easily processed. FM signals are strongly resistant to degradation from noise and from channel fluctuations. For this reason it is preferred in systems where low signal-strengths are expected. As long as the *CNR* is much greater than unity, then the signal-to-noise ratio in an FM system is given by

$$SNR = 1.5\beta^2 CNR\,. \tag{13.22}$$

If the depth of modulation is large (greater than unity), the *SNR* can be appreciable even if the *CNR* is relatively small.

Figure 13-10. Frequency discriminator based on a double-balanced mixer.



There is a practical limit to how low *CNR* may go while maintaining the performance of Eq.(13.22). Frequency discriminators and other frequency demodulators exhibit a threshold effect. As the threshold is approached from the direction of high *CNR,* the noise in the output rises rapidly. The noise takes the form of spikes that yield clicks in the frequency demodulator output. As long as the *CNR* is well above (greater than 3 dB) the threshold level, then the frequency demodulator can maintain a lock on the signal. Typical thresholds are of the order of *CNR=10*.

Phase Modulation (PM)

AM and FM can be used to transmit analog or digital information. **Phase modulation (PM)** is usually used to transmit digital information. A phase shifter is used to impose discrete phase shifts of either 0 or 180 degrees onto a carrier at well-defined intervals (time difference equal to one over the modulation frequency). A “0" bit might be represented by 0 degrees absolute phase (no shift) while a “1" bit might be represented by a 180 degrees absolute phase. Differential phase shift modulation is a variant of PM that involves changing the modulation only when bits change from “0" to “1" or from “1" to “0" in a string of digital data. The instantaneous phase of the carrier can be determined by a phase discriminator (a device that works very similarly to a frequency discriminator). Phase modulation is also highly resistant to noise and channel fluctuation effects.

The Armstrong phase modulation technique is shown in Figure 13-11. First, the carrier is split into two equal parts. One part is then modulated by the modulation signal *m*(*t*) using a simple balanced modulator. For carrier signal

$$E(t) = A_C \sin \omega_C t \tag{13.23}$$

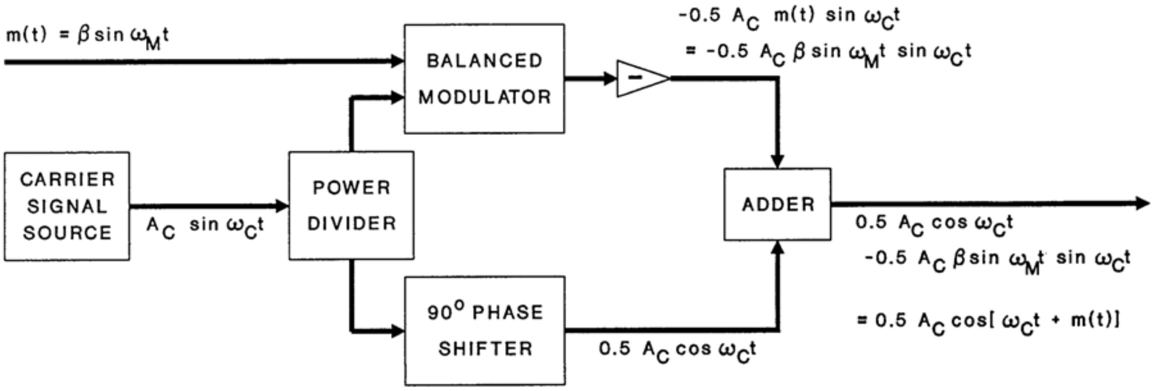
and modulation signal

$$m(t) = \beta \sin \omega_M t \tag{13.24}$$

the output of the balanced modulator is

$$0.5 E(t) m(t) = 0.5 A_C \beta \sin \omega_M t \, \sin \omega_C t \, . \tag{13.25}$$

Figure 13-11. Armstrong phase modulation system.



The second part of the carrier is shifted in phase by 90° and then added to the negative of the output of the balanced modulator. The sum yields

$$\begin{aligned} E_{OUT}(t) &= 0.5A_C \sin\omega_C t - 0.5A_C\beta \sin\omega_M t \sin\omega_C t \\ &= 0.5A_C \cos\left[\omega_C t + \beta \sin\omega_M t\right] \\ &= 0.5A_C \cos\left[\omega_C t + m(t)\right] \end{aligned} \tag{13.26}$$

which is obviously a phase-modulated signal.

Demodulation of phase-modulated signals is accomplished with phase detectors. The simplest phase detector is a doubly-balanced mixer in which the phase-modulated signal is input to the RF port and a signal with the same average frequency but no phase modulation is input to the LO port. The output of the IF port is a signal proportional to the phase difference between the RF and LO signals. In many systems there is an arbitrary total phase accumulation that depends on the exact circumstances of the signal propagation. This phase is not compensated in the simple mixer. However, if a period of known zero phase shift is contained in the signal, then the overall phase can be eliminated.

Since phase and frequency modulation are intimately related it is possible to use a frequency demodulator to recover the phase modulation. If a phase-modulated signal is processed by a frequency demodulator, the output will be proportional to the derivative of the modulation function. Thus, the integral of the output of a frequency demodulator gives the original phase modulation. By adding an integrator to the output of a frequency demodulator we obtain a phase demodulator.

Pulse Modulation

Pulse modulation is the generic class of modulation schemes in which the modulation signal *m*(*t*) to be transmitted is encoded in the form of pulses. The signal is sampled at a regular interval and the measured amplitude of *m* at each sample point is used to select a particular value of some parameter of a train of pulses. For example, the values of *m* may be used to determine the amplitudes of the pulse heights of the pulse train. The frequency of the pulse or the pulse duration or the exact timing of the start of each pulse is irrelevant to the encoded information. Only the pulse amplitude carries information about *m*(*t*). This form of pulse modulation is called **pulse amplitude modulation (PAM)**. NOTE: pulse modulation is a discrete technique in that it requires sampling. However, it may or may not be digital as there is no requirement to quantize the encoding. In the example of pulse amplitude modulation above, there is no constraint placed on the amplitude. It may be continuous and exactly equal to the value of *m*(*t*) at the time of sampling. On the other hand, it may be quantized. This is a choice left to the system designer.

As indicated in Figure 13-3, virtually any parameter of the pulses may be used to encode the data independent of the other parameters. The sampled values of *m*(*t*) may be used to determine the length (duration) of each pulse. This produces what is known as **pulse duration modulation (PDM)** or **pulsewidth modulation (PWM)**. The sampled values of *m*(*t*) may be used to determine the time position (relative to some reference – often the time at which each sample is obtained) at which the pulse is transmitted. The longer the time delay, the larger the amplitude of *m* being represented. This form of pulse modulation is called **pulse-position modulation (PPM)**. The pulse repetition frequency (PRF) of a very high rate pulse train (many pulses are transmitted per sampling interval) may be use to encode the sampled value of *m*(*t*). This form of pulse modulation is called **pulse rate modulation (PRM)**. The four techniques mentioned so far are illustrated in Figure 13-3. There are at least two others that warrant mention but cannot be depicted on the same figure. One of these is **pulse frequency** (or **wavelength) modulation**. In this encoding scheme the frequency of the transmitted radiation making up the pulse is varied in proportion to the sampled value of *m*(*t*). This form of pulse modulation can be considered to a form of sampled frequency modulation. A closely related form of pulse modulation encodes the information on *m*(*t*) into the polarization angle of the radiation transmitted in the pulse. This is called **pulse polarization modulation**.

The author believes this list:

* pulse amplitude modulation (PAM)
* pulse duration modulation (PDM) or pulselength modulation
* pulse-position modulation (PPM) or pulse timing modulation
* pulse rate modulation
* pulse frequency modulation or pulse wavelength modulation
* pulse polarization modulation

to be comprehensive. There are a very limited number of fundamental properties of pulses. One might be tempted to add pulse shape modulation where the shape varies from rectangular to triangular or some other shape as *m* varies. The author thinks this would terribly difficult to implement and is not likely to be attempted. One might consider varying the phase of the transmitted waveform. However, the frequency of radiation in a pulse is easily determined, but phase is not. For example, the phase depends on the separation of the transmitter and receiver.

There is no reference from which the applied phase (proportional to *m*(*t*)) may be determined from the separation phase. In ordinary phase modulation it is changes in phase which are measured. This cannot be done in simple pulses. There are no other unique properties of pulses that the author can identify that could be used to encode the signal *m*(*t*).

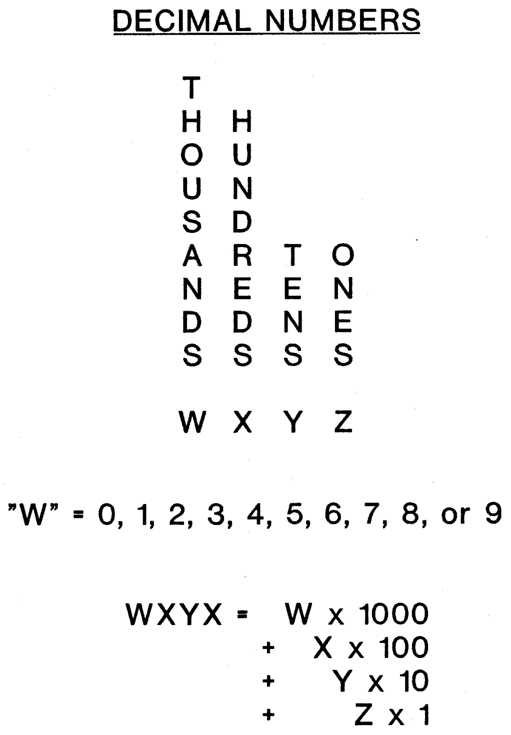
Not all pulse modulation schemes are equal. Some are easier to implement than others. For example, it is easy to delay the emission of a pulse in direct relation to the amplitude of *m*. At the same time, it may be difficult to vary the wavelength or the polarization of the radiation. Conversely, polarization or frequency modulation are not as easily detected as time and amplitude modulation. This may enhance the covertness of a signal. Only those individuals intended to receive the message are likely to have the necessary receivers. Some schemes are more immune to the effects of noise and anomalies in the propagation medium. Pulse amplitude modulation is very susceptible to both noise and propagation effects. Receiver noise will directly increase or reduce the perceived value of *m* because it adds to or subtracts from the pulse amplitude. Propagation anomalies (multipath fading or turbulent scintillation) will cause fluctuation in the received signal strength that will be perceived as real signal fluctuations in *m*. On the other hand, amplitude noise and propagation anomalies will produce almost impact on the pulse repetition rate. Thus, PAM is far more susceptible to noise and anomalies than any of the other techniques.

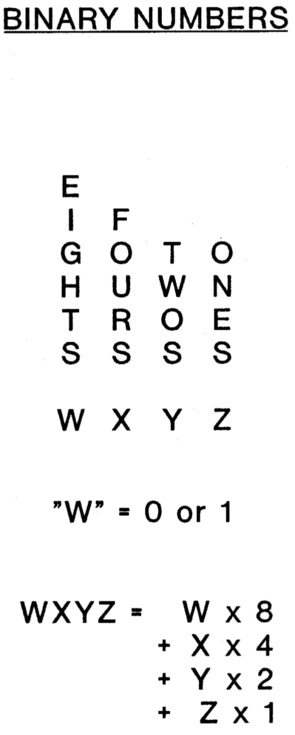
Pulse Code Modulation

Pulse code modulation (PCM) is a truly digital modulation technique. In PCM the signal is sampled, the sampled signal is quantized into two more discrete levels, and the quantized level is transmitted as a unique string of bits, the bits having been encoded into pulses. As we shall see, pulse amplitude, position, duration, pulse rate, frequency, polarization, phase, and polarity among others are capable of being used to encode the bits. PCM is not considered a true form of pulse modulation because the amplitude of the modulating signal is not directly mapped onto some parameter of simple pulses, instead it is mapped into a digital numerical code.

Most PCM encoding techniques use binary numbers to encode the signal. Figure 13-12 compares the decimal number system with which everyone is familiar to the binary number system (which is commonly used in computers and digital communications). In the decimal number system, large numbers are represented by digits. Each decimal digit represents a power of ten. In a four-digit decimal number, the most significant digit represents the number of thousands contained in the number. There are ten possible values *W* for the number of thousands in any number – 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. The next most significant digit represents one power of ten less (hundreds). This digit is the number of hundreds left in the number after all of the thousands have been accounted for. The third most significant digit represents two powers of ten less (tens). The least significant digit

Figure 13-12. Decimal versus binary number systems.



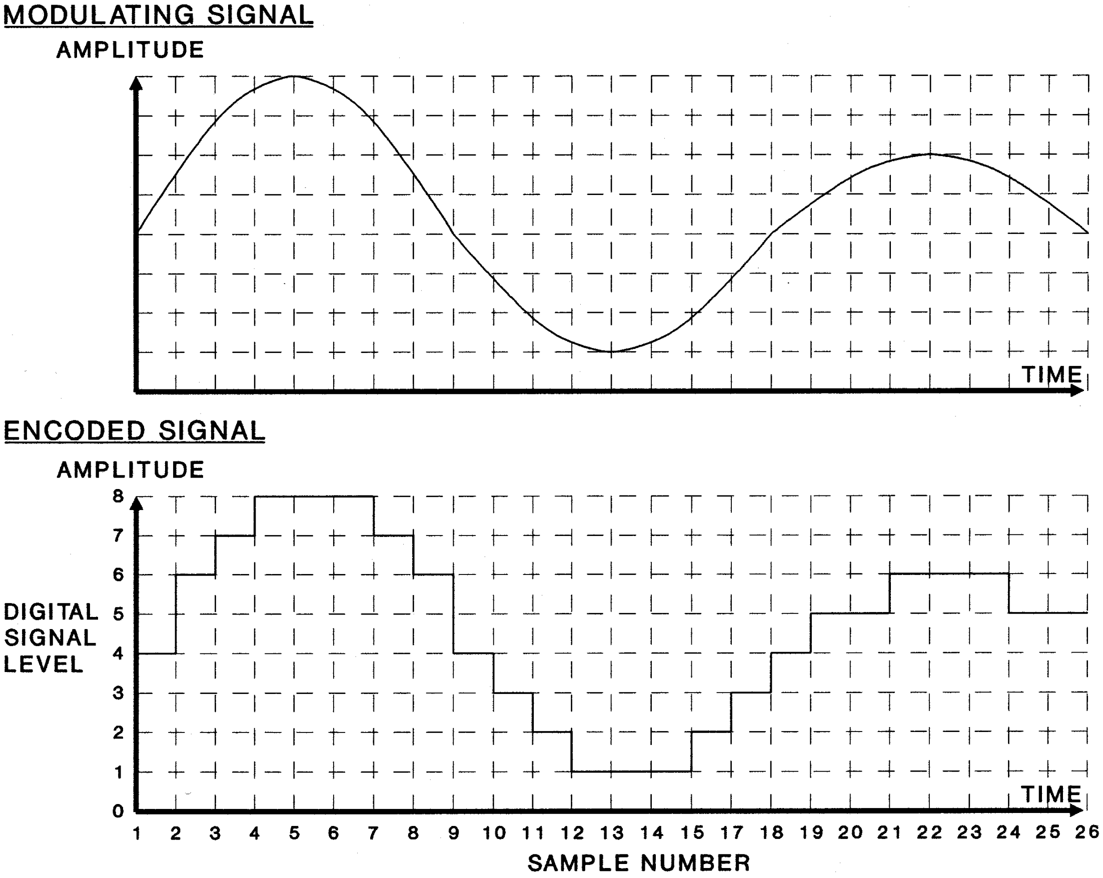


represents the ones. A four-digit decimal number can represent any decimal number between 0 and 9999. Negative numbers can be accommodated by adding a sign. Fractions can be accommodated by adding a "decimal point" and digits which represent tenths, hundredths, thousandths, etc. Decimal numbers of arbitrary size can be generated by adding more digits.

Binary numbers are based on powers of two (as opposed to ten). The binary digits (usually contracted to "bits") represent the number of ones, twos (=2^1), fours (=2^2), eights (=2^3), etc. in the number. There are only two values possible in each bit: 0 or 1. Binary numbers of arbitrary size can be constructed by adding more bits representing still higher powers of two. Negative numbers and fractions can be handled analogous to decimal numbers.

Now consider Figure 13-13. A modulating signal is sampled (at regular intervals in this figure, but irregular sampling is also allowable) and the value at each sample is assigned to the nearest digital signal level. At sample 1, the nearest level is level 4. At sample 2, the nearest level is level 6. The digital signal levels are usually equally spaced (on either a linear or a logarithmic scale) and bear no immediately obvious relationship to the amplitude of the modulation signal. For example, the signal

Figure 13-13. Digital encoding.

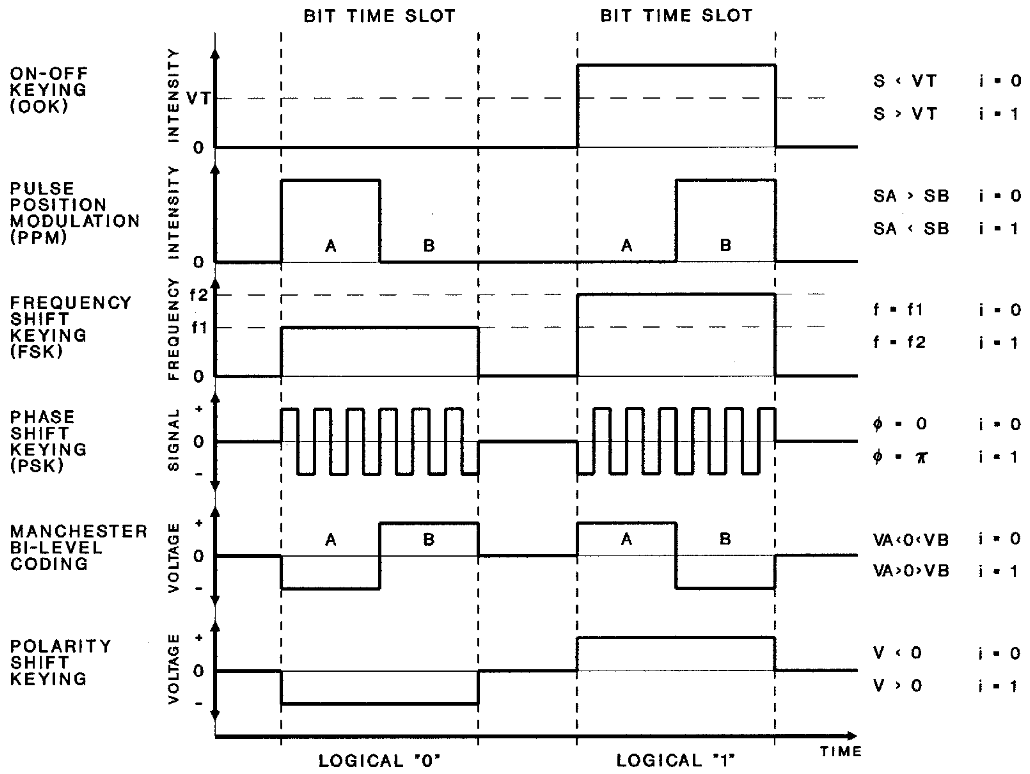


may vary between -10 and -100 volts or between -10 and +10 volts or between -10 and +10 microvolts, yet the same digital levels 0, 1, 2, ... 15 may be used. This is why digitization of a signal is sometimes called **encoding** analogous to the cryptographic practice of assigning some arbitrary symbol or group of letters and numbers to represent a specific letter or number. Here, each real signal level is assigned to a specific digital level (or number).

Digitization is the first step in pulse code modulation. The next step is to convert the digital signal level into a number, usually binary. In the example above, possible digital levels range from 0 to 8. This range can be spanned by four-bit binary numbers. The first value 4 (sample number one) becomes the binary number 0100. The second value 6 (sample number two) becomes the binary number 0110 and so on. Now the binary numbers are translated into pulses which form the pulse code modulation. Binary numbers are usually used because the translation of each bit requires selection between only two possible waveforms. One form of coding might be to represent each bit by a possible pulse. If the pulse is present in a designated time slot, the bit is assumed to be a “1"; if the pulse is absent, then the bit is assumed to be a “0". This is called **on-off keying (OOK)**. In on-off keying, the number four (0100) is represented as NoPulse - Pulse - NoPulse - NoPulse.

Figure 13-14 illustrates a number of the **keying** formats that can be used with pulse code modulation. In usual practice, each bit is assigned a specific time slot. Within that time slot, the

Figure 13-14. Examples of keying (bit encoding) formats



keying technique modulates a pulse or a pulse train in a fashion that the receiver can demodulate and recover the meaning of the original bit (i.e., 0 or 1). The formats shown in Figure 13-14 are not all-inclusive. For example, polarization shift keying (e.g., up linear polarization = 0, sideways linear polarization = 1) is clearly possible but not explicitly shown. Not all keying formats can be applied to all types of signals. For example, polarity shift keying is possible when the transmission system involves voltages transmitted over a wire, but is meaningless when applied to electromagnetic radiation transmitted through space. Similarly polarization shift keying is applicable to electromagnetic radiation but not to voltages on a wire. Some keying techniques such as OOK are universally applicable.

Not all keying formats are equal when it comes to robustness in the presence of receiver noise or transmission anomalies. Let us compare two commonly used techniques, OOK and pulse-position modulation (PPM). In OOK, the optimum demodulator will examine the signal strength *S* in the bit interval (assumed to be time = 0 to time = *T*)

$$S = \int_0^T dt\, s(t) \tag{13.27}$$

and try to determine whether it is closer to “zero” or some “non-zero” value. The received “signal” *s*(*t*) has contributions from the true signal *y*(*t*) and from noise *n*(*t*)

$$s(t) = y(t) + n(t) \tag{13.28}$$

Since the system will always have receiver noise, then $S>0$, even if $y=0$. This bears a strong resemblance to the radar detection problem analyzed in Chapter 11. The optimum decision criterion for OOK is a **threshold test**:

$$S \begin{matrix} \geq \\ < \end{matrix} S_T \qquad \begin{matrix} \text{Choose bit} = 1, & (13.29a) \\ \text{Choose bit} = 0, & (13.29b) \end{matrix}$$

where S_T is chosen to minimize the probability of error.

The **bit error probability** Pr(*e*) is the sum of the probability that a 1 is falsely called a 0 and the probability that a 0 is falsely called a 1

$$\Pr(e) = P(0|1) + P(1|0) . \tag{13.30}$$

Evaluation of the bit error probability will be left to a later volume when communication systems are discussed. At this point we will only say that the threshold is usually adjusted to equalize $P(0|1)$ and $P(1|0)$. This will produce bit error probabilities that are almost if not absolutely minimum. Nevertheless it should be obvious that if sensor noise goes up significantly, then the bit error rate will also go up because $P(1|0)$ will increase. Similarly, if there is significant fading of the signal, then the bit error rate will increase because $P(0|1)$ will increase.

Now consider pulse position modulation. A pulse is always present in one or the other half of the bit time. If we associate a pulse being present in the first half of the bit time with a 0 and a pulse being present in the second half of the bit time with a 1 then we may define two quantities

$$S_0 = \int_0^{T/2} dt\ s(t) \tag{13.31}$$

and

$$S_1 = \int_{T/2}^{T} dt\ s(t) . \tag{13.32}$$

In the absence of noise, a logical 0 should produce a large S_0 and a zero S_1, while a logical 1 should produce a large S_1 and a zero S_0. The presence of noise will cause the "zero" values to assume non-zero levels. The presence of fading will tend to affect both S_0 and S_1 more or less equally. The optimum decision criterion in this instance is a **ratio test**

$$S_0 \geq S_1 \qquad \text{Choose bit = 0,} \tag{13.33a}$$

$$S_0 < S_1 \qquad \text{Choose bit = 1.} \tag{13.33b}$$

Using this test, errors only occur if noise becomes comparable to the signal level and if the fades are so deep as to make the noise significant. Any keying technique that uses a ratio test will have substantially lower bit error rates at a given carrier-to-noise ratio than any technique that uses a threshold test. Similarly, any technique that uses a ratio test will be less susceptible to bit errors produced by fading than any technique that uses a threshold test. Although OOK is the simplest pulse code keying scheme, it is seldom used because of noise and fading problems. PPM, PSK, or FSK all provide superior performance because it is possible to use ratio tests in the decision process.

Delta modulation is an important form of PCM. Figure 13-15 shows the modulator and demodulator portions of a delta modulation transmission system. The output Δ(t) of a difference amplifier is modulated by a simple pulse train $p_i(t)$. The output is a train of pulses of constant amplitude but varying polarity. The sign of the output polarity depends only on the sign of the input function Δ(t). Note: in practice any two discernible parameters (polarization, frequency, pulse-position, etc.) could be used to transmit the sign of the difference. The output pulse train is simultaneously transmitted over the communication channel to the receiver and directed into an integrator. The output of the integrator $q(t)$ is subtracted from the input modulation $m(t)$ to generate the difference signal Δ(t). Thus,

$$p_o(t) = p_i(t) \cdot \mathrm{Sgn}(\Delta(t)) \tag{13.34}$$

where Sgn(x) is the signum or sign function and

$$\Delta(t) = m(t) - q(t) = m(t) - \int dt \; p_o(t) . \tag{13.35}$$

The name delta modulation derives from the fact that the transmitted signal is proportional to the difference between the modulation signal $m(t)$ and an approximation to the modulation $q(t)$. Figure 13-16 shows an example of a modulation waveform, its delta function approximation, and the transmitted pulse train.

At the receiver end of the system a quantizer analyzes the signal and determines whether each pulse is positive or negative and generates a pulse train that reproduces the original $p_o(t)$. The quantizer serves to minimize the effects of noise and channel propagation characteristics on the signal. An integrator then adds (or subtracts) the output pulse train. This signal should be the same as the approximate modulation $q(t)$. A filter removes the high frequency components associated with the pulse modulation yielding a smoothed approximation $m'(t)$ to the original modulation signal.

Figure 13-15. A delta modulation transmission and reception system.

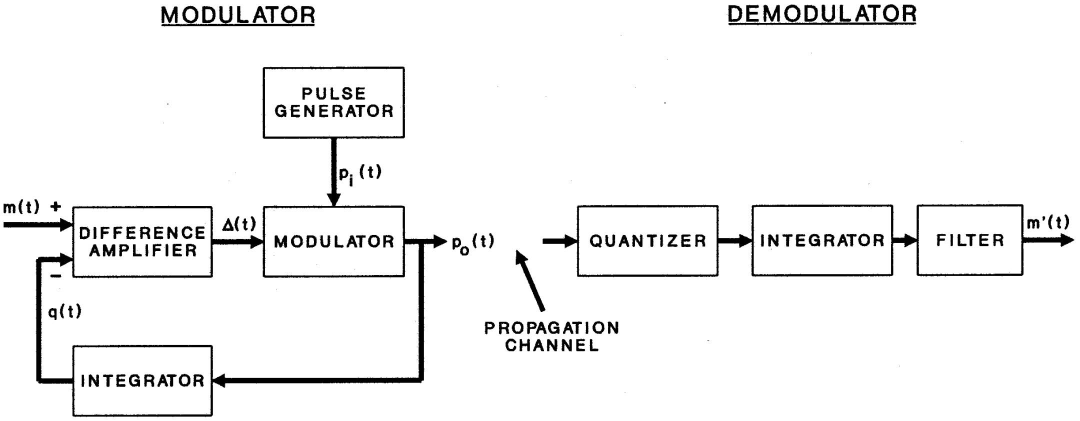
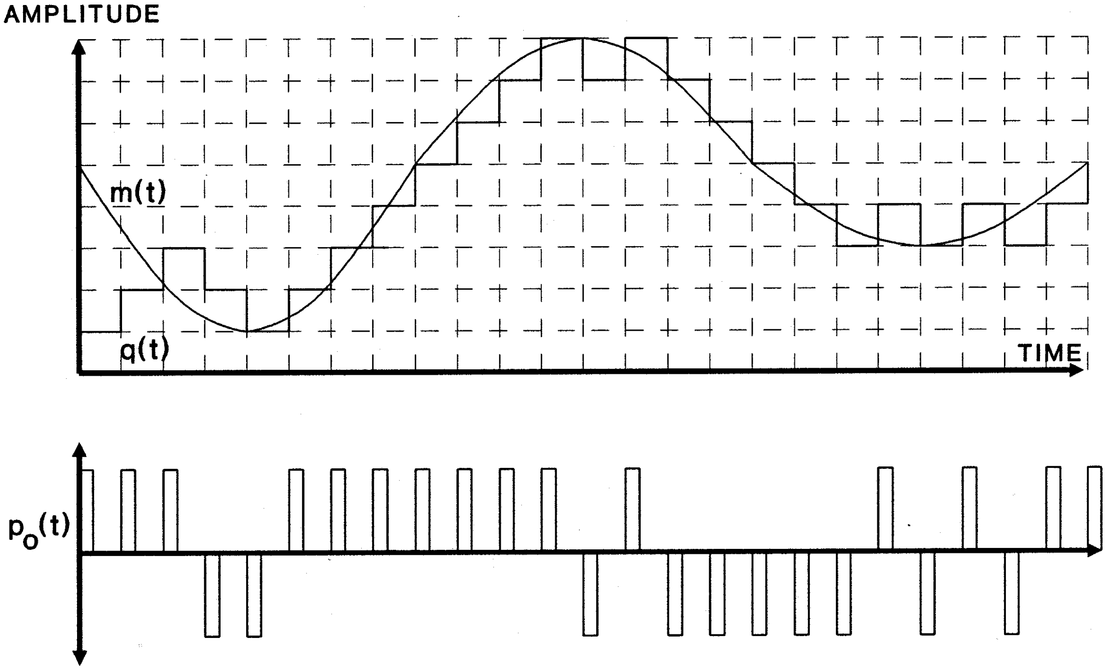


Figure 13-16. The waveforms associated with delta modulation.



References

[1] Taub, Herbert and Schilling, Donald L., Principles of Communication Systems 2nd Ed. (McGraw-Hill, New York NY, 1986).

[2] Straw, R. Dean, Ed., The ARRL Handbook for Radio Amateurs (American Radio Relay League, Newington CT, 1999).

[3] Mini-Circuits, RF/IF Designers Handbook (Scientific Components, Brooklyn NY, 1992).

Problems

13-1. A modulated signal is investigated with a spectrum analyzer. The spectrum consists of a strong spike surrounded by two equal and symmetric sidebands of much lower amplitude. What form of modulation is most likely to be responsible for the observed spectrum?

13-2. A modulated signal is investigated with a spectrum analyzer. The spectrum consists of a series of equally spaced lines of widely varying amplitude. What form of modulation is most likely to be responsible for this observed spectrum?

13-3. Describe in simple mathematical terms the difference between frequency modulation and phase modulation.

13-4. Edwin Armstrong's introduction of frequency modulation led to major improvements in the quality of radio transmissions compared to de Forest's amplitude modulation. Why?

13-5. A binary sequence 1001101011100011 is to be transmitted by a pulse code modulation system. Assume a 1 MHz bit transmission rate with no dead time between bits and 1 μsec bit time. Draw with appropriate scale the transmitted waveform corresponding to this binary sequence if on-off keying is used. On the same drawing draw the transmitted waveform if pulse-position modulation is used.

13-6. An atmospheric propagation channel is subject to multipath fading. Which PCM technique will give better performance, FSK or OOK? Why?

CHAPTER 14

IMAGING AND IMAGE-BASED PERCEPTION

Generalized Images

Nearly everyone has a firm idea of what constitutes an image. Most would describe an image as a two-dimensional array (horizontal by vertical angular dimensions) of brightness, and possibly color, data that is sensed by the eye and interpreted by the brain. This is indeed a form of imagery. A few other individuals would use the same physical description but would include acquisition by sensors such as film, television, or possibly even infrared imagers. In fact, all of these are images, but the aggregate does not approach a more generalized concept of images. The author prefers the more general definition of an **image** as

> *"a distribution in at least two spatial dimensions of one or more parameters related to physical properties of an object or scene."*

In this definition there are many possible spatial dimensions that can be used. There are also many different physical properties that can be distributed over those dimension. Table 14-1 lists a number of these potential spatial dimensions and a number of possible physical properties.

Table 14-1. Potential object characteristics that can form images.

POTENTIAL SPATIAL DIMENSIONS	POTENTIAL PHYSICAL PROPERTIES
AZIMUTH (or BEARING)	COLOR
ELEVATION ANGLE	REFLECTIVITY
RANGE	REFLECTED INTENSITY
CARTESIAN COORDINATES (x, y, z)	RADIANCE
DEPTH (or ALTITUDE)	CONCENTRATION
MAP COORDINATES	TRANSMITTANCE (or ABSORPTANCE)
CROSS-RANGE	VELOCITY
	TEMPERATURE
	RANGE
	RADAR CROSS SECTION

For example, a classic visual image is a distribution of reflected intensity over azimuth and elevation angles. However, a distribution of radar cross section as a function of range and cross-range also forms an image. Such an image could be produced by a synthetic aperture radar (SAR) or by a range-Doppler imaging radar. A distribution of infrared radiance versus Cartesian coordinates (x = along track distance, y = cross-track distance) is another form of image that could be produced by a line scan thermal reconnaissance camera. An AN/KAS-1 sensor essentially produces an image of path-integrated infrared absorptance (of chemical vapors in the air) as a function of azimuth and elevation angle. In essence, any combination of coordinates and physical parameters can form an image.

An image may have more than one physical parameter. An active-passive infrared imager [1] can produce a distribution of range, velocity, reflectivity, and infrared radiance as a function of azimuth and elevation angle. Obviously, such an "image" is not readily understandable by human visual processing – human vision is really only capable of processing roughly four dimensions of data (horizontal, vertical, brightness, and color) not six. However, every sub-image (one parameter versus the spatial dimensions) can be readily viewed and understood.

Resolution versus Coverage

Images contain information. Whether the image is analog or digital, the spatial context is ultimately essentially quantized. We use the term **pixel** (a contraction of "picture element") to denote any one of the discretely-addressable regions in the image space from which an image is constructed by assignment of parameter values. Pixels may represent actual data samples or they may be discrete elements of the display (such as individual cells in a gas discharge (plasma panel) display or individual RGB (red-green-blue) phosphor clusters in a color TV monitor).

Resel is a contraction of "resolution element". A resel is any region of image space whose size in each dimension is equal to the system resolution in that dimension. Resolution may be limited by external factors (e.g., atmospheric turbulence, microwave multipath, etc.), the measurement physics (e.g., pulse duration, beamwidth, collection aperture, etc.), data sampling limitations (e.g., number of detector elements, digitization rate, etc.), or the display (e.g., electron beam spot size greater than the line spacing in a television display, etc.).

An image may have pixels which are smaller than its resels. However, it is arguable as to whether it is really possible to have resels that are smaller than the pixels. One group will argue that the finite spacing of pixels effectively prohibits the acquisition of information with a scale smaller than the pixel spacing. This can be considered to limit the resolution. If one sample is all that is obtained, the author will agree with this first group. A second group argues that if additional looks are obtained with sub-pixel offsets relative to the first, then the aggregate of multiple looks will yield smaller scale information down to the limit of the resel. The author agrees with the assertion but would suggest that the aggregated data has a reduced pixel size. Thus, it is likely that the pixel size must be smaller than or equal to the resel size in any system. There are two common units of resolution that need explanation. The first is **TV line (TVL)**. A TVL is the spacing between resolvable, fully-interlaced raster lines that equals the projected size of a resolution element. That is,

$$1\ TVL = 1\ resel \tag{14.1}$$

The second unit is the **line pair (LP)**. A line pair is the width of a light bar-dark bar pair in a resolution chart that would exhibit a depth of modulation of only 50% when placed near the target and observed by the system. One line pair is the wavelength of one cycle of a light-dark square wave. It is often used interchangeably with the unit called a **cycle**. That is,

$$1\ LP = 1\ cycle = 2\ resels = 2\ TVL \tag{14.2}$$

These units are usually combined with a unit length or unit angle, such as LP/mm, cycles/mrad, TVL/inch when describing the resolution of a system.

When dealing with images and imaging systems, there are several terms that must be understood. These are field of regard, field of view, and instantaneous field of view. The relationship of these terms to each other is illustrated in Figure 14-1. Images are typically generated by scanning a detector array over a limited range of angles. The **field of view (FOV)** is defined by the range of angles that is scanned and/or subtended by the detector array. Each frame of image data covers one field of view. At any given instant of time, the detector elements each look at a subset of the field of view. The **instantaneous field of view (IFOV)** is usually defined as the angular portion of the field of view that can be seen by one detector at any instant in time. Occasionally, the IFOV is defined as that subset of the field of view that is viewed by the entire detector array at any one instant. Context usually removes any ambiguity in the assignment. **Field of regard (FOR)** is the range of angles that can be imaged if the FOV is moved around by a pointing system.

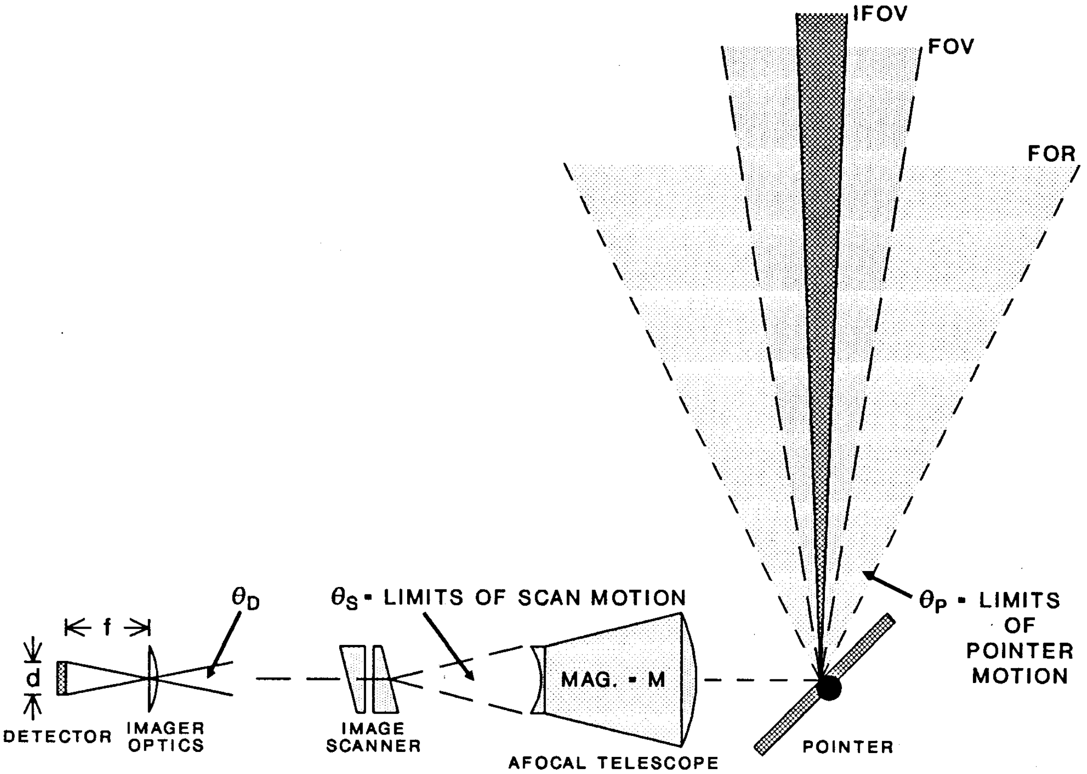
The IFOV can be determined by the dividing the detector dimension *d* by the effective focal length (image focal length *f* times any afocal magnification *M)*

$$\theta_{IFOV} = d / f\,M = \theta_D / M\,. \tag{14.3}$$

The FOV is determined by the limits of the scanner motion, modified by any afocal magnification between the scanner and the output aperture, plus one IFOV,

$$\theta_{FOV} = \left(\theta_S / M\right) + \theta_{IFOV} \approx \theta_S / M\,. \tag{14.4}$$

Figure 14-1. Graphic representation of image terms of reference.



In many instances, where θ_S is large, then θ_{IFOV} can be neglected. The FOR is determined by the limits of full aperture pointer motion plus one FOV,

$$\theta_{FOR} = \theta_P + \theta_{FOV} . \tag{14.5}$$

The preceding equations assume that the optical system does not have any field stops that would artificially limit the "field of view" to smaller values.

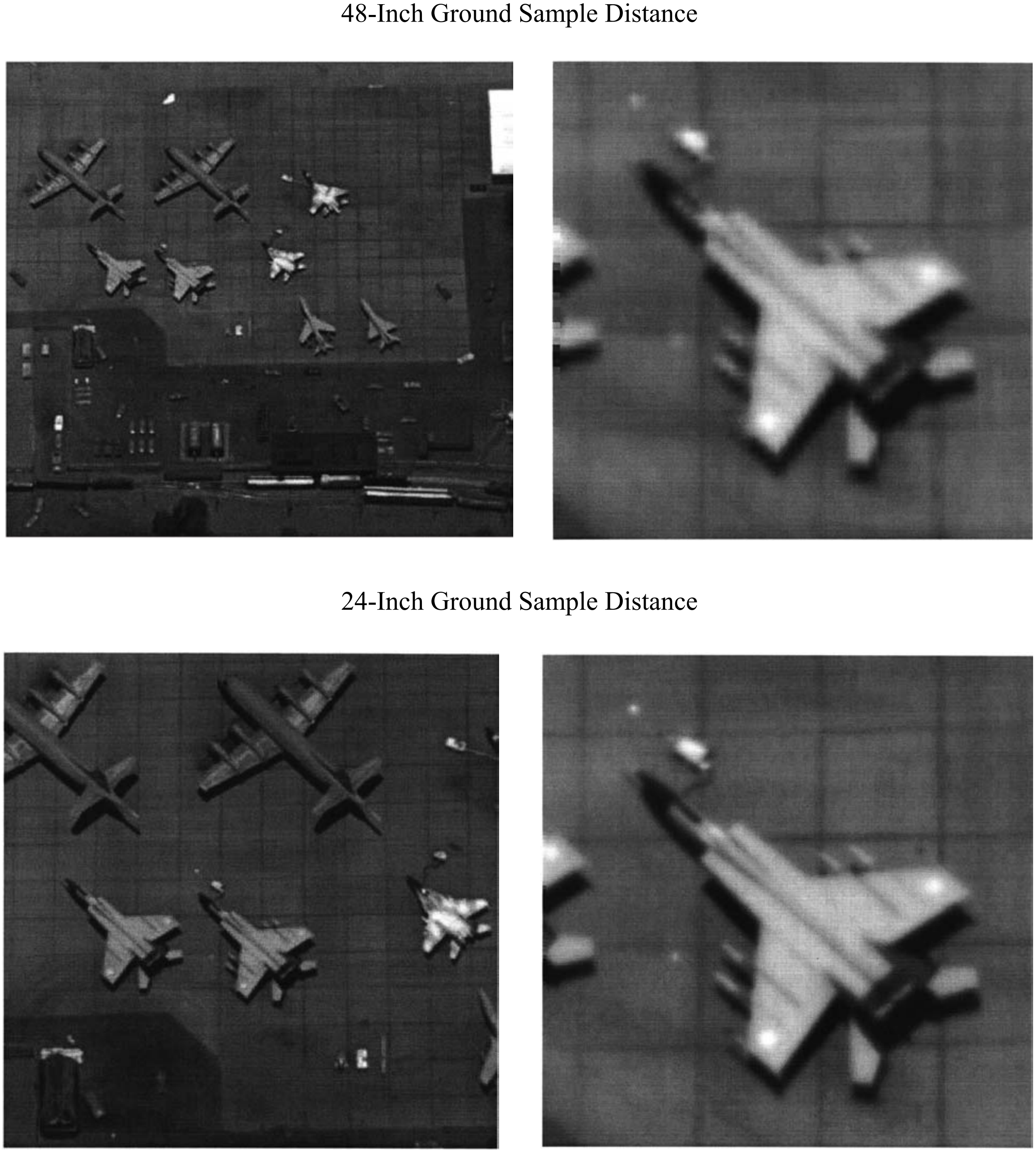
In almost every imaging system there is a tradeoff between FOV and resolution. For example, a modern digital camera may have a focal plane array with 1024 x 1024 detector elements. If this array takes an image with resolution α (in μrad), then the FOV of the system can be no larger than 1024α. Projected to the ground at range R, the resolution covers linear dimension (called the **ground sample distance** – GSD) $l = \alpha R$ and the FOV covers linear dimension $L = N\ \alpha R$, where N is the number of detectors in one dimension of the detector array (= 1024 in the example above). If $N = 1024, \alpha = 10\ \mu rad$, and $R = 100\ km$, then $l = 1\ m$ and $L = 1\ km$. If higher resolution is required, for example, $\alpha = 1\ \mu rad$, then $l = 10\ cm$, but $L = 100\ m$. For a fixed number of detector elements in a staring array, resolution and FOV are proportional quantities. Their ratio is fixed. If one is increased, the other must increase by the same factor. This is illustrated in Figure 14-2. Four images of aircraft on an airfield are presented with ground sample distances of 48 inches, 24 inches, 12 inches, and 6 inches, respectively. The FOV is concomitantly reduced by a factor 2 between each image to maintain the fixed ratio.

The proportionality between FOV and resolution is not just valid for staring array detectors. Consider a imager using a scanned linear detector array. Any detector element must dwell on any pixel in the scene for a minimum period of time in order to acquire sufficient target radiation to be accumulated. The quotient of the number of detectors in the array divided by the dwell time gives the maximum number of pixels per unit time that may be acquired. The area coverage per unit time will be the product of the pixel area and the maximum number of pixels per unit time If high resolution pixels are obtained, then the area corresponding to each pixel will be small, and the area coverage will be small. If low resolution pixels are obtained, then the pixel area will be larger, and the area coverage will be larger.

Even photographic film is subject to a finite product between resolution and area coverage. The resolution of film is determined by the size of the silver halide grains in its emulsion. Each film type will have a specified resolution (usually specified in line pairs/mm). When exposed by a high quality optical system, the resolution will be determined by the film resolution. To obtain higher spatial resolution, the image size at the film must be increased. Since film only comes in fixed sizes, increased image size must come at the expense of reduced field of view.

Thus, it is generally not possible to design an imaging system with arbitrarily large FOV and arbitrarily small resolution. For this reason many imaging systems are provided with switchable optics that can several different fields of view at proportional resolutions. Initial search may be done with a wide FOV (perhaps 250 mrad at 0.25 mrad resolution) while target recognition is done with a narrow field of view (perhaps 25 mrad at 25 μrad resolution).

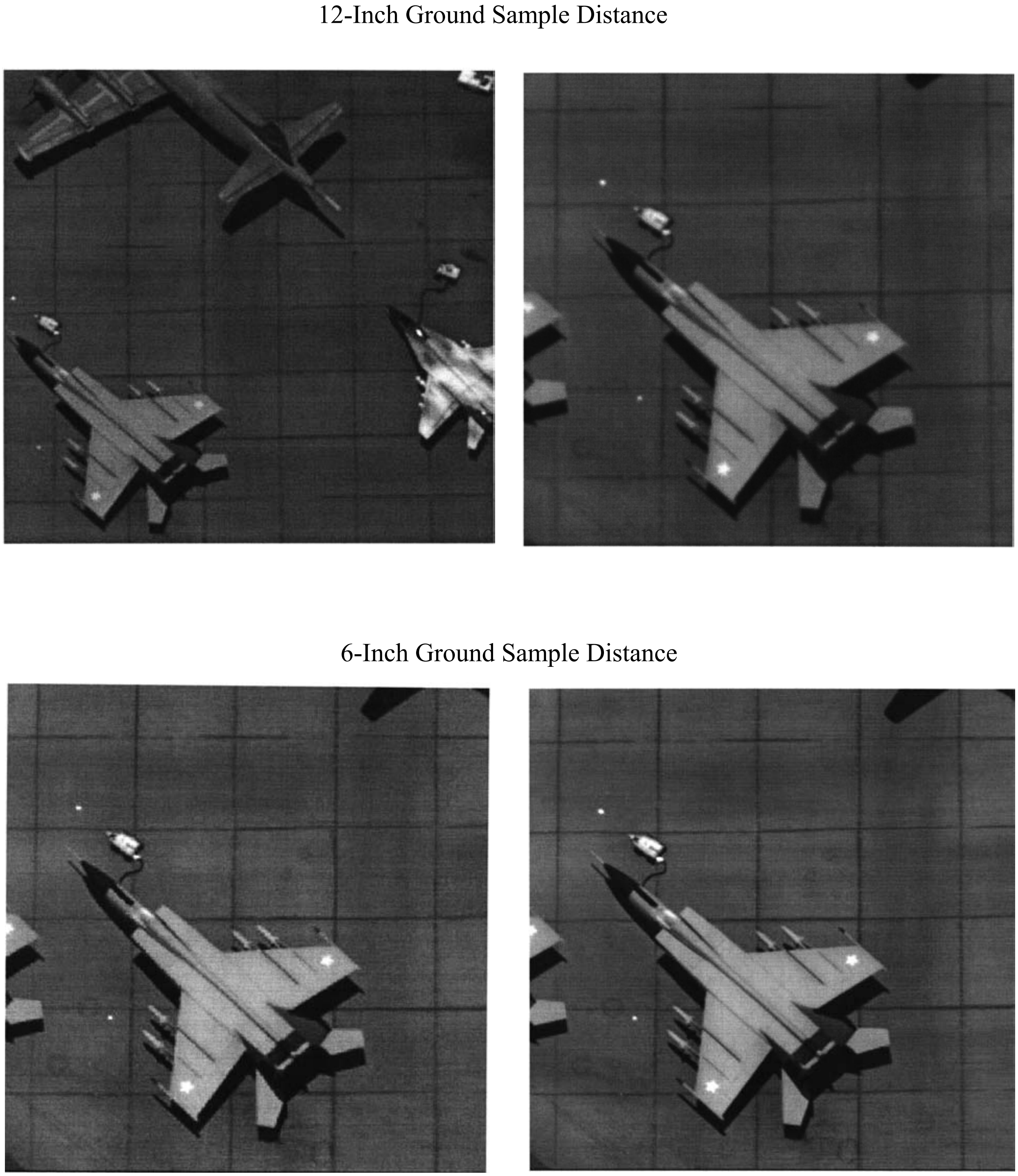
Figure 14-2. Tradeoff between coverage and resolution.[2]



48-Inch Ground Sample Distance

24-Inch Ground Sample Distance

Figure 14-2 (continued). Tradeoff between coverage and resolution.[2]



12-Inch Ground Sample Distance

6-Inch Ground Sample Distance

Image-Based Perception

Visual images are major source of information used by human beings to perform just about any function. We can use our vision to detect objects, to recognize specific objects, and to determine the probable class of unknown objects. The term "recognition" is somewhat loosely defined in common usage. It is sometimes used to denote the entire class of human perceptual processes that lay beyond target "detection". In other instances it carries a quite specific meaning and refers to a specific level of perceptual performance. Table 14-2 lists one hierarchical ordering of perceptual processes and their definitions.[3] Neither this hierarchy nor the definitions are universally accepted. However, they are in sufficient agreement with a large body of other sensor designers that the author sees no reason to alter it. For the remainder of this book, we will use the narrow definitions as shown in Table 14-2, unless it is specifically indicated that a broader definition is being used..

In an earlier paper, the author has related this hierarchy to increasing amounts of information available to the observer, although he has not proven this relationship. Although unproven, the concept is intellectually satisfying and of demonstrated utility. We will continue to make this association despite its lack of proof. The hierarchy begins with detection – meaning simply that something is present in the image and we can perceive its presence, but no other information is extractable. If slightly more information is available, the observer may be able to discern a level of symmetry in the object (or lack thereof) and a rough orientation (horizontal or vertical). With more

Table 14-2. Levels of image-based object perception.

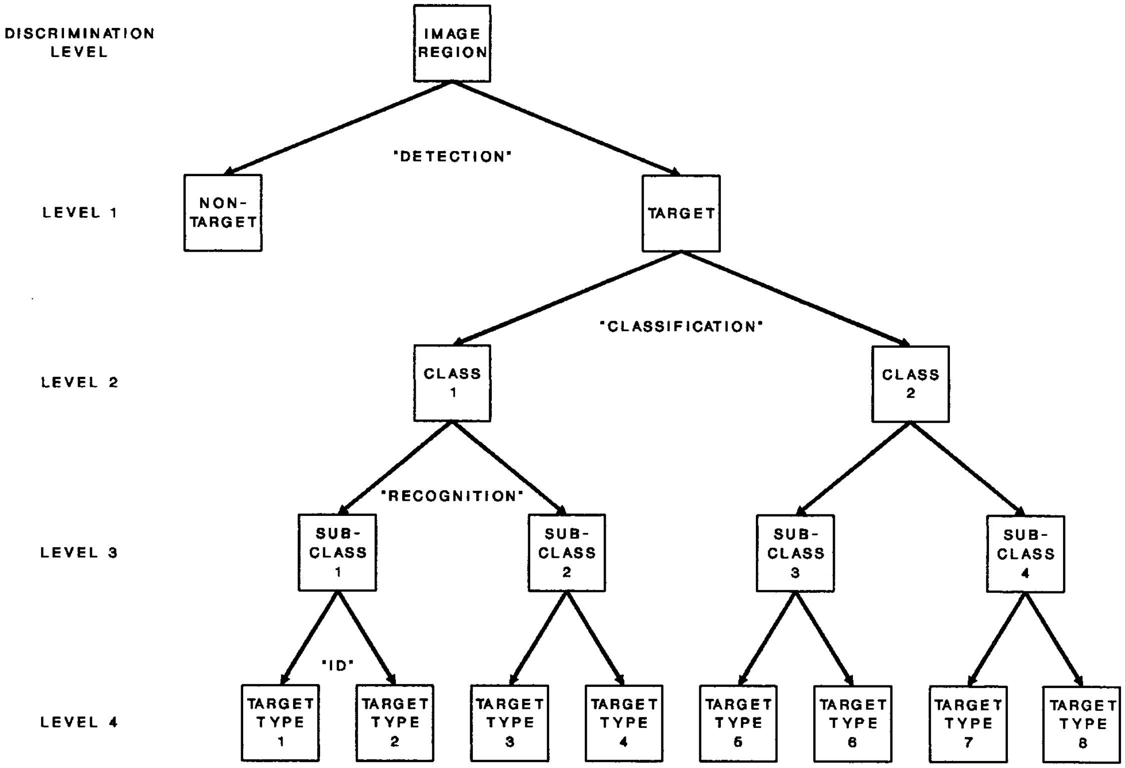
PERCEPTUAL LEVEL	MEANING	EXAMPLE
DETECTION	AN OBJECT IS PRESENT.	OBJECT
ORIENTATION	THE OBJECT IS APPROXIMATELY SYMMETRIC OR ASYMMETRIC AND ITS ROUGH ORIENTATION MAY BE DISCERNED.	HORIZONTAL RECTANGLE
CLUTTER REJECTION	THE OBJECT IS A POTENTIAL TARGET AND NOT A CLUTTER OBJECT.	TARGET
CLASSIFICATION	THE BROAD CLASS OF OBJECT TYPES TO WHICH THE OBJECT BELONGS MAY BE DETERMINED.	TRACKED VEHICLE
RECOGNITION	THE SUBCLASS OF OBJECT TYPES TO WHICH THE OBJECT BELONGS MAY BE DETERMINED.	TANK
IDENTIFICATION FRIEND-OR-FOE	THE "COUNTRY OF MANUFACTURE" OF THE OBJECT MAY BE DETERMINED.	WARSAW PACT TANK
IDENTIFICATION	THE SPECIFIC OBJECT TYPE WITHIN A CLASS OF OBJECTS MAY BE DETERMINED.	T-72 TANK
DISCRIMINATION	REAL TARGETS MAY BE DISTINGUISHED FROM REPLICA DECOYS.	REAL T-72
INTENT DETERMINATION	OWNERSHIP OF AN OBJECT MAY BE DETERMINED AND/OR PROBABLE HOSTILE/NON-HOSTILE INTENT MAY BE ESTABLISHED.	IRAQI T-72 (HOSTILE)

information, the observer may be able to tell that the object is not likely of natural origin – it is probably man-made and therefore a potential target and not a clutter object. At a higher level of perception we can determine the broad class of objects to which the object belongs. As we progress up the hierarchy, we can determine the subclass, and specific type of object. At the highest level of perception we can discriminate a real object from a decoy intended to deceive the observer and possibly even determine the likely intent of the object from its detailed configuration and behavior.

A governmental entity called the Automatic Target Recognizer Working Group (ATRWG) published a slightly different hierarchy [4] shown in Figure 14-3. It has fewer levels but is in significant agreement with the trend of Table 14-2. The four levels which equate to detection, classification, recognition, and identification in Table 14-2 are used extensively in discussions of imaging sensor performance. The perceptual level of orientation was used in early work (described in the next section), but is seldom used today.

How we perceive and "recognize" objects is the study of the science of cognition. We have made great strides in our understanding how the visual system works in transforming stimuli into nerve signals, but to date, we still do not have even the simplest workable concept for describing how we recognize objects. Nevertheless, we have been able to perform numerous experiments that

Figure 14-3. ATRWG image-based discrimination hierarchy.[4]



have allowed us to develop empirical relationships for predicting perceptual performance and quantifying aspects of the visual system relevant to sensor design. The direct relationships to perceptual processes are described in the next section. Before doing this we will present some of the results that have influenced that work.

Rosell & Willson [5] found in empirical studies that the probability of image-based detection by human observers of small, rectangular, normal contrast targets was a function of signal-to-noise ratio as shown in Figure 14-4. A detection probability of 50% requires SNR of 2.8. Note: in this instance the author means SNR not CNR. If the SNR is less than this, the target to be recognized will often not be detected in the first place. Electro-optical systems designed to have human observers will typically satisfy this criterion at their design range.

Another empirical observation is that human observers can only utilize a finite number of gray levels (shades of gray) in images. Figure 14-5 shows a sequence of images made with increasing numbers of gray levels.[6] Somewhere around 64-128 gray levels image quality does not continue to improve. Thus, designing a system to provide more 128 gray levels will not improve perceptual performance and is likely to be overdesigned.

Figure 14-4. Rosell-Willson rectangular target detection probability.[5]

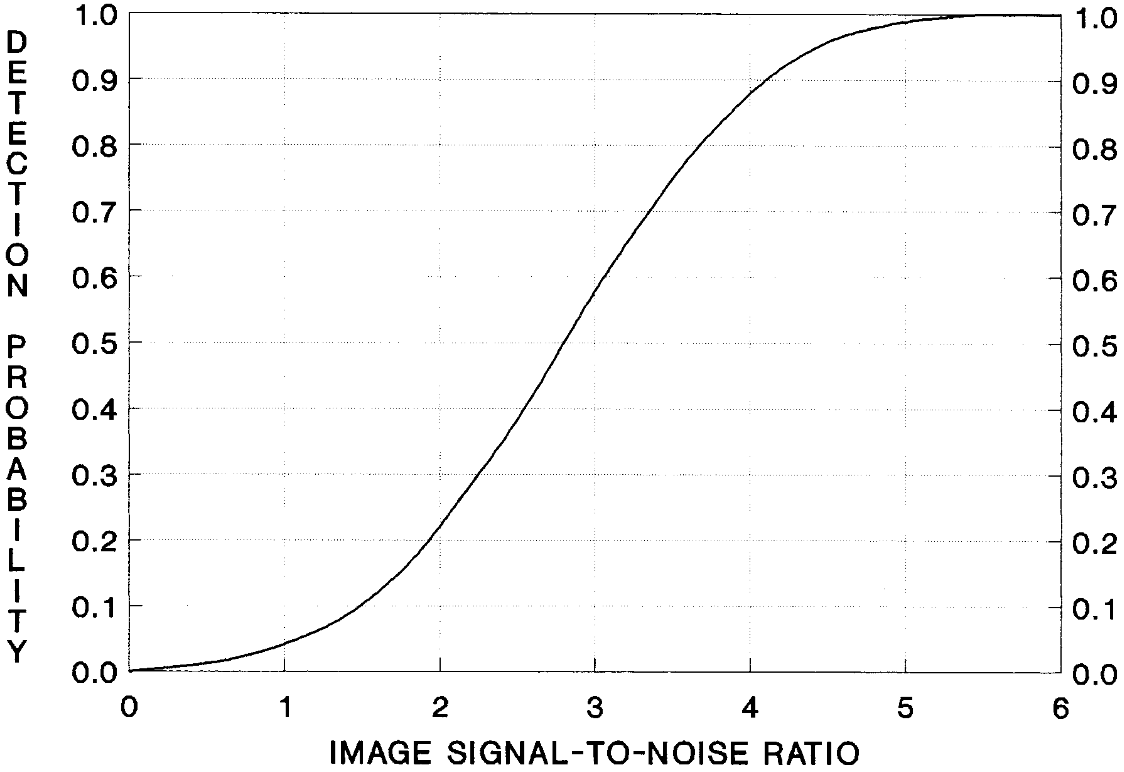
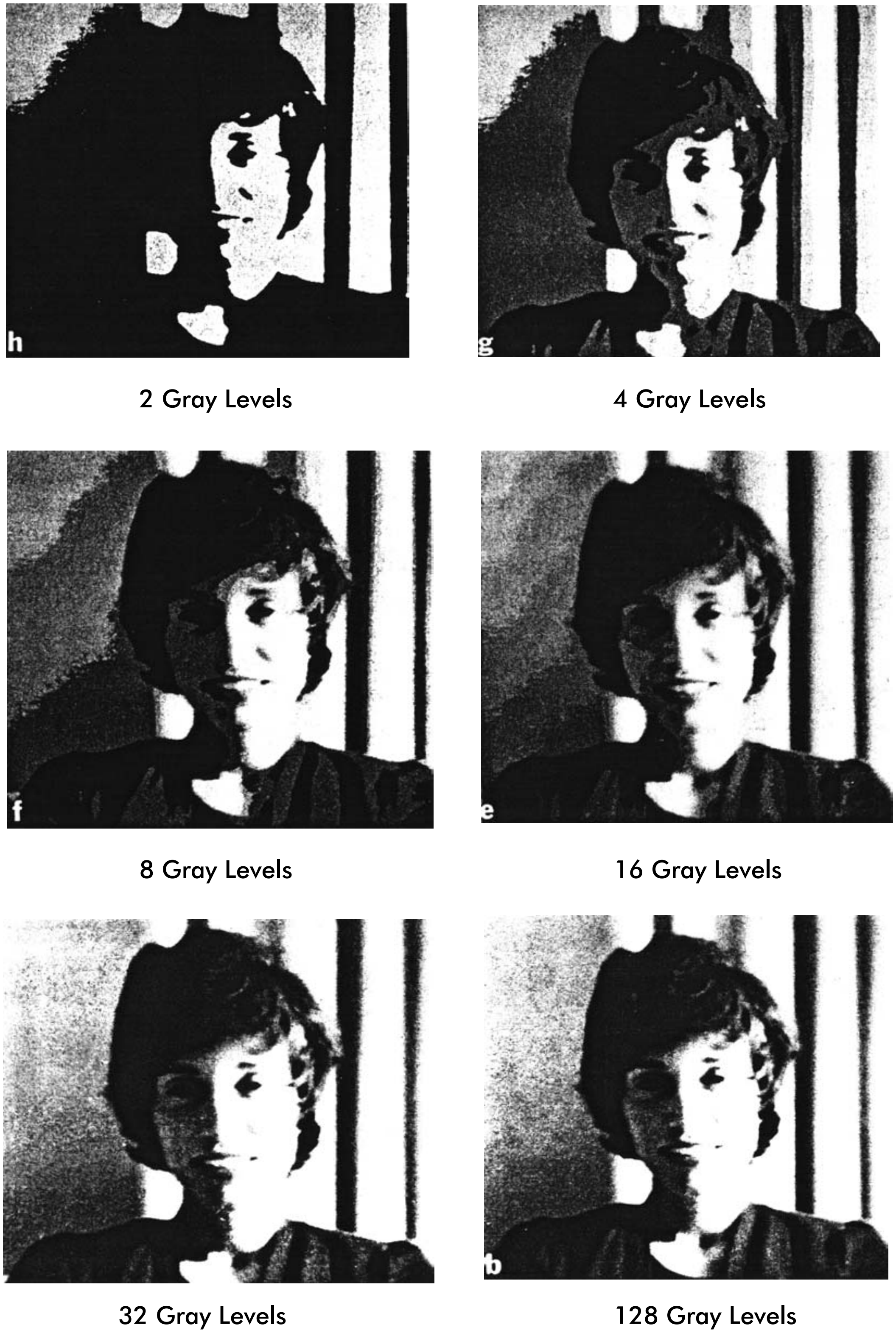


Figure 14-5. Comparison of images displayed using different numbers of gray levels.[6]



2 Gray Levels

4 Gray Levels

8 Gray Levels

16 Gray Levels

32 Gray Levels

128 Gray Levels

Johnson's Criteria

In the mid-1950's, John Johnson at the U. S. Army Engineer Research and Development Laboratories at Fort Belvoir (at last account now called the Night Vision and Electronic Sensors Directorate) performed a series of psychophysical experiments on perception of image intensifier images.[7] Targets were drawn from nine classes were placed at different ranges in a benign terrain context and were viewed with an image intensifier by a large number of observers. Each observer was asked what he saw in the image:

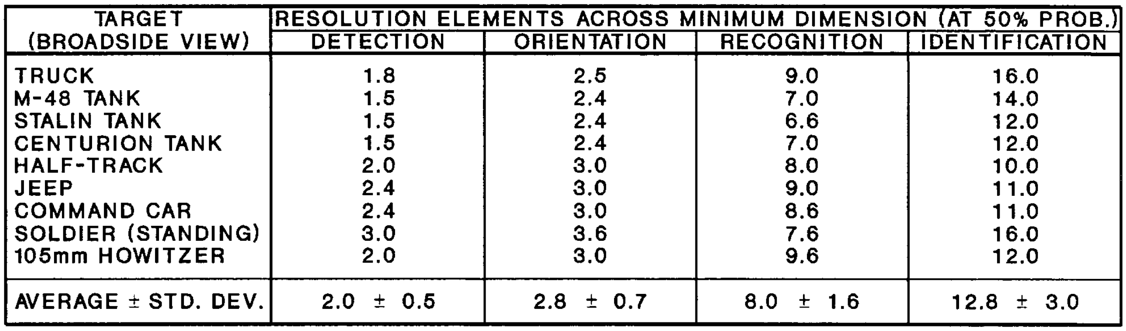
Was an object present?
If yes, what did you see?
If yes, what was the object?

At each range, 4-bar test patterns (similar to that in Figure 12-3) were also imaged and the minimum number of resolvable lines across the minimum dimension of each target was determined. After collecting a large number of data, Johnson statistically analyzed them. He found a strong correlation between how much could said about the target and the minimum number of resolvable lines detected. The results are shown in Figure 14-6. The average over all target classes was also determined. Statistically, he found that if 2.0 ± 0.5 resolvable lines across the minimum dimension could be perceived then the observer had a 50% probability of detecting the target. If 8.0 ± 1.6 resolvable lines could be distinguished then the observer had a 50% probability of recognizing the target. The four values

2.0 resels per min. dimension	→	50% detection probability
2.8 resels per min. dimension	→	50% orientation probability
8.0 resels per min. dimension	→	50% recognition probability
12.8 resels per min. dimension	→	50% identification probability

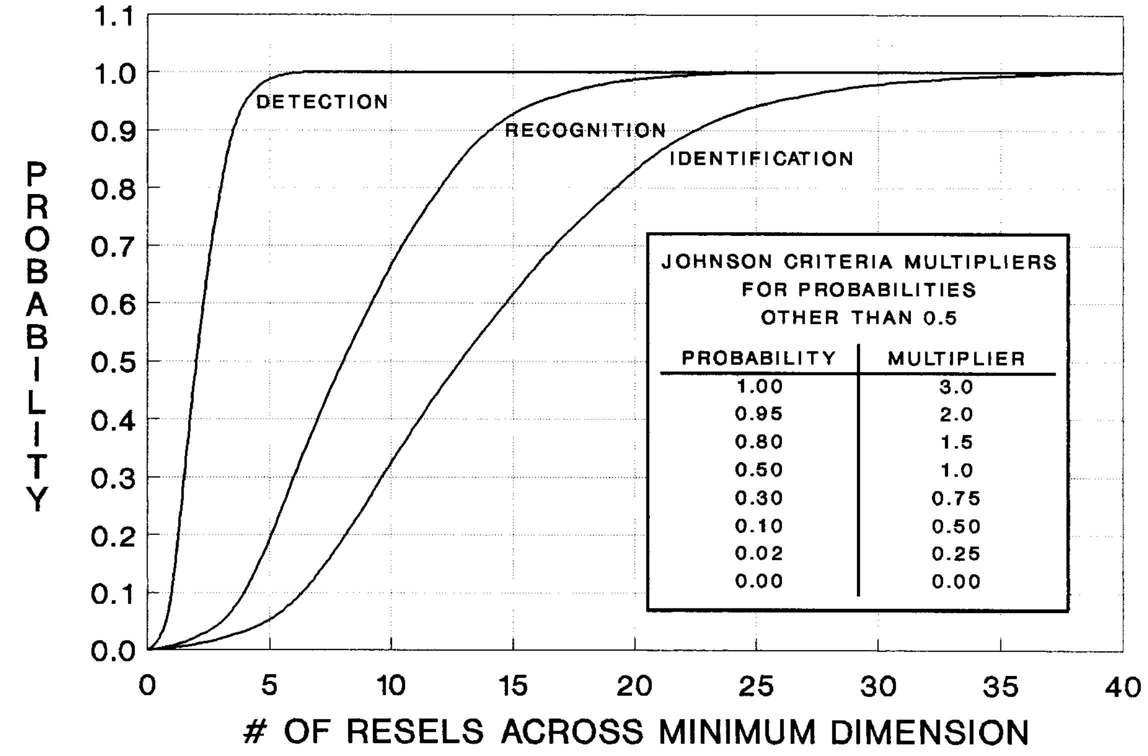
have become known as Johnson's criteria. These values were observed to be essentially independent of scene signal-to-noise ratio and target-to-background contrast as long as the scene values and the resolution chart values for these parameters were roughly the same. Similar studies using other

Figure 14-6. Johnson's criteria.



TARGET	RESOLUTION ELEMENTS ACROSS MINIMUM DIMENSION (AT 50% PROB.)			
(BROADSIDE VIEW)	DETECTION	ORIENTATION	RECOGNITION	IDENTIFICATION
TRUCK	1.8	2.5	9.0	16.0
M-48 TANK	1.5	2.4	7.0	14.0
STALIN TANK	1.5	2.4	6.6	12.0
CENTURION TANK	1.5	2.4	7.0	12.0
HALF-TRACK	2.0	3.0	8.0	10.0
JEEP	2.4	3.0	9.0	11.0
COMMAND CAR	2.4	3.0	8.6	11.0
SOLDIER (STANDING)	3.0	3.6	7.6	16.0
105mm HOWITZER	2.0	3.0	9.6	12.0
AVERAGE ± STD. DEV.	2.0 ± 0.5	2.8 ± 0.7	8.0 ± 1.6	12.8 ± 3.0

Figure 14-7. Resolution dependence of Johnson's criteria probabilities.



JOHNSON CRITERIA MULTIPLIERS FOR PROBABILITIES OTHER THAN 0.5

PROBABILITY	MULTIPLIER
1.00	3.0
0.95	2.0
0.80	1.5
0.50	1.0
0.30	0.75
0.10	0.50
0.02	0.25
0.00	0.00

target sets have not contradicted Johnson's original results.

Johnson also determined the basic resolution versus probability behavior of the detection, recognition, and identification probabilities. This behavior is shown in Figure 14-6.[8] A small table inset to the figure summarizes the multiplier (of the appropriate Johnson's criterion) to convert to probabilities other than 50%.

These criteria or their most recently defined values [10] are used as heuristics by design engineers in estimating required aperture sizes and are utilized by most modern electro-optical performance prediction codes [8-11] as the principal criteria for determining detection, recognition, and identification probabilities as functions of range. They have been assumed to be applicable across the whole range of electro-optical sensor systems. Little evidence has arisen to invalidate this assumption.

Johnson's criteria were developed from studies of rasterless image intensifier imagery. Color (which results in significant added information) complicates their applicability to direct view optical systems. In systems where color is distorted (such as produced by the use of yellow or orange haze filters) or at lower light levels (where scotopic vision predominates) they are usually accepted as valid. They are commonly accepted as valid for monochrome television and thermal imaging systems, despite the presence of a raster on most displays. Limited personal experience with multiple- frame-averaged laser radar reflectivity images does not contradict the validity of Johnson's criteria, if the higher angular resolution of the laser radar (see the Appendix for a discussion of angular resolution in sensors) is properly considered. Speckled images such as single-frame laser radar reflectivity images and synthetic aperture radar images are much less useful than speckle-averaged images, and are not believed to obey Johnson's criteria, probably due to their low SNR (=1). The last two sections of this chapter will elaborate on image SNR and its significance.

Applications of Johnson's Criteria

The simplest and most common use of Johnson's criteria is to perform back-of-the-envelope estimates of minimum aperture sizes for imaging sensors. If we combine any appropriate angular resolution expression, such as,

$$\alpha = X\lambda / D \tag{14.6}$$

where X is the constant in the resolution expression (X= 1.220 for the Rayleigh criterion and an Airy function response) with an expression for the linear size l of an object of angular size θ at range R

$$l = \theta R \tag{14.7}$$

and with the appropriate Johnson's criterion N_{50} (the number of resolution elements across the minimum target dimension to give 50% probability of performing a specified perceptual task – N_{50} = 2 for detection) we obtain

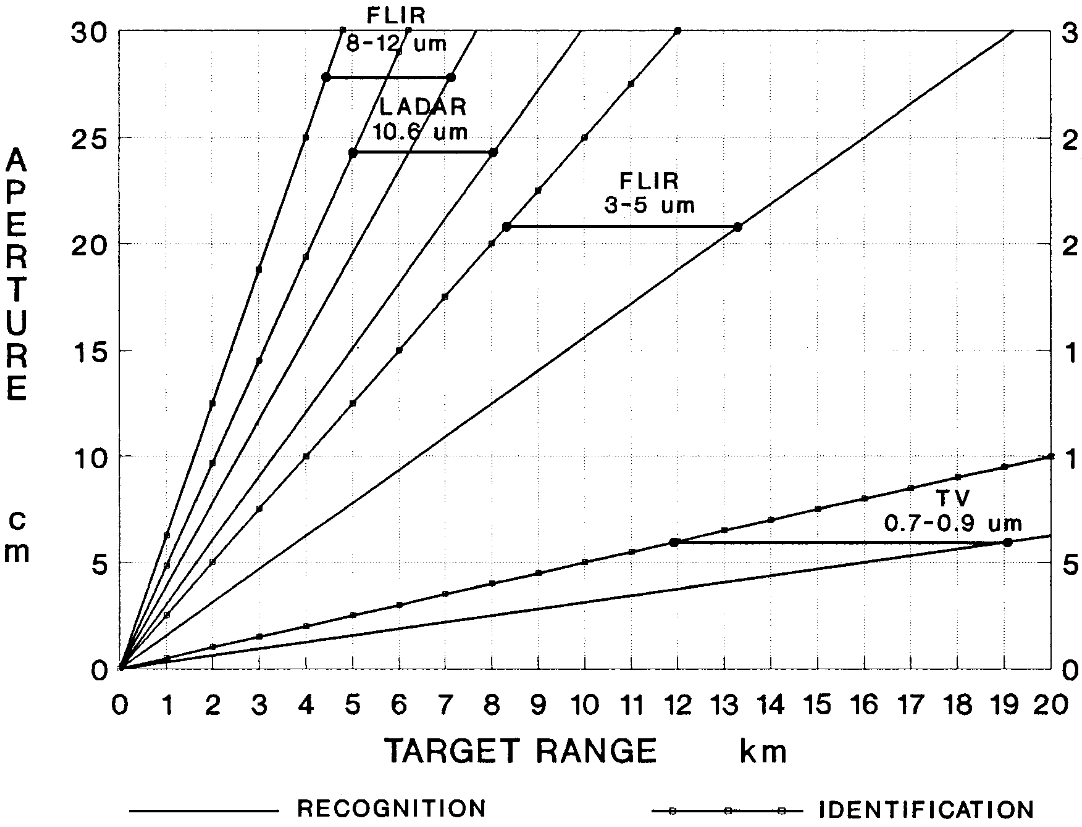
$$R = \frac{l}{\alpha N_{50}} = \frac{l\,D}{X\lambda N_{50}} \tag{14.8}$$

Equation (14.8) gives the maximum range at which a particular Johnson's criterion N_{50} can be satisfied for a target of minimum dimension l using a particular resolution criterion $X\lambda/D$ for a sensor with wavelength λ and aperture diameter D. This expression is plotted in Figure 14-8 for a number of sensors. The criterion for both recognition (N_{50} =*8*) and identification (N_{50} = *12.8*) have been used. The Rayleigh resolution criterion was used for all sensors and the target dimension was assumed to be 2.5 m.

The curves in Figure 14-8 represent the minimum aperture sizes that are consistent with being able to perform a given perceptual function at a specified range. These estimates will be reasonably accurate if the aperture is the limiting factor in determining the total sensor resolution. If the number or size of the elements in the detector array or the bandwidth of a filter in the video processing electronics is the limiting factor, then Figure 14-8 is not applicable. However, in a well-designed system, the aperture is usually the limiting factor.

Examination of Figure 14-8 shows a number of interesting facts. First, is that 8-12 μm thermal imagers cannot perform recognition or identification at ranges greater than 5 km with reasonably sized apertures (20 cm or less). Nevertheless, we have sent such sensors into combat with weapons of longer effective range than 5 km. Such cases have resulted in fratricide (casualties produced by friendly fire). Thermal imagers operating in the 3-5 μm perform much better in terms of perceptual performance. If recognition performance is desired at ranges of 20-40 km (desirable ranges for air combat) then the imaging sensor must operate in the visible or near infrared. This is the primary reason that television cameras are still used in weapons systems despite limitations on

Figure 14-8. Aperture size requirements for different sensors and imaging functions.



their performance in adverse weather and at night. A final observation is that laser radars operating at 10 μm should outperform thermal imagers operating at 10 μm in terms of recognition performance.

A significant step up in sophistication in the application of Johnson's criteria is the "Night Vision Laboratory Search Model".[11] This is a computer model of the entire visual search process. The search field of regard is broken into *n* separate sensor fields of view. Images are assumed to be obtained for each field of view. Each field of view image is viewed for an arbitrary period of time until the searcher decides to view the next image. This time is assumed to be roughly the same for each field of view. The fields of view may or may not be revisited.

If an infinite amount of time could be expended on searching the field of regard, one expects a certain probability of detection, that is given by the expression

$$p_\infty = \frac{\left(N / N_{50}\right)^E}{1 + \left(N / N_{50}\right)^E} \tag{14.9}$$

where E has been empirically derived to be

$$E = 2.7 + 0.7(N / N_{50}) . \tag{14.10}$$

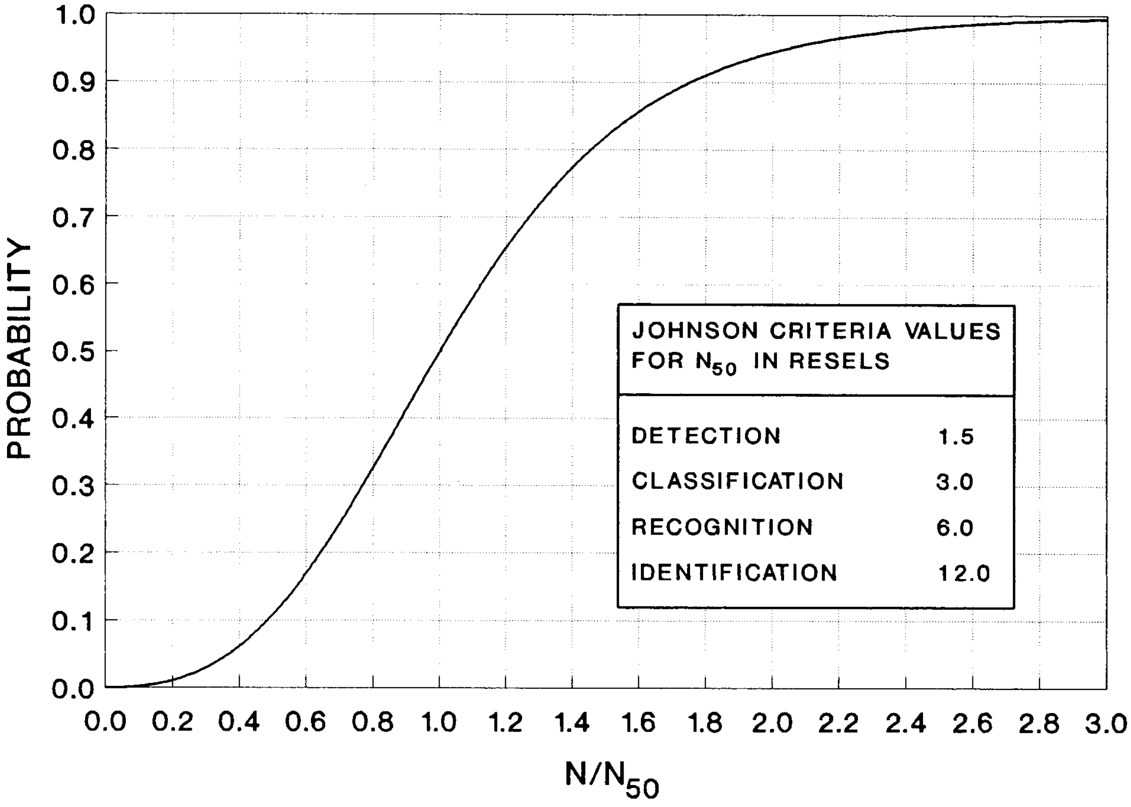
N is the number of resolvable cycles (one cycle equals two resels) across the target dimension and N_{50} is the number of cycles across the target for 50% detection probability (essentially Johnson's criterion for detection). Eq. (14.10) is qualitatively and quantitatively very similar to the curve in Figure 14-7, the original Johnson criterion result. Equations (14.9) and (14.10) are plotted in Figure 14-9.

If a finite amount of time T can be expended on searching the field of regard, the detection probability is less than the infinite time probability by the relation

$$p(T) = p_{\infty}\left[1 - e^{-T/n\tau}\right] \tag{14.11}$$

where n is the number of fields of view within the field of regard and τ is the mean time to detect

Figure 14-9. The infinite time probability function used in the NVL Search Model.



JOHNSON CRITERIA VALUES FOR N_{50} IN RESELS	
DETECTION	1.5
CLASSIFICATION	3.0
RECOGNITION	6.0
IDENTIFICATION	12.0

a target within a single field of view (if a target is found). This time (in seconds) is given by

$$\tau = 3.4 / p_\infty . \tag{14.12}$$

The search model is a static model. That is, it assumes that there is no target motion. Target motion produces much higher detection probabilities and shorter search times than predicted by the search model.

In current versions of imaging system performance codes, Johnson's criteria are used in slightly modified format. The modification is based on a recognition that pixels lying across the maximum dimension of the target also contribute to detection, recognition, and identification. Obviously, knowledge of whether a target is "square" (minimum and maximum dimensions roughly equal) or "rectangular" (minimum dimension much smaller that the maximum dimension). Using two-dimensional criteria requires slightly different numbers.

A critical target dimension is defined by the geometric mean of the target's vertical and horizontal dimensions

$$L_C = \left(L_X L_Y \right)^{1/2} . \tag{14.13}$$

The new values for the Johnson's criteria are:[10]

1.5 resels (0.75 cycles) per critical dimension	→ 50% detection probability
3.0 resels (1.5 cycles) per critical dimension	→ 50% classification probability
6.0 resels (3.0 cycles) per critical dimension	→ 50% recognition probability
12.0 resels (6.0 cycles) per critical dimension	→ 50% identification probability

The most sophisticated application of Johnson's criteria is in the imaging system performance models developed by the "Night Vision Laboratory". Consider the model called FLIR92/ACQUIRE which is designed for thermal imaging systems but can be adapted to other types of sensors. FLIR92 allows the sensor characteristics to be calculated and incorporated into a set of "minimum resolvable temperature difference" (MRTD) curves. We will discuss this aspect of the calculation in more detail in the chapter on thermal imaging systems later in this volume. At this point all we need to know is that these curves describe the minimum resolvable temperature difference between target and background as a function of the spatial frequency for a hypothetical resolution pattern viewed by the sensor. The ACQUIRE portion of the code calculates probability of performing a specific perceptual function in the following manner.

The target is characterized by horizontal and vertical dimensions L_H and L_V and a temperature difference ΔT. Atmospheric extinction (defined by extinction coefficient α) is assumed to reduce the real target signature to an apparent value at range R given by

$$\Delta T_{APPARENT} = \Delta T\, e^{-\alpha R} \qquad (14.14)$$

The angular size of the target at range R is defined by the equations

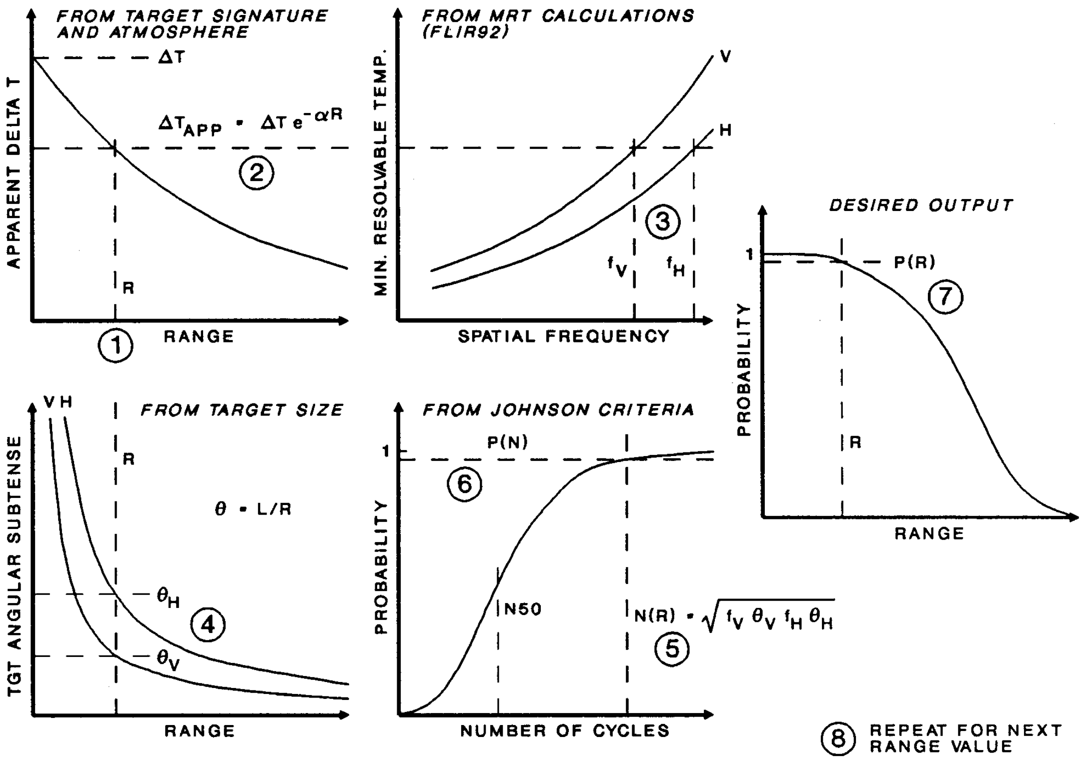
$$\theta_H = L_H \,/\, R \qquad \text{and} \qquad \theta_V = L_V \,/\, R \qquad (14.15)$$

The model begins at a small value of the range. At this range $\Delta T_{APPARENT}$ is calculated. This value of ΔT is compared to the MRTD curves to determine the horizontal and vertical spatial frequencies (in cycles/mrad) that match the apparent temperature difference. At the same range we calculate the angular sizes of the target using Eq. (14.15). The products of the spatial frequencies and the angular sizes gives the number of resolvable cycles across the target at range R

$$N(R) = \sqrt{f_V\, f_H\, \theta_V\, \theta_H} \qquad (14.16)$$

The number of resolvable cycles is translated into a probability using the search model Eqs. (14.9) and (14.10). This process is repeated for a number of different range values until a plot of probability of "perceptual function" versus target range can be plotted. Figure 14-10 illustrates the steps ((1) thru (8)) in this iterative process. The approach can be adapted to model any imaging system.

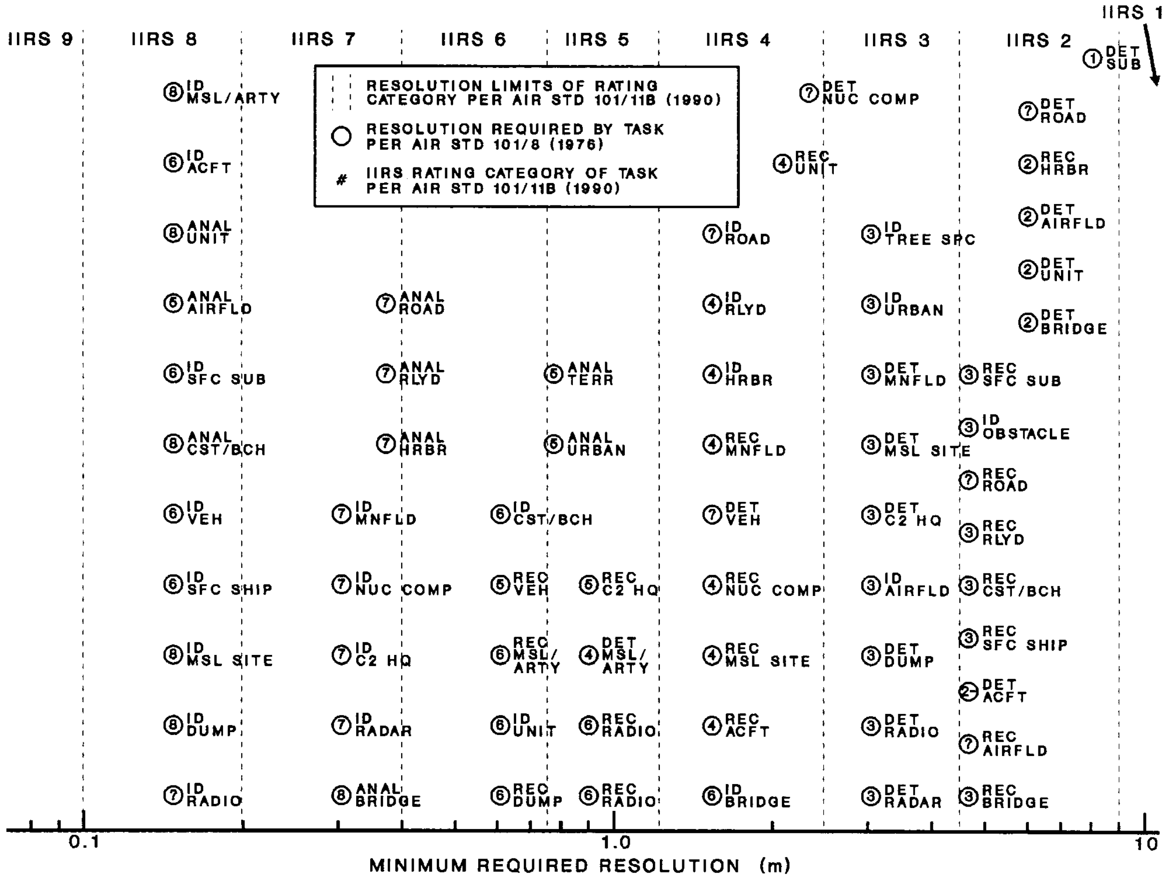
Figure 14-10. The algorithm used in the ACQUIRE model for imaging system performance.



Although it is not an application of Johnson's criteria, and Johnson's criteria were not used in its formulation, there is another quantitative measure of perceptual performance versus resolution that is in reasonable agreement with Johnson's criteria and serves the validate the general approach. This measure is the Image Interpretability Rating Scale (IIRS) used by photo-interpreters. The scale and its data requirements are set forth in two Air Standardization Coordination Committee standards [12]-[13] and summarized in Figure 14-11. The basic IIRS tasks are detect, recognize, identify, and analyze. The first three are essentially identical to the Johnson's criteria tasks with the same name. Analysis is the ability to determine critical target parameters to sufficient accuracy to support technical intelligence requirements. Analysis would probably lie well below the bottom of the list in Table 14-2.

Although it is difficult to make a close comparison, examination of the resolution requirements for detection, recognition, and identification as shown in Fig 14-11, indicates that the results are not in obvious disagreement with Johnson's criteria. It should be noted that IIRS was developed without any consideration (or probably even knowledge of) Johnson's criteria.

Figure 14-11. Tasks and their required ground resolution included in the Image Interpretability Rating Scale.



Information-Based Johnson's Criteria

The material in this section is conjecture. Despite significant intellectual appeal and apparent agreement with limited predictions, it still lacks unequivocal and incontrovertible substantiation. It concerns a methodology which permits the "Johnson's criteria" approach to perception to be extended to all forms of image-based perception from all kinds of generalized imaging sensors and possibly to perception based on non-imaging sources. Based on the author's personal research, it has a philosophical aspect and a practical aspect. The philosophical aspect involves the author's conviction that the nearly universal applicability of Johnson's criteria derives from a basis in fundamental principles. That is, that the metrics used in Johnson's criteria are directly relatable to total information content and that total available information drives perceptual ability. The philosophical and empirical justifications behind the conjectures are explained in detail in Reference [3]. In this section we will confine the discussion to the mechanics of using the **information-based Johnson's criteria**.

The first step is the determination of the total information content of an image. The total information H_T content may contain contributions from intensity data (the individual pixel brightnesses H_I and their arrangement in space H_2), the context in which the image is taken H_C, color (the variations in spectral content of the signal in each pixel) H_H, range (if "three-dimensional" image data is obtained) H_R, polarization/depolarization of the image radiation H_P, and/or velocity (associated with each pixel) H_V.

$$H_T = H_0 + H_1 + H_C + H_H + H_R + H_P + H_V \tag{14.17}$$

Obviously, not every image contains every kind of potential information.

The pixel **intensity information** is assumed to be given by the product of the number of pixels on the target N (background pixels contribute to context information) and the information content per pixel B_I.

$$H_0 = NB_I \tag{14.18}$$

In angle-angle images (human vision, television, thermal imagers, etc.) the number of target pixels is approximately given by

$$N = N_X N_Y = \left(L_X / \alpha_X R\right)\left(L_Y / \alpha_Y R\right) \tag{14.19}$$

where R is the range to the target, L_X and L_Y are the transverse dimensions of the target, and α_X and α_Y are the horizontal and vertical angular resolutions of the total sensor system. In range-cross range images (e.g., synthetic aperture radar images) the number of target pixels is approximately given by

$$N = N_X N_Z = \left(L_X / \Delta X\right)\left(L_Z / \Delta R\right) \tag{14.20}$$

where L_Z is the target size in the range dimension, ΔX is the cross-range resolution, and ΔR is the range resolution. The information content should be related to the signal-to-noise ratio of the data. From information theory we expect the information content in bits to be given by

$$B_I \approx \log_2\left(1+SNR\right) \tag{14.21}$$

Ideally B_I could be as large or as small as the *SNR* dictated. However, there are practical limits. In order for Johnson's criteria to be strictly applicable, the *SNR* must exceed Rosell's SNR (=2.8). A $SNR = 2.8$ gives B_I~2 bits. Later in this section we will allow for lower values of *SNR* (such as *SNR =1* for speckled laser radar images or synthetic aperture radar images) by an appropriate modification of Johnson's criteria. At the upper end, the intrinsic variability of materials will provide an effective limit. If one looks at a supposedly uniform surface, there will be small variations in intensity. This may be due to small variations in surface texture, reflectivity, or the cleanliness of the surface. These variations cannot provide information useful for target detection or recognition because two otherwise identical objects will show these differences, and because those differences will not necessarily be constant over time. The surface variability limit is given by

$$B_I \le \log_2\left(\mu / 2\sigma\right) \tag{14.22}$$

where μ is the mean intensity and σ is the standard deviation of the surface variations. For surface variability of 1% to 3% of the mean, this provides an upper limit of 4-5 bits.

The "**shape**" information derivable from the arrangement of the pixels is rather limited. It can be shown to be approximately given by the number of ways N pixels can be arranged as rectangles, which is approximately

$$H_1 \approx \log_2\left(N_X N_Y\right) \tag{14.23}$$

Context also conveys very limited information content. The maximum number of contextual situations that aid in the detection or recognition process is limited. For example, being in the sky, on a road, on a lake, or in a field conveys useful meaning. Ground vehicles are seldom found in lakes or in the air. Orientation of the target with respect to the terrain is also of interest. However, whether the target is in front of one tree, two trees, or any number of trees conveys the same amount of useful information as being in front of a forest. In practice, contextual information is expected to convey at most

$$H_C \approx 9-11 \text{ bits}. \tag{14.24}$$

Sensors that measure information in multiple wavelength bands can provide **color** information. The magnitude of this information is provided by

$$H_H = N\ N_P\ B_H \tag{14.25}$$

where N is the number of pixels, N_P is the effective number of "primary colors" that are detected, and B_H is the number of bits of information content per primary color. There is a trend towards use of multispectral imagers. In such devices, images may be obtained at dozens if not hundreds of separate wavelengths. The number of wavelengths measured is not the same as the number of primary colors available in the data. If the world were covered with standard three-color camouflage, there would be only three meaningful colors, no matter how many spectral bands into which a sensor tried to partition the signals. Any observed spectral pattern could be shown to be produced by a linear combination of the responses to the three primary colors. In practice the number of primary colors is probably of the order of 10. The bits per color will be subject to the same intrinsic variability limits as intensity information. Thus, B_H is likely 4-5 bits unless limited by low *SNR*.

Some sensors can provide range information. **Absolute range** information is to be differentiated from **relative range** information. Knowing that a target is at 2501 meters versus 2500 meters tells us very little that is different from knowing that the target is at a range between 2400 and 2600 meters. However, knowing that the front of a vehicle is 6.5 m less range than the rear of the vehicle tells a great deal about the target. The utility of absolute range information is in transforming angular information into linear information ($X=\alpha R$). The information provided by absolute range is related to the number of range bins necessary to estimate target sizes to an accuracy of roughly 1 resel. This can be shown to be equivalent to

$$H_{AR} = \log_2\left(4\, N_X\, N_Y\right) \ . \tag{14.26}$$

Relative range can yield considerably more information. If δR is the range precision of the sensor then the information per pixel provided by breaking a target of range depth L_Z into range bins of size δR is

$$B_{RR} \approx \log_2\left(L_Z / \delta R\right) . \tag{14.27}$$

Velocity information is much the same as range information. Gross velocity (**body motion**) provides only limited information about target class. For example, breaking the velocity spectrum into stationary, walking speed, vehicle speed, helicopter speed, subsonic air vehicle speed, supersonic air vehicle speed, hypersonic speed, and orbital velocity coupled with direction (towards, away, or transverse to the sensor line of sight) provides about all the discrimination information that one can practically use. If we can define N_M motional velocity regimes then the gross velocity information content is

$$H_M = \log_2 N_M \ . \tag{14.28}$$

At most this will contribute 3-5 bits of information. Details about internal motions of a target are possible more useful. By analogy to the relative range information, we may define the **internal motional information** content per pixel to be

$$B_V \approx \log_2\left[\left(V_{\max} - V_{\min}\right) / \delta V\right] \tag{14.29}$$

where $V_{\max}$ - $V_{\min}$ is the spread in velocities across the target and δV is the velocity measurement precision. The total internal velocity information is

$$H_V = NB_V . \tag{14.30}$$

If a system is capable of using polarization, then the bits per pixel of **polarization** information is obviously

$$B_P = \begin{cases} 0 & (1 \text{ xmt} / 1 \text{ rcv}) \\ 1 \times \log_2\left(1 + SNR\right) & (1 \text{ xmt} / 2 \text{ rcv or } 2 \text{ xmt} / 1 \text{ rcv}) \\ 2 \times \log_2\left(1 + SNR\right) & (2 \text{ xmt} / 2 \text{ rcv}) \end{cases} \tag{14.31}$$

depending on how polarization is used and measured.

The **total information** available from an image is the sum of all of the relevant terms above. That is,

$$\begin{aligned} H_{IMAGE} = N_X\ N_Y\left[B_I + N_P B_H + B_{RR} + B_V + B_P\right] \\ + H_1 + H_C + H_{AR} + H_M \end{aligned} \tag{14.32}$$

The utility of the information-based approach comes only after we define a new set of Johnson's criteria. In reference [3] the author derived the following criteria (H_{50})

$$H_{DET} \approx 5 \text{ bits} \tag{14.33}$$

$$H_{REC} \approx 88 \text{ bits} \tag{14.34}$$

$$H_{ID} \approx 306 \text{ bits} \tag{14.35}$$

as the amount of information needed to perform detection, recognition, and identification at the 50% probability level. When properly applied, these information-based criteria predict performance that is almost identical to that of the Johnson's criteria based solely on resolution. The information content would be determined and compared to the information-based criteria. Taking the square root of the ratio yields a quantity roughly equivalent to the ratio N/N_{50} used in the search model. Making this equivalence explicit, we may use the search model to predict the infinite time probabilities. Thus,

$$\frac{N}{N_{50}} \approx \left(\frac{H_{IMAGE}}{H_{50}} \right)^{1/2} . \tag{14.36}$$

To illustrate the use of the information-based approach, let us first compare a thermal imager and a three-dimensional imaging laser radar. Both systems have the same aperture and operate at the same wavelength and must produce 50% recognition at some range. We wish to evaluate the relative range at which this occurs. For the thermal imager we assume $B_I = 4$ bits/pixel. For the laser radar, we have $B_I = 1$ bit/pixel and $B_{RR} = 5$ bits per pixel (an easily achievable range precision). We will assume that the resolution of the radar is 0.73 times the resolution α_0 of the thermal imager (this is the ratio of resolutions at the Rayleigh criterion). The thermal imager will have pixel-dependent (terms proportion to the number of pixels) information content of

$$H_{TI} \approx 4\left(L_X L_Y / \alpha_0^2 R^2 \right) \tag{14.37}$$

while the laser radar will have information content

$$H_{LR} \approx 6\left(L_X L_Y / \left(0.73\alpha_0\right)^2 R^2 \right) \tag{14.38}$$

Since equal information content should yield equal recognition performance, the effective range of the laser radar is much larger than that of the thermal imager

$$R_{LR} \approx 1.68 R_{TI} . \tag{14.39}$$

Another example involves estimating the recognition potential of a synthetic aperture radar (SAR), which is not amenable to treatment using the ordinary Johnson's criteria. We assume a target size of 3 x 7 meters. For an assumed information per pixel of $B_I = 1$ bit/pixel, the information content of the SAR is

$$H_{SAR} \approx 1\left(L_X L_Z / \Delta X \Delta R \right) + 16 \text{ bits} \tag{14.40}$$

The 16 bits accounts for spatial arrangement and context information. Using the 88 bits recognition criterion, we calculate that the SAR can achieve 50% recognition if the resolution product obeys the relation

$$\Delta X \Delta R \le \left(7 \times 3 \right) / 72 = 0.29 \text{ m}^2 \ . \tag{14.41}$$

Example - Atmospheric Turbulence Effects on Laser Radar Images

As discussed in Chapters 4 and 11, turbulent scintillation causes an otherwise non-fluctuating (glint) target to have amplitude fluctuations of magnitude [14]

$$I_g = I_0 e^{4\chi} \tag{14.42}$$

where χ is the log-amplitude. The log-amplitude is described by a normal (Gaussian) distribution with a mean equal to minus the variance, i.e.,

$$p_\chi(\chi) = \frac{1}{\sqrt{2\pi\sigma_\chi^2}} e^{-\frac{\left(\chi+\sigma_\chi^2\right)^2}{2\sigma_\chi^2}} . \tag{14.43}$$

Recalling the definition of the saturation signal-ro-noise (Eq. (11.17))

$$SNR_{SAT} = \frac{\left\langle \left|\underline{y}\right|^2 \right\rangle^2}{\mathrm{var}\left(\left|\underline{y}\right|^2\right)} \tag{14.44}$$

we can easily calculate [14]

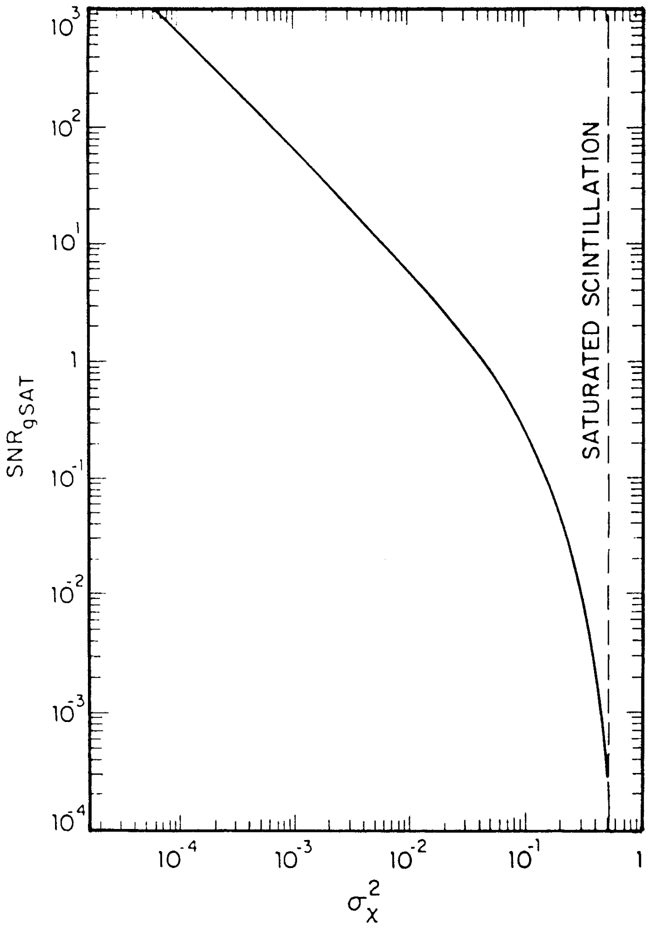
$$SNR_{gSAT} = \frac{1}{e^{16\sigma_\chi^2} - 1} \tag{14.45}$$

This function is plotted in Figure 14-12. It takes on an infinite value at zero turbulence level. This is expected, as the target is otherwise non-fluctuating and by definition has a zero variance. However, as the scintillation approaches saturation, the SNR_{SAT} plunges to extremely low values, reaching as low as 3 x 10^{-4} at saturation. There is virtually no usable signal left when the *SNR* is that low.

As discussed in Chapter 11, the presence of Rayleigh-fading and aperture-averaging alters the probability distribution of fluctuations on a speckle target (Rayleigh-fading target) to

$$I_s = I_0 e^{4\sigma_\chi^2} V e^{2U} \tag{14.46}$$

Figure 14-12. Glint-target saturation signal-to-noise ratio vs turbulence strength.[14]



where V is a unit-mean, exponentially distributed random variable with distribution

$$p_V(V) = e^{-V} \tag{14.47}$$

which accounts for Rayleigh fading and U is a log-normal variable given by

$$p_U(U) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(U+\sigma^2\right)^2}{2\sigma^2}} \tag{14.48}$$

where σ^2 is the aperture-averaged log-amplitude variance. In the absence of turbulence ($\sigma^2 = 0$) then mean of the remaining exponential term is 1 as is the variance. Thus, the saturation signal-to-noise ratio for a Rayleigh-fading target without turbulence is

$$SNR_{sSAT} = 1 \tag{14.49}$$

With turbulence included it can be shown that the saturation signal-to-noise ratio for the speckle target is [14]

$$SNR_{sSAT} = \frac{1}{2e^{4\sigma^2} - 1} \tag{14.50}$$

This function is shown in Figure 14-13. As expected, when the turbulence log-amplitude variance goes to zero, Eq. (14.50) approaches unity. It stays near unity until the scintillation log-amplitude variance becomes comparable to one. Then, the saturation *SNR* falls to a value less than 2 x 10^{-4} as the scintillation reaches saturation.

Signal-to-noise ratios much less than one are difficult enough to comprehend when dealing with simple signals and noise. When dealing with images it can be harder to comprehend the results. Additional insight can be obtained by studying synthetic images degraded by high noise levels. In Figure 14-14 a synthetic glint (non-fluctuating) target is shown against a supposedly uniform background (the reproduction of the synthetic images is less than perfect). Each pixel in the original

Figure 14-13. Speckle-target saturation signal-to-noise ratio vs turbulence strength.[14]

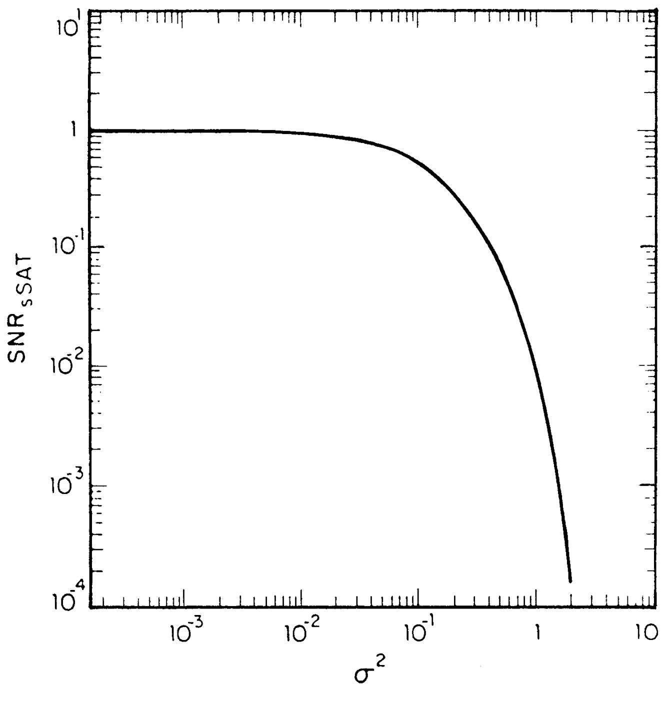
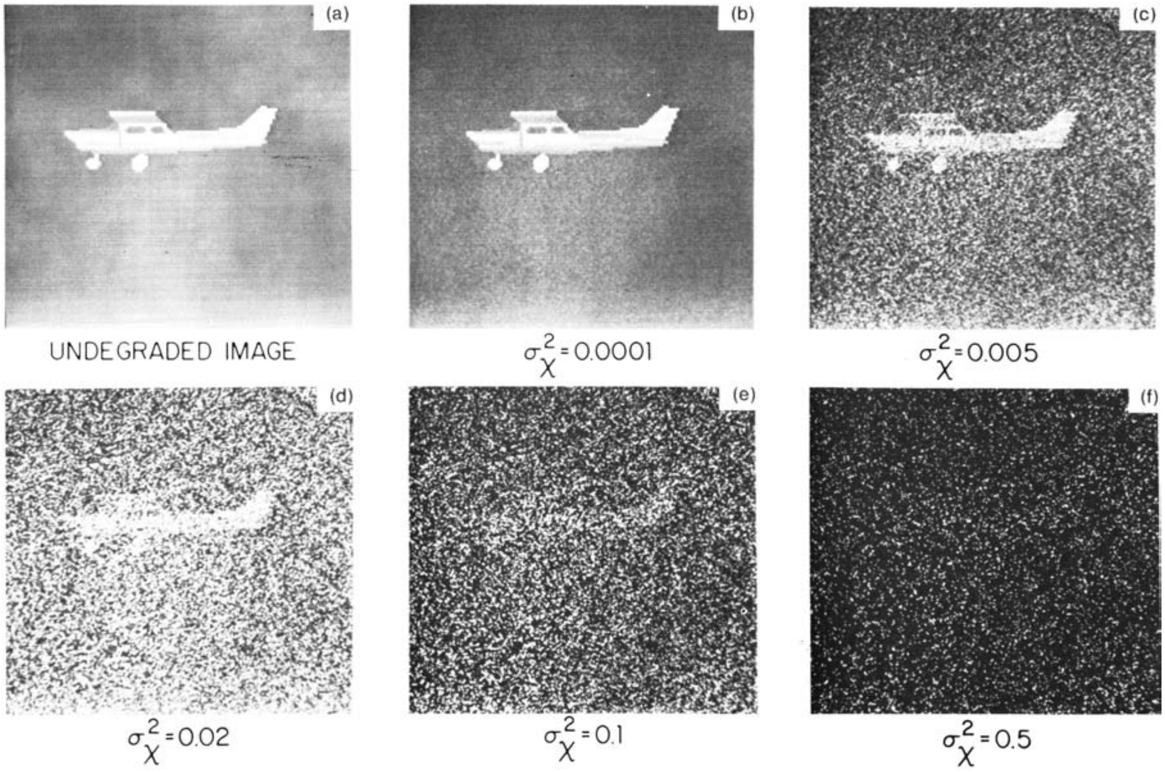


Figure 14-14. Simulated glint target images showing original image and turbulence-degraded images for five values of the log-amplitude variance.[14]



UNDEGRADED IMAGE

$\sigma_\chi^2 = 0.0001$

$\sigma_\chi^2 = 0.005$

$\sigma_\chi^2 = 0.02$

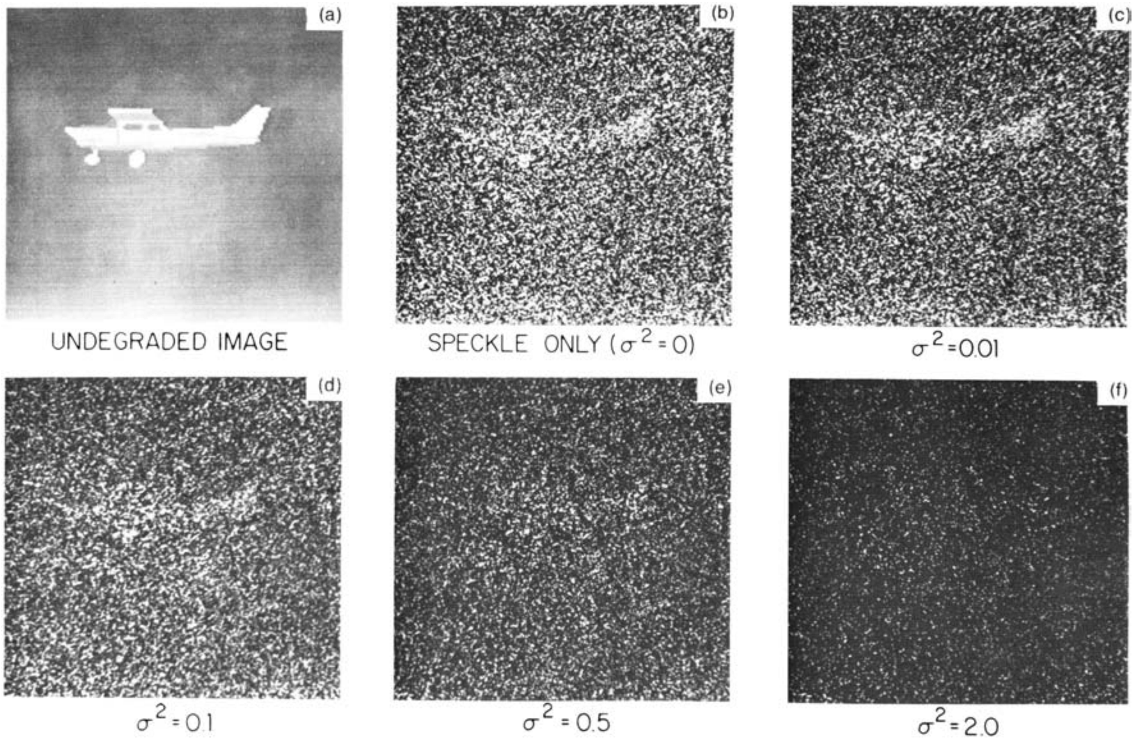
$\sigma_\chi^2 = 0.1$

$\sigma_\chi^2 = 0.5$

image is then multiplied by a random number $e^{2\chi}$ where χ is chosen from a Gaussian distribution with mean - σ_χ^2 and variance + σ_χ^2 . As the turbulence increases, the target remains recognizable until the log-amplitude variance becomes quite large ($\sigma_\chi^2 > 0.1$). However, by the time the turbulent scintillation reaches saturation, the image has virtually disappeared into what appears to be a field of stars. It should be noted that if the target had been imaged against zero background, the target would have remained detectable. Scintillation is unable to produce bright spots if the reflectivity is zero. The few bright spots produced by the target would easily stand out against the non-existent background. "Hot spot" detection would essentially replace image detection that obeyed Johnson's criteria. In these synthetic images, the number of resolvable spots across the typical target dimension is approximately 35. The corresponding value of N/N_{50} ~24 is so large that we would expect to detect the target even if the signal-to-noise ratio were quite small. This appears to be the case. Although we will not examine the detection transition in detail. However, it appears that these results provide additional qualitative support for the information-based perception theory.

Similar images can be generated for speckle targets. In this case each pixel is multiplied by two random numbers, one selected from a unit-mean exponential distribution (speckle) and one selected from a log-normal distribution (scintillation). These images are reproduced in Figure 14-15.

Figure 14-15. Simulated speckle target images showing the original image, the image with speckle only and turbulence-degraded images for four values of the log-amplitude variance.[14]



The image quality remains almost constant until the log-amplitude variance exceeds 0.1. Then it degrades rapidly until the image is unrecognizable at saturation of scintillation.

Fluctuations in traditional radar detection could often be improved by pulse integration. The equivalent process in imaging is frame-averaging in which the results from *N* separate images are added on a pixel-by-pixel basis and then normalized by the number of frames averaged. That is, if I_k(i,j) denotes the intensity of the pixel in the ith row and the jth column of the kth image, then the i,jth pixel of the frame-averaged image is given by

$$I(i,j) = \frac{1}{N}\sum_{k=1}^{N} I_k(i,j) . \tag{14.51}$$

The expression for the signal-to-noise ratio of a frame-averaged image is [14]

$$SNR(N) = \frac{N \cdot CNR/2}{1 + \dfrac{N \cdot CNR}{2\ SNR_{SAT}(N)} + \dfrac{1}{2\ CNR}} \tag{14.52}$$

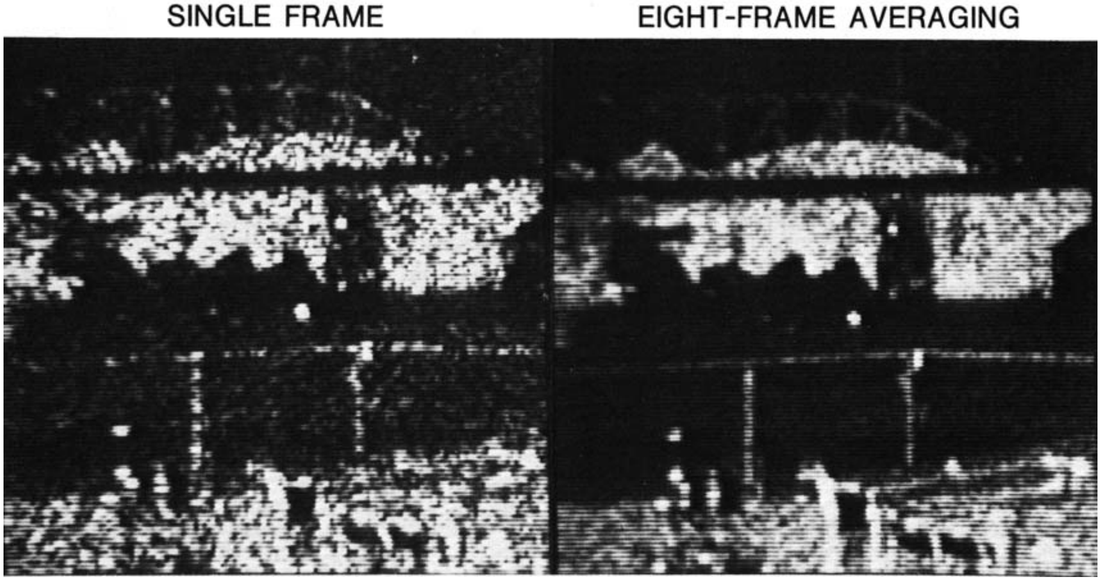
where the frame-averaged saturation signal-to-noise ratio is given by

$$SNR_{SAT}(N) = N \cdot SNR_{SAT}(1) \tag{14.53}$$

Essentially frame-averaging improves the signal-to-noise ratio by roughly the number of frames that are averaged.

Although it would take an inordinate amount of frame-averaging (thousands of frames averaged) to significantly improve images obtained at saturated scintillation, moderate amounts of frame averaging (of the order of 10 frames averaged) would dramatically improve images that are only moderately degraded by scintillation. Speckle-degraded images can also be improved by frame averaging. Figure 14-16 shows a single-frame of laser radar data clearly exhibiting speckle and an image made by averaging eight frames of laser radar data. The speckle effect is clearly reduced producing a much more recognizable image. Please note that the original images contained only eight bits of intensity data. Given that the original scene had components spread over a dynamic range of at least 14 bits, the dynamic range compression coupled with nonlinearities in the reproduction process make it difficult to see the detail present in the darker areas of the original images.

Figure 14-16. Comparison of a single frame laser radar image showing the degradation due to laser speckle (Rayleigh fading) and the improvement obtained by averaging 8 frame of data.



References

[1] Harney, R. C., “Dual active/passive infrared imaging systems”, *Optical Engineering*, 20, #6, 976-980 (December 1981).

[2] Pike, John, “Coverage - Resolution and Swath”, Federation of American Scientists (9 June 1996). Pictures from Itek Optical Systems. Available on the Internet at http://www.fas.org/irp/imint/cover3.htm.

[3] Harney, R. C., “Information-based approach to performance estimation and requirements allocation in multisensor fusion for target recognition”, *Optical Engineering*, 36, #3, 789-798 (March 1997).

[4] Automatic Target Recognizer Working Group, “Target Recognizer Definitions and Performance Measures”, ATRWG No. 86-001 (February 1986).

[5] Rosell, F. A. and Willson, R. H., "Performance Synthesis of Electro-Optical Sensors", U. S. Air Force Avionics Laboratory Report AFAL-TR-74-104, April 1974.

[6] Gonzalez, R. C. and Wintz, P., Digital Image Processing (Addison-Wesley, Reading MA, 1977) pp. 26-27.

[7] Johnson, J., "Analysis of Image Forming Systems", in Image Intensifier Symposium, Warfare Vision Branch, U. S. Army Engineer Research and Development Laboratories, Fort Belvoir VA, pp. 249-273, 6-7 Oct. 1958.

[8] Ratches, J., Lawson, W. R., Obert, L. P., Bergemann, R. J., Cassidy, T. W., and Swenson, J. M., "Night Vision Laboratory Static Performance Model for Thermal Viewing Systems", U. S. Army Electronics Command Report ECOM Report 7043, Fort Monmouth NJ, 1975.

[9] Scott, L. and D'Agostino, J., "NVEOD FLIR92 thermal imaging systems performance model", in Infrared Imaging Systems: Design, Analysis, Modeling and Testing III, G. C. Holst, Ed., *Proceedings of the SPIE*, 1689, pp. 194-203 (1992).

[10] Friedman, M., Tomkinson, D., Scott, L., O’Kane, B. and D’Agostino, J., “Standard night vision thermal modeling parameters”, in Infrared Imaging Systems: Design, Analysis, Modeling and Testing III, G. C. Holst, Ed., *Proceedings of the SPIE*, 1689, pp. 204-212 (1992).

[11] Howe, J. D., “Thermal imaging systems modeling – present status and future challenges”, in Infrared Technology XX, B. F. Andresen, Ed., *Proceedings of the SPIE,* 2269, pp. 538-550 (1994).

[12] Air Standardization Coordination Committee, “Image Interpretability Rating Scale”, Air Standard 101/11B (1 March 1990).

[13] Air Standardization Coordination Committee, “Minimum Ground Object Sizes for Imagery Interpretation”, Air Standardization Agreement 101/8 (31 December 1976).

[14] Shapiro, J. H., Capron, B. A., and Harney, R. C., “Imaging and target detection with a heterodyne-reception optical radar”, *Applied Optics*, 20, #19, 3292-3313 (1 October 1981).

Problems

14-1. Acoustic daylight is a term used to describe imaging in which a noise source (e.g., wave noise) acts as a directional form of illumination. Reflections of this illumination from targets can be detected by sonars. What sort of image (physical property vs. dimension types) might be formed by an acoustic daylight system?

14-2. An electro-optical imaging sensor has 1000 x 1000 detector elements. If the ground sampled distance is 1 meter, how big is the field of view of the sensor?

14-3. Image plane scanners can realistically scan through angles of 90° x 90°. If the sensor with such a scanner must cover a 10° x 10° field of view, what is the maximum allowable magnification of the afocal telescope through the image is observed?

14-4. Name six levels of image-based perceptual performance. Rank these levels in order of difficulty or amount of target information provided.

14-5. What are the current numerical values of Johnson's criteria for detection, recognition, and identification?

14-6. Arrange the following in order of increasing maximum range at which they can do target recognition (neglecting atmospheric effects):

a. A 10 μm laser radar with a 1 m diameter aperture?

b. A 5 μm thermal imager with a 50 cm aperture?

c. A 1 μm television with a 20 cm aperture?

14-7. A new destroyer is to be fitted with a 3-5 μm thermal imager for target identification. If the infrared aperture is limited to 15 cm due to stealth requirements, what recognition and identification ranges can be expected against the following targets using the newest versions of Johnson's criteria:

a) patrol boat (bow aspect -- 8 m wide x 8 m high

b) floating mine (0.8 m wide x 0.4 m high)

c) small boat (bow aspect -- 2.7 m wide x 0.7 m high)

d) frigate (bow aspect -- 14 m wide x 10 m high)

What is the probability of identifying any of the targets if 24 resels can be obtained across the target?

14-8. If only aperture size is a variable, which thermal imaging band is more favorable for high resolution imaging (3-5 μm or 8-12 μm)? Why? The number of resolvable cycles across a target is 18, what is the probability of identification? What is the probability of recognition?

14-9. An observation post at the Korean DMZ requires a TV system to identify targets at 15 km range. The primary criteria are detection and recognition. Determine the minimum size aperture for a 0.7-0.9 μm camera if the nominal target size is 5 meters. How many resolvable cycles across the target are required to detect this target with 95% probability. Is this detection consistent with a camera capable of recognition at 50% probability?

14-10. Which is more likely to permit recognition at a longer range: a visible camera with a 25 mm aperture lens or a 10 μm thermal imager looking through a telescope with a 1 meter diameter? Atmospheric propagation effects are assumed to be negligible.

14-11. A thermal imager has an MRT function given by: MRT (in K) = f (in cy/mr) in both horizontal and vertical directions. Plot the probability of detection and probability of recognition vs range (every 1km from 0km to 10km) for a 3m x 3m target with a 3K ΔT and atmospheric extinction coefficient of 0.22314km^{-1}. Use the latest NVL numbers for Johnson's criteria.

14-12. An imager for use on a remotely-piloted vehicle (RPV) must produce IIRS 6 quality imagery over a down-looking scan swath that is 60° wide. If the RPV flies at 15 km altitude, what is the worst case angular resolution that the imager can tolerate? If the imager wavelength is 10 μm, what is the minimum aperture size that is acceptable?

14-13. Assuming an intensity-imaging laser radar can perform 8-frame averaging, what single factor causes the recognition performance of the laser radar to be considerably better than the performance of a thermal imager operating at the same wavelength and with the same aperture size?

14-14. A spotlight synthetic aperture radar (SAR) can target multiple independent images of a target. The information-theoretic Johnson's criteria say that if SNR can be improved, then the resolution required for target identification can be relaxed. If 16 SAR images can be obtained and averaged, what range-cross range resolution product would be compatible with recognition of a 3 x 7 meter target?

14-15. A laser radar can integrate up to 100 frames of target data. If a saturation SNR of 4 is required for a recognition function to be performed, what turbulence log-amplitude variance can be tolerated?

CHAPTER 15

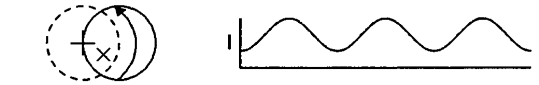
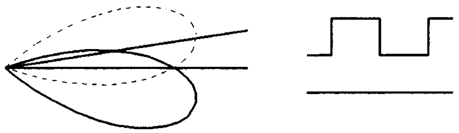
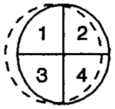
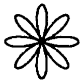
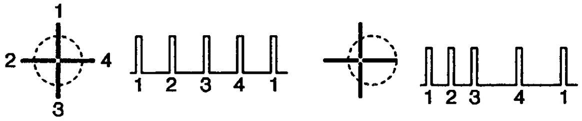
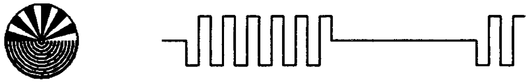
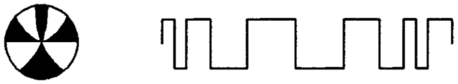
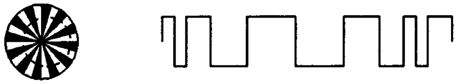
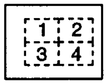
TRACKING

Tracking Systems

The last "fundamental" sensor function is tracking. It is a higher order function than detection or estimation, involving multiple detection and estimation processes. However, the function is so important and involves such different processes that it deserves separate discussion. Tracking is performed by tracking systems. The performance of such systems is usually differentiated in three distinct way: the characteristics of the signal being tracked, the size of the target being tracked relative to the tracking window, and the number of targets being tracked. Signal characteristics include acoustic vs. electromagnetic, active vs. passive, coherent vs. non-coherent, high frequency vs. low frequency. Targets which are unresolved by the tracking sensor are called **point targets**. Point targets produce signals at a single possible tracking location. Targets which are resolved or large enough to produce signals over a range of possible track positions are called **extended targets**. In general, tracking systems that perform well against point targets do not perform as well against extended targets, while tracking systems that perform well against extended targets do not perform well against point targets, although there are a few exceptions. Many tracking systems work well against a **single target** but are incapable of addressing multiple targets. Other tracking systems can handle **multiple targets** as easily as they handle single targets.

In Table 15-1, a number of single-target tracking systems are described. Note: some of these techniques will be described in more detail in later chapters. Here we desire only to summarize the different major approaches to tracking and to elucidate certain key underlying principles. Conical scan trackers are commonly used in radar applications although the technique may in fact be utilized by almost any signal type.[1] Conical scan trackers scan their response pattern in a tight cone. Any "target" at the center of the "con-scan" produces no modulation in the received signal. Targets not at the center produce an amplitude modulation that is proportional to the distance away from the center. The phase of the modulation can be used to establish the angular (azimuthal) position of the target. Extended targets tend to smear out the response reducing the amplitude modulation and degrading the track. Thus con-scan trackers are poor extended target trackers. Sequential lobing is a tracking technique employed predominantly in the radar regime.[1] It can be considered to be a conical scan that has first been reduced to a one dimensional scan and then reduced to a digital scan between angular positions. Any target not located at the midpoint between the two positions will produce a square-wave amplitude modulation. As in con-scan systems, the larger the angular misalignment, the large the amplitude modulation. Monopulse tracking systems are like two (orthogonal) sequential lobing systems combined and looking at all four lobes at once.[1] Either the phase or the amplitude of the incoming signal may be measured in each of the four channels. Error signals are generated by taking sums and differences of the four lobes (as indicated). Phase

Table 15–1. Single target tracking techniques.

TECHNIQUE		MECHANICS	APPLICABLE TO					
			RADAR	IR/VIS SENSOR	ACOUST SENSOR	PASSIVE RF	POINT TARGET	EXT'D TARGET
CONICAL SCAN			YES	POSS.	POSS.	POSS.	GOOD	POOR
SEQUENTIAL LOBING			YES	POSS.	POSS.	POSS.	GOOD	POOR
MONOPULSE (QUADRANT DETECTOR)	•PHASE	 $\Delta X = (1+3) - (2+4)$	YES	LADAR	YES	YES	GOOD	POOR
	•AMPLITUDE	$\Delta Y = (1+2) - (3+4)$	NO	YES	NO	POSS.	GOOD	POOR
ROSETTE SCAN		 PSEUDO-IMAGING	NO	YES	NO	NO	GOOD	FAIR
CON-SCANNED CROSS ARRAY			NO	YES	NO	NO	GOOD	POOR
RETICLE •SPIN-SCAN	- AM		NO	YES	NO	NO	GOOD	POOR
	- FM		NO	YES	NO	NO	GOOD	POOR
•CON-SCAN	- FM		NO	YES	NO	NO	GOOD	POOR
IMAGE •CENTROID	-DIGITAL	CALCULATE INTENSITY CENTROID	NO	YES	NO	NO	FAIR	GOOD
	-BINARY	THRESHOLDED (BINARY) IMAGE CENTROID	NO	YES	NO	NO	FAIR	GOOD
	-GATED VIDEO	 $\Delta X = (1+3) - (2+4)$ $\Delta Y = (1+2) - (3+4)$	NO	YES	NO	NO	FAIR	GOOD
•CORRELATION		AGAINST PRIOR IMAGE OR REFERENCE	NO	YES	NO	NO	FAIR	GOOD
•EDGE		EDGE DETECT, THEN CENTROID TRACK	NO	YES	NO	NO	POOR	GOOD
•FEATURE-BASED		TARGET RECOGNIZER	NO	YES	NO	NO	FAIR	GOOD

monopulse is the most accurate and the most widespread technique. However, most visible and infrared sensors cannot implement phase monopulse techniques. They can implement amplitude monopulse in the form of a quadrant detector array.

Rosette scan is a pseudo-imaging technique (pseudo-imaging in the sense of covering a large angular field at high angular resolution – enough to create an image – without actually forming or using an image).[2],[3] A single detector is scanned in a rosette (flower-like) pattern. Target location is determined by the times at which target detections are made. Rosette scanning is used almost exclusively in the visible and infrared regions. A similar time based approach uses four long thin detector elements arranged in an array. When a scene is conically-scanned around the four arrays, the timing of individual detections in the four detections can yield both the radial and azimuthal positions.

Reticles are optical elements that act to produce light/dark modulations.[2]-[5] Some reticles are transparent with the light/dark modulation produced by alternating opaque absorbing or reflecting regions with transparent regions. Other reticles operate in reflection with the light/dark modulation being produced by alternating reflecting regions with absorbing or transparent regions. Reticle-based scanning can take several different forms. Either the scene can be conically scanned around a fixed reticle (con-scan reticles) or the scene can be imaged onto a spinning reticle (spin-scan reticles). Depending on their design, either AM or FM modulation can be produced by reticle systems. Appropriate demodulation can uniquely determine the radial and azimuthal position of a target.

All of the preceding techniques tend to work better against point targets than against extended targets. They are also only capable of tracking a single target at a time. Imaging trackers tend to work better against extended targets than against point targets, although they can track point targets fairly easily. Imaging systems are also conceptually capable of tracking multiple targets simultaneously, although this capability must be explicitly included in the tracking design.

The simplest form of imaging tracker determines the centroid of the target and tracks the motion of the **centroid**. The centroid (*I, J*) may be determined digitally from the image *I*(*i, j*) by the relations

$$I = \mathrm{Int}\left(\sum_{i,j=1,1}^{N,M} i\, I(i,j) \Big/ \sum_{i,j=1,1}^{N,M} I(i,j) \right) \tag{15.1}$$

and

$$J = \mathrm{Int}\left(\sum_{i,j=1,1}^{N,M} j\, I(i,j) \Big/ \sum_{i,j=1,1}^{N,M} I(i,j) \right) \tag{15.2}$$

where *i* and *j* are indices denoting specific pixel locations in the image. If the target is small compared to the field of view, it is desirable to gate out most of the background to avoid having the background bias the centroid. Ideally, the image processing system would **segment** the image, i.e., identify which pixels are background and which pixels are target, and carry out the integration only over the target pixels. An adaptation of centroid tracking has the image processing system generate a **binary image** (an image whose intensity at any pixel is either 0 or 1 – typically determined by comparing each pixel intensity against a specified threshold) before calculating the centroid. Binary image formation could be performed before or after normal segmentation. Binary image formation can act as a form of image segmentation; it will definitely prevent bright areas on the target from biasing the centroid. An analogue form of centroid tracking can be accomplished by using gated integrators on the video signal. In a **gated video** tracker, four integrators integrate different parts of the video signal. Timing of the integrator gates is such that the four integrators each integrate on quadrant of a rectangular area containing the target. The four integrators act just like a quadrant detector (amplitude monopulse) system. The centroid can be determined from the sums and differences of the integrated signals recorded by the four integrators.

Another image tracker has the image processing system segment the image and then apply an edge detector to the segmented target. The centroid of the edges is then tracked. Proponents of **edge trackers** claim superior performance. The author has not seen convincing justification for these claims.

A different image tracking technique is **correlation tracking**. In a correlation tracker, the image *I*(*i, j*) is cross correlated with a reference image *R*(*i, j*) producing a correlation function *C*(*i', j'*)

$$C(i',j') = \sum_{i,j=1,1}^{N,M} I(i,j) R(i-i', j-j') . \tag{15.3}$$

The coordinates (*i', j'*) at which the cross correlation function is a maximum describes the shift (in pixels) between the image and the reference. If the assumption is made that any shift is due to target motion, then this shift is also the number of pixels that the imager must be "moved" to reposition the target at the "reference point". An *a priori* stored image could be used as the reference. However, in most systems, the reference image is the immediately preceding frame of image data. This has the benefit of allowing slow variations in target scale and orientation to be accommodated. The frame-to- frame changes will be small enough that the peak of the correlation function is only minimally affected.

Feature-based trackers assume the existence of an automatic recognizer. For the most part they use one of the image-based tracking schemes described above. However, they also tag each track file with key identifying features and compare the features in each new image frame with those of the prior frames. This permits a target to be tracked even if it passes behind an obscuration. The loss of the key features indicates the target has disappeared and the tracker should go into a "**coast mode**" in which it follows the last valid tracking commands and gradually loosens up any tracking gates. If the target reappears from behind the obscuration, the reappearance of the identifying features, guarantees that the proper target is being tracked and a new track file is unnecessary.

If properly designed image-based trackers can handle tracking of multiple targets. This is not true of the other tracking techniques in Table 15-1. They can only handle multiple targets by fielding multiple tracking systems (one per target to be tracked). In Table 15-2 the techniques amenable to multiple target tracking are described. As with single target trackers, some multiple-target tracking techniques work better against point targets; others work better against extended targets.

The dominant multiple-target tracking technique is **track-while-scan (TWS)**. The name accurately describes the way it works. The sensor scans its field of view continuously. No special attention or action is required to permit tracking. After a scan of the complete field of a set of "detections" or threshold crossings has been established. On the next scan, each detection is associated with a detection in the set from the preceding scan. If an old detection cannot be associated with a current detection, it becomes a candidate to be dropped as a current **track file**. If a new detection cannot be associated with and old detection, it becomes a candidate for establishment of a new **track file**. After several scans and associations, enough data exists to initialize a **tracking filter** for each set of scan-to-scan persistent detections. On successive scans, detections are associated with track files (assisted by predicted positions from each tracking filter) and used to update the tracking filters, or they are tagged to potentially establish new track files.

Track-while-scan can be implemented in a number of forms. **Azimuth-only TWS** can be generated by radio-frequency interferometers (direction finders), passive sonars, and radars that are being affected by range-denial jammers. Most surface search radars and some air search radars can be used to perform **azimuth-plus-range TWS**. Infrared search and track systems without associated rangefinders perform **azimuth-plus-elevation TWS**. Some sophisticated air defense radars, aircraft fire control radars, laser radars, and infrared search and track systems with associated rangefinders there is the ability to perform **three-dimensional TWS** (range-plus-azimuth-plus-elevation).

Systems with phased array antennas have the potential to do multiple target tracking using **interferometric beam forming**. Using either the whole array or a small subset of the whole array an antenna lobe can be constructed in any desired direction by properly shifting the phases of the signals received by the array elements involved. An object in such an antenna lobe can be tracked by the equivalent of a monopulse technique or by a scanning technique analogous to sequential lobing or conical scanning by suitable modulation of the phases and observing the modulations they produce in the signal. As target motions are detected, the element phases may be varies to keep the target in the lobe. As more targets need to be tracked, more antenna lobes can be synthesized. This is easier to say than it is to do. If the entire array is used to synthesize the tracking lobes, then gain will be reduced as the number of lobes increases. If a subset is used for each lobe, then each lobe will suffer a gain reduction proportional to the fraction of the array elements used.

Most image-based tracking techniques can be used to track multiple targets just as they can track single targets. Point targets can be detected in images and tracked using tracking filters just as in an azimuth-plus-elevation TWS system. The limitation on the number of extended targets that can be tracked by other imaging techniques depends on the amount of image processing power that can be brought to bear. Calculation of a correlation function is computationally intensive. The degree of computation difficulty will scale with the number of targets that must be processed.

Table 15–2. Multiple target tracking techniques.

TECHNIQUE		MECHANICS	APPLICABLE TO					
			RADAR	IR/VIS SENSOR	ACOUST SENSOR	PASSIVE RF	POINT TARGET	EXT'D TARGET
TRACK-WHILE-SCAN	•AZIMUTH ONLY	•DIRECTIONAL SENSOR BEAM IS REPEATEDLY SWEPT OVER THE SEARCH FIELD OF VIEW AND ALL DETECTIONS ARE STORED.	POSS.	NO	YES	YES	GOOD	FAIR
	•AZIMUTH + RANGE	DETECTIONS ARE CORRELATED WITH THOSE OF PRIOR SWEEPS TO ESTABLISH TRACK FILES.	YES	NO	NO	NO	GOOD	FAIR
	•AZ-EL	TRACKING FILTER (ALPHA-BETA-GAMMA OR KALMAN) PROVIDES CONTINUOUS ESTIMATES OF TARGET STATE.	POSS.	YES	POSS.	POSS.	GOOD	FAIR
	•3-D (AZ-EL-RG)		YES	YES	NO	NO	GOOD	FAIR
INTERFEROMETRIC BEAM-FORMING		•A TRACKING BEAM IS SYNTHESIZED FOR EACH TARGET BY PHASING AN ARRAY ANTENNA. THE POSITION OF EACH TARGET IS TRACKED BY CON-SCAN, SEQUENTIAL LOBING, OR OTHER TECHNIQUE.	YES	NO	YES	YES	GOOD	POOR
IMAGE	•MULTIPLE POINT	•SIMILAR TO TRACK-WHILE-SCAN USING THRESHOLDED IMAGE FRAMES	NO	YES	NO	NO	GOOD	POOR
	•CENTROID	•MULTIPLE CENTROID TRACKING GATES ARE ESTABLISHED IN THE IMAGE	SAR?	YES	NO	NO	FAIR	GOOD
	•EDGE	•MULTIPLE EDGE TRACKING FILTERS ARE APPLIED TO THE IMAGE	SAR?	YES	NO	NO	POOR	GOOD
	•FEATURE-BASED	•ALL REGIONS WITH TARGET-LIKE CHARACTERISTICS ARE SEGMENTED AND TRACKED WITH CENTROID OR EDGE TRACKERS. FEATURES EXTRACTED FROM REGION AID THE FRAME-TO-FRAME CORRELATION.	SAR?	YES	NO	NO	POOR	GOOD

Tracking Filters

Tracking filters are mathematical post-processing that is performed to improve a tracking systems ability to estimate the state vector of targets. The state vector $\vec{X}$ is a quantity describing the dynamics of the target at the level of concern to the problem at hand, i.e.,

$$\vec{X} = \left\{ X_1, X_2, \cdots, X_N \right\} \tag{15.4}$$

where the X_i are elements of dynamic motion. The simplest form of state vector might be the location of the target and its velocity. If the sensor were only capable of measuring angle to the target, the state vector might consist of the target angle and the angular velocity,

$$\vec{X} = \left\{ \theta, \dot{\theta} \right\} \tag{15.5}$$

where the dot denotes a derivative with respect to time. If acceleration were deemed important to the problem, the state vector would consist of the target angle, the angular velocity, and the angular acceleration

$$\vec{X} = \left\{ \theta, \dot{\theta}, \ddot{\theta} \right\} . \tag{15.6}$$

If the system were able to measure range and angle, then the state vector might be

$$\vec{X} = \left\{ R, \theta, \phi, \dot{R}, \dot{\theta}, \dot{\phi}, \ddot{R}, \ddot{\theta}, \ddot{\phi} \right\} . \tag{15.7}$$

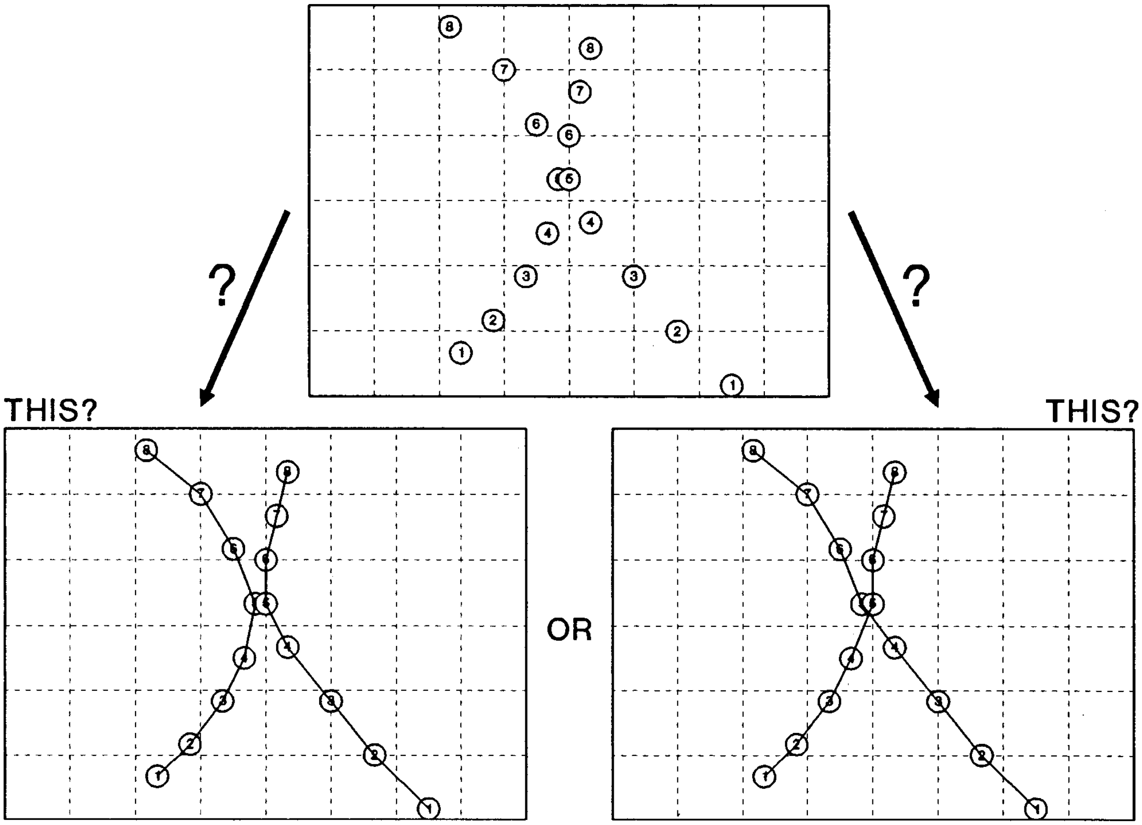
The state vector for a target maneuvering relative to a sensor on a maneuvering platform might get very complicated.

Track-while-scan systems suffer from three fundamental difficulties: track precision limitation, track loss due to missed detections or false alarms, and difficulty in frame-to-frame target association. In any real sensor system there will be errors in measuring the state vector. Without a track filter, the track error can be no smaller than the measurement error (and this is often not good enough). Note that it is not really possible to completely eliminate a tracking filter. If a human being looks at the data, even if only to connect the dots on a plot, that person will act as a tracking filter. However, a computer-based filter can usually improve performance more than the human.

Track-while-scan systems rely on the ability to detect an object time and again and associate a detection in one data interval (e.g., an image frame which equals one scan of the field of view) with a detection in another interval. If the object is not detected in a frame, what decision is to be made? Did the target exit the scene? Did it crash (not an impossible outcome)? Did the sensor simply fail to detect the target? What if a false alarm (remember that the sensor cannot tell a false alarm from a real target) occurs in a frame? Is this a new target? Which of the two detections is the target? Without a tracking filter, it is difficult to make such decisions. With a filter, they may not need to be made at all. The problem lies in the need to associate detections between frames.

Consider the sequence of detections shown in Figure 15-1. Two targets are converging. They appear to be following slightly curve trajectories, but because of sensor noise it is not possible to rule out straight line motion. Which of the two possible sets of associations is correct? A tracking filter might shed light on this question by (1) providing the best estimate available of target position when the critical association decision needs to be made (at frame 5), and (2) by having estimates for higher order terms in the dynamics of the target which would shed light on whether or not a maneuver appeared to be underway. This is not to say that a tracking filter cannot make a mistake in such "pathological" situations, but that is will likely do better than the human processor would do (barring luck or good intuition).

Figure 15-1. A confusing sequence of detections of two targets. Did the target trajectories cross or did they swerve to avoid a potential collision? Which target is which at the end of this sequence?



Consider the plot shown in Figure 15-2. This plot contains 9 frames of track data with 10 real targets. The 10 real targets are detected with a detection probability of 95%. Thus there are roughly 5 missed detections. The plot also has an average of 1 false alarm per frame. The reader is invited to attempt his own fit to the points. The actual tracks are overlaid on this plot in Figure 15-3. It is obvious that in more than one instance, it would be difficult to make an unequivocal association of detections to a track. For example, in the upper right hand corner, just as two tracks cross, there is a missed detection. Which track actually had the missed detection? In the center of the plot there are several track crossings that are actually close encounters. Again the question arises of which detection belongs to which track. Although a mistaken association might not cause a loss of track in this example, it will increase the error in predicting the track. Now consider the problem facing an air controller which may have several hundred targets being tracked simultaneously. It is problems like this that have made tracking filters a necessity.

Consider the generic tracking filter described in Figure 15-4.[6] The actual **target dynamics** will produce a **state vector** $\vec{X}$. The **observation vector** $\vec{Y}$ consists of those components of the state vector that the sensor measures plus any noise

$$\vec{Y} = \mathrm{H}\vec{X} + \vec{N} \tag{15-8}$$

Figure 15-2. A sequence of 9 frames of track data with 10 real targets, 5 missed detections ($P_D = 0.95$), and an average of 1 false alarm per frame.

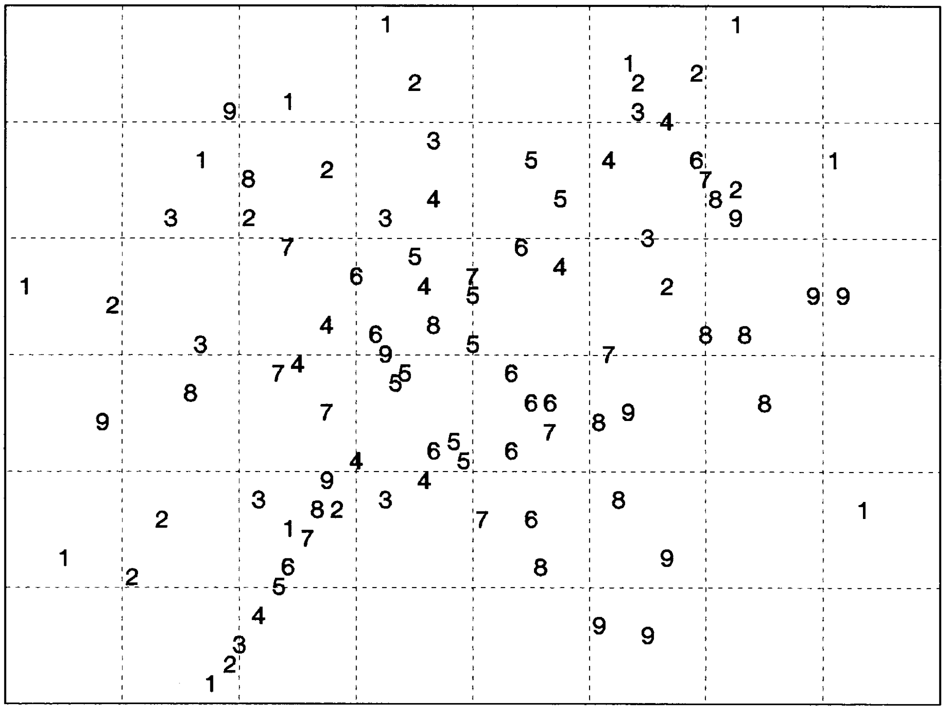


Figure 15-3. The plot of Figure 15-2 overlaid with the actual tracks.

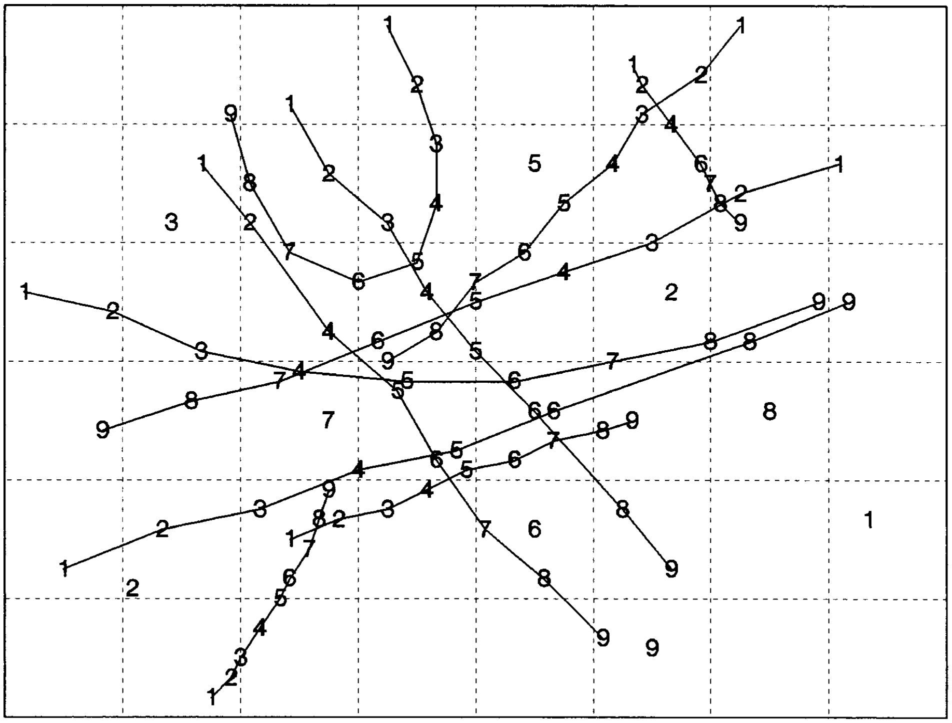
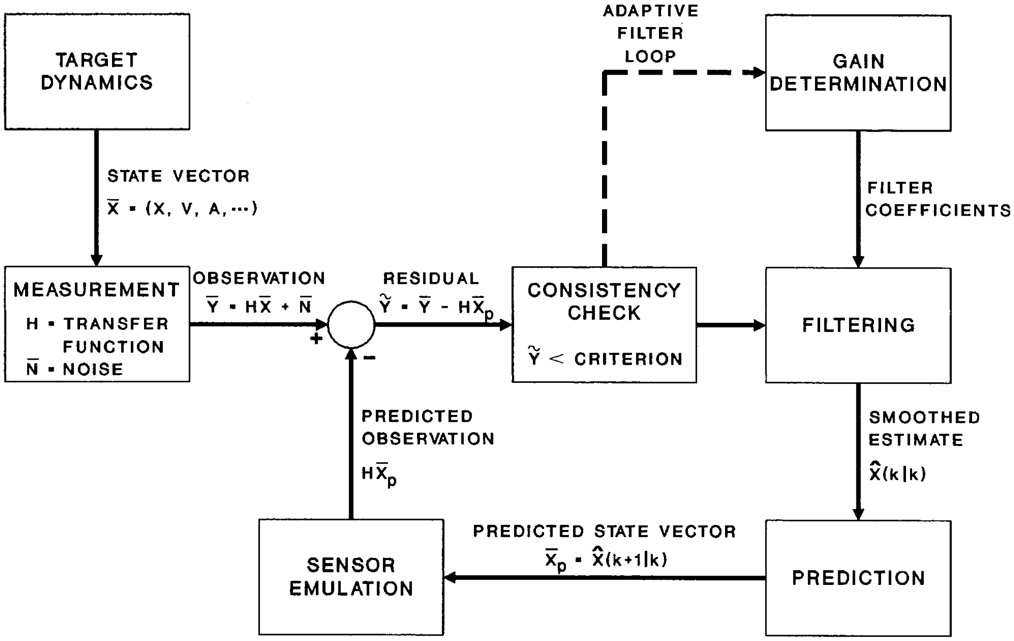


Figure 15-4. Functional flow diagram of a generic tracking filter.



where H is the **transfer function** of the sensor and $\vec{N}$ is the sensor **noise vector**. The measurement transfer function is the mathematical description of the manor in which the sensor translates actual state vector information into measurement information. For example, a pure Doppler radar would transform a state vector by ignoring position and acceleration terms, and then projecting the velocity terms onto the vector describing the line of sight to the target. That is, it converts the three-dimensional velocity into a radial velocity.

A **residual** $\tilde{Y}$ is the difference between a **predicted observation** $H\vec{X}_P$ and the actual observation. The residual is usually checked for consistency; it should be small. If the residual is so large that it cannot reasonably be due to valid target motion (e.g., surface targets seldom travel at Mach 1 and aircraft do not normally travel at Mach 10), then the current observation may be due to a false alarm or a new target. Such inconsistent observations should not be included in the track file. The residual might also be used to adjust filter gains in an adaptive fashion. The predicted and actual observations and the residual are inputs to the filtering processor. In a filter, a fraction of the residual is added to the predicted observation to establish a **smoothed estimate** $\hat{X}(k|k) = \vec{X}_S$ of the current state vector. The fraction of the residual added to the predicted observation is determined by the gain coefficient. Addition of less than the complete residual means that only a portion of any measurement error will be included in the new estimate. Over time, the increased reliance on prediction versus actual measurement will lead to a reduction in the variance of the estimated state vector. This reduction in variance of the estimated position versus the variance in the observed position is called **smoothing**. The smoothed estimate of the state vector is combined with the expected target dynamics (all filters must assume a fixed type of target motion such as uniformly accelerated linear motion, constant velocity circular motion, etc.) to produce a **predicted state vector** $\vec{X}_P$ for the next measurement time. The predicted state vector is transformed using the transfer function to establish the new predicted observation. After a new measurement has been made, the cycle is ready to repeat.

There are two distinct forms that tracking filters can take: fixed-coefficient or adaptive coefficient filters. In a fixed-coefficient filter, the gains (coefficients in the smoothing algorithm) are fixed in advance. In an adaptive coefficient filter the gains are gradually varied up and down depending on the residuals. When the residuals are small, the gain coefficients are reduced until the residuals start to increase. If the residuals are large, the gain coefficients are increased until the residuals stop decreasing after a coefficient increase. The Kalman filter is the most common implementation of an adaptive coefficient algorithm.

As indicated above, all tracking filters assume a set of target dynamics. The filters use the dynamical equations to predict ahead in time to estimate what the new state vector should be based on the old estimate. If the target motion does not match the dynamics for any extended period of time, the prediction will get worse and worse and the residuals will get larger and larger. For example, a common assumption is straight-line, uniformly accelerated motion. The dynamical equations are

$$X(t) = 0.5A(0)t^2 + V(0)t + X(0) \tag{15.9}$$

$$V(t) = A(0)t + V(0) \tag{15.10}$$

$$A(t) = A(0) \tag{15.11}$$

or in terms of a dynamics matrix D and state vector $\vec{X} = \{X, V, A\}$

$$\vec{X}(t) = \mathrm{D} \cdot \vec{X}(0) = \begin{bmatrix} 1 & t & 0.5t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X(0) \\ V(0) \\ A(0) \end{bmatrix} . \tag{15.12}$$

A ballistic missile would be assumed to follow a ballistic trajectory after motor burnout. A non-maneuvering satellite would be assumed to follow a trajectory consistent with orbital dynamics. A target aircraft performing a breakaway maneuver might be expected to follow a uniform centripetally accelerated trajectory for a short period of time. If the precise dynamics cannot be determined then it often becomes necessary to simultaneously run parallel tracking filters with different estimated target dynamics and select the output at each sample that gives the smallest residual.

Fixed-Coefficient Filters

The simplest tracking filters to understand are the fixed-coefficient filters. The simplest fixed-coefficient filters are the *α-β* and *α-β-γ* filters. Before describing these filters in detail, we must define a number of terms:

$\hat{X}(k|j) =$ Estimate of target state vector at sample *k* given samples through *j*

$X_S(k) =$ Smoothed estimate of state vector at sample *k*

$X_O(k) =$ Observed value of state vector at sample *k*

$X_P(k+1) =$ Predicted value of state vector at sample *k+1*

$qT =$ Time since last valid sample (T = sample time interval)

$\alpha, \beta, \gamma =$ Fixed (but selectable) filter coefficients

We will use these definitions throughout this section.

Now let us begin by describing the ***α-β* tracking filter**. This filter assumes simple unaccelerated motion (in a straight line). The dynamical equations are

$$X(t) = V(0)t + X(0) \tag{15.13}$$

and

$$V(t) = V(0). \tag{15.14}$$

Measurements are made only of the X component of the state vector $\vec{X} = \{X, V\}$. The filter algorithm goes as follows [6]:

The smoothed estimate of the position X is developed by adding a fraction α of the residual (X_O-X_P) to the original prediction

$$X_S(k) = \hat{X}(k|k) = X_P(k) + \alpha\left[X_O(k) - X_P(k)\right]. \tag{15.15}$$

A smoothed estimate of the velocity V is obtained by developing a velocity "residual" (position difference divided by time increment) and adding a fraction β of that velocity residual to the prior smoothed estimate of velocity

$$V_S(k) = \hat{\dot{X}}(k|k) = V_S(k-1) + \beta\left[X_O(k) - X_P(k)\right]/qT\,. \qquad (15.16)$$

The predicted position of the next observation is determined by using the dynamical equations and inserting the smoothed estimates of the parameters

$$X_P(k+1) = \hat{X}(k+1|k) = X_S(k) + T\,V_S(k)\,. \qquad (15.17)$$

This prediction is combined with the next observation to determine a residual that begins the process again. This loop is repeated over and again until tracking is no longer required. This iterative algorithm needs initiation. To initiate the loop, one assumes the following initial conditions. The initial smoothed position estimate is assumed equal to the first prediction and both are set equal to the first measurement

$$X_S(0) = X_P(1) = X_O(0)\,. \qquad (15.18)$$

The initial smoothed velocity is assumed to be zero, with the subsequent smoothed estimate being calculated from the finite difference approximation to the first derivative using the two position observations as inputs,

$$V_S(0) = 0 \qquad (15.19)$$

and

$$V_S(1) = \left.\frac{dX_O}{dt}\right|_{(1)} \approx \left[X_O(1) - X_O(0)\right]/T\,. \qquad (15.20)$$

Once initialized the α-β tracking algorithm can be cycled from observation *k* to observation *k*+1 as long as desired.

The ***α-β-γ* tracking filter** assumes uniformly accelerated straight line motion. The dynamical equations are Eq.(15.9) through (15.11). Again measurements are made only of the position component of the state vector $\vec{X} = \{X, V, A\}$. The filter algorithm goes as follows [6]:

The smoothed estimate of the position *X* is developed by adding a fraction α of the residual (X_O-X_P) to the original prediction

$$X_S(k) = X_P(k) + \alpha\left[X_O(k) - X_P(k)\right]. \qquad (15.21)$$

A smoothed estimate of the velocity *V* is obtained by developing a velocity “residual” (position difference divided by time increment) and adding a fraction β of that velocity residual to the sum of the prior smoothed estimate of velocity and the estimated velocity change due to acceleration

$$V_S(k) = V_S(k-1) + qTA_S(k-1) + \beta\left[X_O(k) - X_P(k)\right] / qT\,. \tag{15.22}$$

A smoothed estimate of the acceleration A is obtained by developing an acceleration "residual" (based on the velocity residual divided by the time increment) and adding a fraction γ of that acceleration residual to the prior smoothed estimate of the acceleration.

$$A_S(k) = A_S(k-1) + \gamma\left[X_O(k) - X_P(k)\right] / (qT)^2\,. \tag{15.23}$$

The predicted position of the next observation is determined by using the dynamical equations and inserting the smoothed estimates of the parameters

$$X_P(k+1) = X_S(k) + TV_S(k) + 0.5\,T^2\,A_S(k)\,. \tag{15.24}$$

This prediction is combined with the next observation to determine a residual that begins the process again. This loop is repeated over and again until tracking is no longer required. This iterative algorithm needs initiation. To initiate the loop, one assumes the following initial conditions. The initial smoothed position estimate is assumed equal to the first prediction and both are set equal to the first measurement

$$X_S(0) = X_P(1) = X_O(0)\,. \tag{15.25}$$

The initial smoothed velocity is assumed to be zero, with the subsequent smoothed estimate being calculated from the first two position observations,

$$V_S(0) = 0 \tag{15.26}$$

and

$$V_S(1) = \left[X_O(1) - X_O(0)\right] / T\,. \tag{15.27}$$

The first two initial smoothed accelerations are assumed to be zero, with the subsequent smoothed estimate being calculated from the finite difference equation for the second derivative using the first three position observations,

$$A_S(0) = A_S(1) = 0 \tag{15.28}$$

and

$$A_S(2) = \left.\frac{d^2 X_O}{dt^2}\right|_{(2)} = \left[\left.\frac{dX_O}{dt}\right|_{(2)} - \left.\frac{dX_O}{dt}\right|_{(1)}\right] / T$$

$$. \qquad = \left[\left(\left[X_O(2) - X_O(1)\right]/T\right) - \left(\left[X_O(1) - X_O(0)\right]/T\right)\right]/T . \qquad (15.29)$$

$$= \left[X_O(2) + X_O(0) - 2X_O(1)\right]/T^2$$

Once initialized the α-β-γ tracking algorithm can be cycled from observation k to observation k+1 as long as desired.

How good is the performance of these simple filters? It can be shown that the optimum choice of β given an arbitrary choice of α is

$$\beta = 2(2-\alpha) - 4\sqrt{1-\alpha} \qquad (15.30)$$

for either the α-β or the α-β-γ filters.[4] Given both α and β the optimum choice of γ in an α-β-γ filter is

$$\gamma = \beta^2 / 2\alpha . \qquad (15.31)$$

The optimum β and γ coefficients are shown as functions of the α coefficient in Figure 15-5.

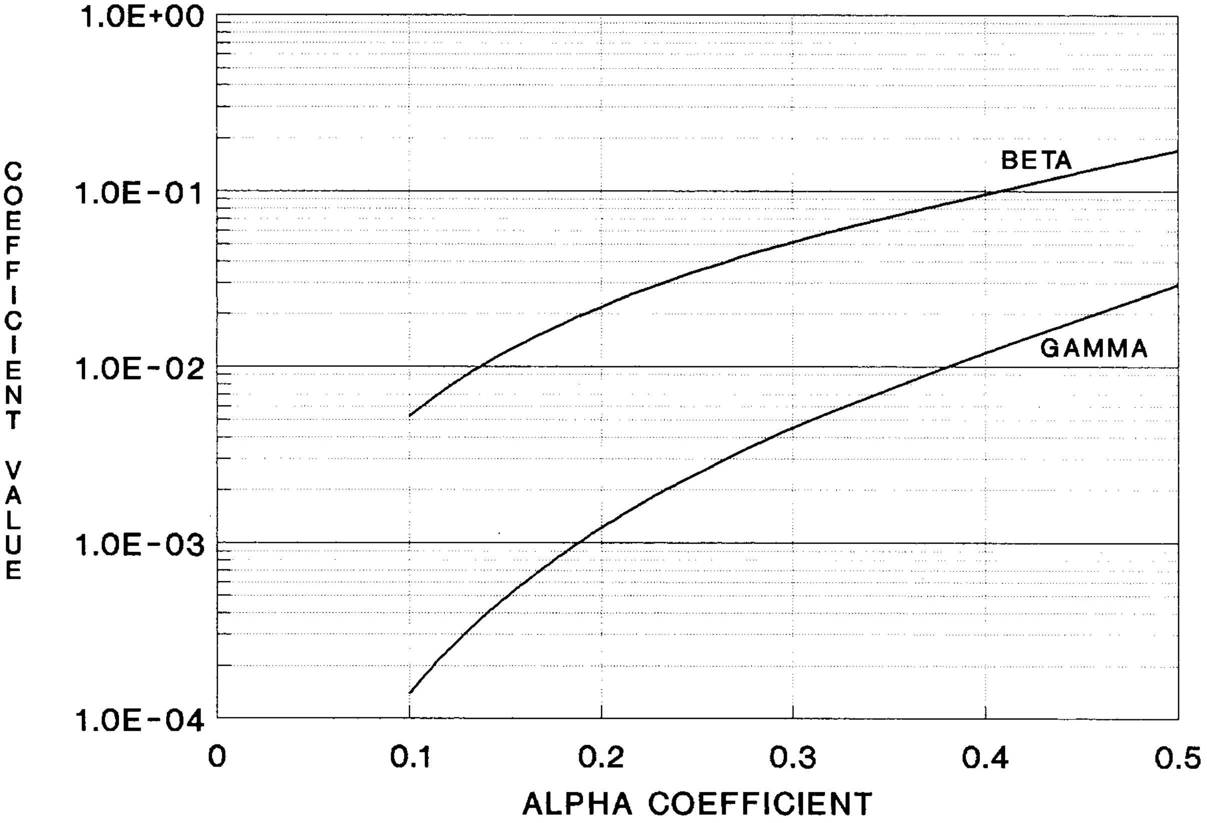
The smoothing efficiency for a component of the state vector is defined as the variance of the smoothed estimates of that component divided by the variance of the observations of that component. In the absence of direct observations of a component, then the finite difference approximation to the derivative defining that component is substituted along with the observations to determine the smoothing efficiencies. For the α-β-γ filter the position smoothing efficiency (variance reduction) is

$$K_X = \frac{\sigma^2_{x_S}}{\sigma^2_{x_O}} = \frac{2\alpha^2 + \beta(2-3\alpha)}{\alpha[4-\beta-2\alpha]} \qquad (15.32)$$

and the velocity smoothing efficiency (variance reduction) is

$$K_V = \frac{2\beta^2}{T^2\alpha[4-\beta-2\alpha]} . \qquad (15.33)$$

Figure 15-5. Optimum choices of beta and gamma coefficients given an arbitrary selection of the alpha coefficient.

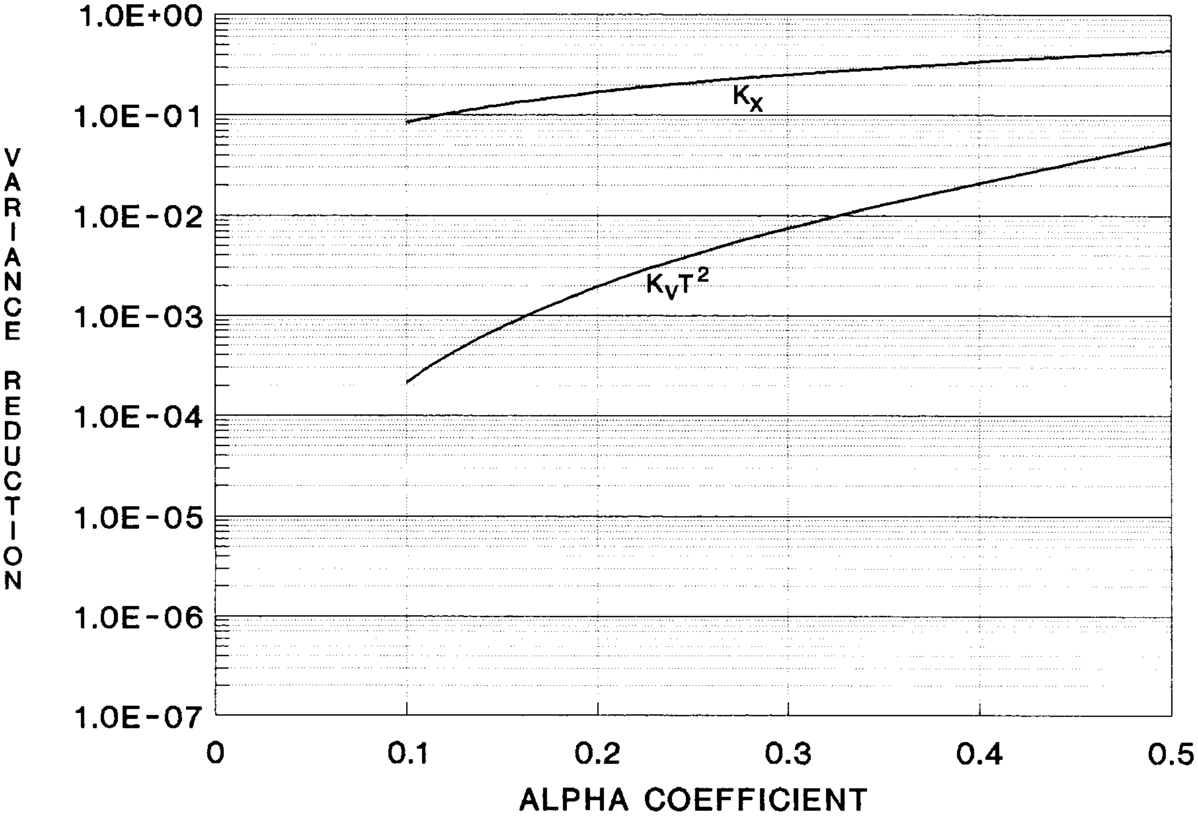


These two smoothing efficiencies are plotted as functions of the yet to be selected α coefficient in Figure 15-6. As α decreases the position smoothing efficiency becomes smaller and smaller (meaning the variance is reduced more and more). It is a quite respectable 10% at $\alpha = 0.1$. The velocity smoothing falls even faster than the position smoothing.

This would suggest that the optimum performance would be obtained by making α as small as possible. Unfortunately, with the good (smoothing) there comes the bad. In a tracking filter, the "bad" takes the form of response time and hangoff error. The gain coefficient α is the fraction of the residual that is added to the prediction to make the smoothed estimate. This means that any observation takes $1/\alpha$ measurement intervals before its effects become negligible. An alternative view would be that it takes $1/\alpha$ measurement intervals before any change becomes apparent. There is a non-zero **response time** t_0 associated with a tracking filter that is roughly given by

$$t_0 = T / \alpha . \qquad (15.34)$$

Figure 15-6. Degree of smoothing (variance reduction) achievable in an α-β tracking filter.



Filters with good smoothing characteristics also tend to have long response times. Response time can become critical if the target has any ability to maneuver or change a parameter of its dynamics. If the target motion changes, the filter will make no appreciable change in its output or predictions until a response time has elapsed. If the motion change is too great, a slow response time can permit the residual to grow to a value deemed unphysical and the track will be lost.

Hangoff is another potential problem. If the filter designer has mis-estimated the dynamics of the target, then hangoff may occur. Consider the α-β filter. If the target actually has a constant non-zero acceleration in motion, then the filter will never accurately predict the future position. The acceleration will cause the target to stay one jump ahead of the tracker. This is hangoff. The hangoff error in an α-β tracker for a constant acceleration $\ddot{X}$ is given by

$$\lim_{k\to\infty}\left[X(k)-X_S(k)\right]=\frac{(1-\alpha)T^2}{\beta}\ddot{X}\,. \tag{15.35}$$

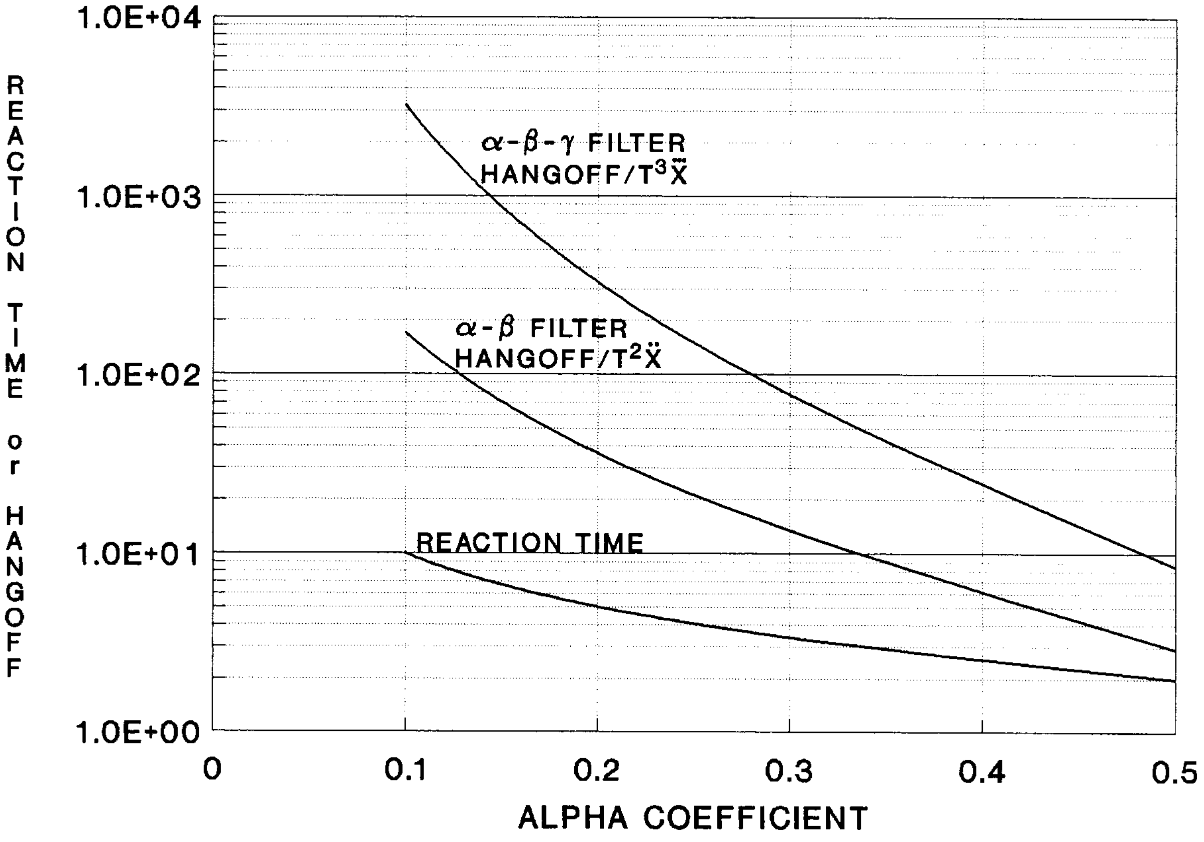
Hangoff error and response time for an α-β tracker are shown in Figure 15-7. Both get worse as α gets smaller. The choice of α must result from a tradeoff between the degree of smoothing desired, the acceptable response time, and the magnitude of any potential hangoff error that may be present. In many cases a choice of $0.1 \le \alpha \le 0.2$ is a reasonable compromise.

The optimum coefficients for the α-β-γ filter are the same as those for the α-β filter and have already been given in Figure 15-5. The smoothing efficiencies for this filter can be determined from

$$K_X = \frac{\sigma^2_{x_S}}{\sigma^2_{x_O}} = \frac{2\beta\left[2\alpha^2 + \beta(2-3\alpha)\right] - \gamma\alpha\left[4-\beta-2\alpha\right]}{\left[4-\beta-2\alpha\right]\left[2\alpha\beta+\alpha\gamma-2\gamma\right]}, \tag{15.36}$$

$$K_V = \frac{4\beta^3 - 4\beta^2\gamma + 2\gamma^2\left[2-\alpha\right]}{T^2\left[4-\beta-2\alpha\right]\left[2\alpha\beta+\alpha\gamma-2\gamma\right]}, \tag{15.37}$$

Figure 15-7. Hangoff errors and reaction time limitations associated with fixed-coefficient tracking filters.



and

$$K_A = \frac{4\beta^2\gamma}{T^4[4-\beta-2\alpha][2\alpha\beta+\alpha\gamma-2\gamma]} . \tag{15.38}$$

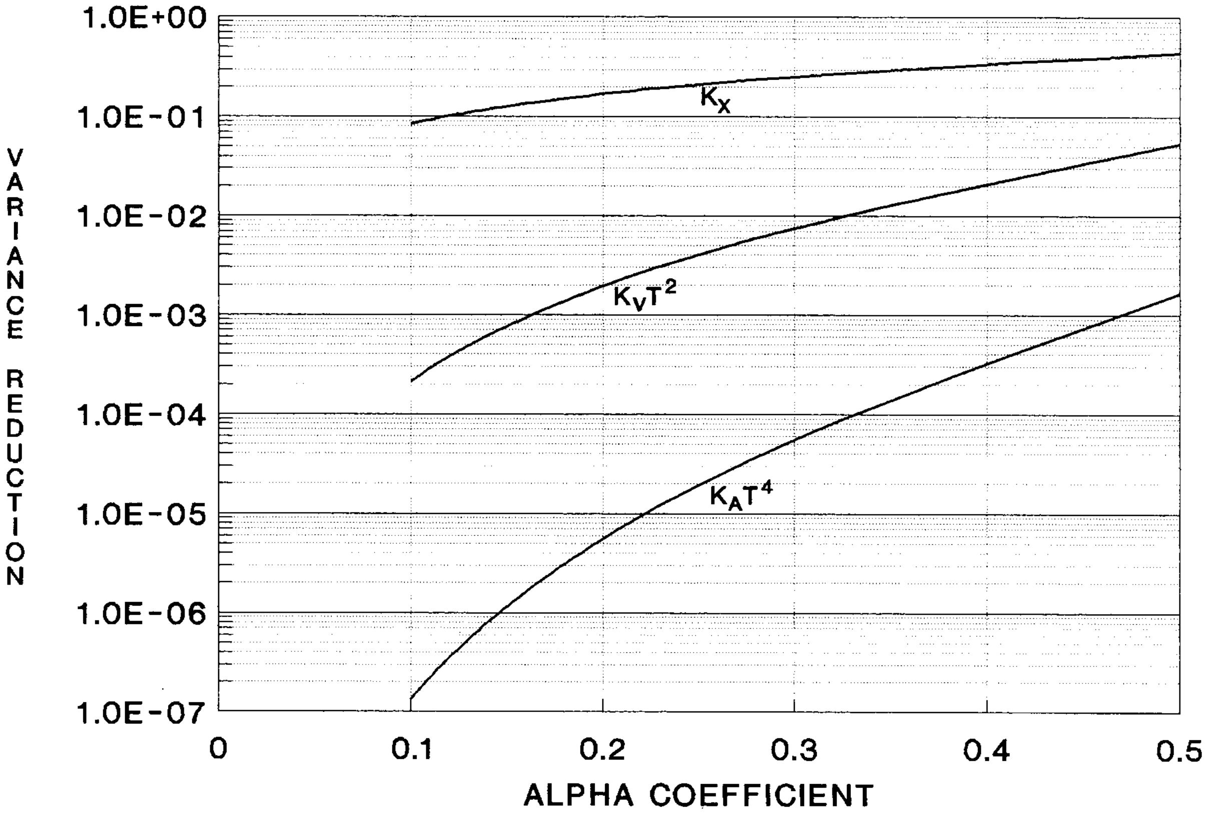
These efficiencies are plotted in Figure 15-8.

As with the α-β tracker the α-β-γ tracker also suffers from response time limitations and hangoff errors. The response time for the two trackers is identical and given by Eq. (15.31). The hangoff error for an α-β-γ tracker following a target with a constant jerk, $\dddot{X}$, is given by

$$\lim_{k\to\infty}[X(k)-X_S(k)] = \frac{(1-\alpha)T^3}{2\gamma}\dddot{X} . \tag{15.39}$$

As before, α must be determined through a trade between smoothing efficiency, hangoff, and response time.

Figure 15-8. Degree of smoothing (variance reduction) achievable in an α-β-γ tracking filter.



Kalman Filters

A Kalman filter is a tracking filter whose coefficients are allowed to vary based on the measured residuals. When residuals are low, the "gains" are increased to provide improved smoothing. When residuals become large, the gains are reduced significantly to improve the filter's ability to respond rapidly to the changes. If properly implemented, Kalman filters will provide optimum smoothing for a given level of sensor noise and disturbance inputs.

In a Kalman filter, the target motion is assumed to be a discrete Markov process. By this it is assumed that knowledge of the state vector of a system at an instant of time is adequate to describe the future behavior of that system. Specifically, the state vector evolves according the equation [6]

$$\vec{X}(k+1) = G\vec{X}(k) + \vec{Q}(k) + \vec{F}(k+1|k) \tag{15.40}$$

where

$$\vec{X}(k) = \begin{bmatrix} X(k) \\ V(k) \\ A(k) \end{bmatrix} \tag{15.41}$$

is the state vector at measurement point *k*. Note we have assumed a specific form for the state vector. The form of the state vector for a real problem will be dictated by the nature of that problem. *G* is a presumed known transition matrix describing the gross dynamics of the target. For example, if we assume that the dynamics are uniformly accelerated linear motion, then

$$G = \begin{bmatrix} 1 & T & T^2/2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix} \tag{15.42}$$

$\vec{Q}$ is a vector (with covariance Q – see below for a definition and Appendix C for a more detailed discussion of the covariance of a random vector) describing noise in the target motion. Often it is assumed to be zero-mean, white, Gaussian noise. A common source of this noise is turbulence buffeting. The vector function $\vec{F}$ is a deterministic input accounting for motion of the tracking platform.

The variance of a function is

$$\mathrm{Var}(X) = E\{(X - E(X))(X - E(X))\}. \tag{15.43}$$

By analogy the covariance of two variables *X* and *Y* is

$$\mathrm{Cov}(X,Y) = E\{(X - E(X))(Y - E(Y))\}. \tag{15.44}$$

The covariance of a vector

$$\vec{X} = \{X_1, X_2, \cdots, X_N\} \tag{15.45}$$

is defined by

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) \tag{15.46}$$

$$\Sigma_{ij} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_N) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_1, X_N) & \mathrm{Cov}(X_2, X_N) & \cdots & \mathrm{Var}(X_N) \end{bmatrix}. \tag{15.47}$$

Sensors measure components of the state vector. They may or may not measure the components directly. The measurement process is modeled by the equation

$$\vec{Y}(k) = H\vec{X}(k) + \vec{N}(k). \tag{15.48}$$

H is a transfer function describing the measurement process. $\vec{N}(k)$ is a zero-mean, white, Gaussian measurement noise vector with covariance R_C.

The Kalman filter algorithm has two parts: a set of initial values and a sequence of recursion relations.[6] The initial values are:

observation number:

$$k = 0, \tag{15.49}$$

estimated error (zero-mean Gaussian):

$$\begin{aligned} P(0\,|\,0) &= E\left\{\left[\vec{X}(0) - E\{\vec{X}(0)\}\right]\left[\vec{X}(0) - E\{\vec{X}(0)\}\right]^T\right\} \\ &= \mathrm{var}\{\vec{X}(0)\} \end{aligned} \tag{15.50}$$

where the superscript T denotes the transpose of a matrix, and estimated state vector at observation 0 given no measurements:

$$\hat{X}(0\,|\,0) = \mathrm{E}\left\{\vec{X}(0)\right\} = \left\{\mathrm{E}\left(X_1(0)\right), \mathrm{E}\left(X_2(0)\right), \cdots, \mathrm{E}\left(X_N(0)\right)\right\}. \qquad (15.51)$$

The covariances Q and R_C are constants.

The recursion relations in order of execution are:

estimated error given *k* observations:

$$P(k+1|k) = GP(k|k)G^T + Q \qquad (15.52)$$

Kalman filter gain:

$$K(k+1) = \frac{P(k+1|k)H^T}{HP(k+1|k)H^T + R_C} \qquad (15.53)$$

estimated error given *k+1* observations:

$$P(k+1|k+1) = \left[I - K(k+1)H\right]P(k+1|k) \qquad (15.54)$$

estimated state vector given *k* observations:

$$\hat{X}(k+1|k) = G\hat{X}(k\,|\,k) + \vec{F}(k+1|k) \qquad (15.55)$$

estimated state vector given *k+1* observations:

$$\hat{X}(k+1|k+1) = \hat{X}(k+1\,|\,k) + K(k+1)\left[\vec{Y}(k+1) - H\hat{X}(k+1|k)\right] \qquad (15.56)$$

incremented observation number:

$$k \to k+1 \qquad (15.57)$$

At the conclusion of Eq. (15.57), the recursion loops back to Eq. (15.52) for the next measurement.

Comparison of Fixed-Coefficient and Kalman Filters

It is instructive to compare and contrast fixed-coefficient and Kalman filters for tracking applications. Both types of filters are fading-memory, predictor-corrector, recursive filters. That is, the filter uses previous measurements but after a period of elapsed time, the influence of any individual measurement becomes less and less significant. The measurement fades from the filter memory. The filter predicts a position and corrects its prediction process based on observation. The algorithm is recursive in that it involves a fixed set of rules that are applied repeatedly once each measurement cycle.

An obvious difference between the two is that fixed-coefficient filters employ fixed gains while Kalman filters use time-variable gains that are adaptively varied based on past performance to maximize future tracking performance. Both techniques provide significant smoothing, but the smoothing from the Kalman filter should usually exceed that of the fixed-coefficient filters.

Fixed-coefficient filters have simple solutions for predicted performance that can assist design analyses. They are computationally simple and efficient. Thus they are ideal for use in dense multiple-target, high data rate applications. The performance of a Kalman filter cannot really be predicted, although it can be obtained through extensive Monte Carlo simulations. It is not uncommon for early performance estimates to be performed assuming a fixed-coefficient filter, with full knowledge that the final solution will be a Kalman filter with somewhat better performance. Kalman filters are considerable more computationally complex than fixed-coefficient filters. Significant processing capabilities may be required for multiple-target, high data rate applications.

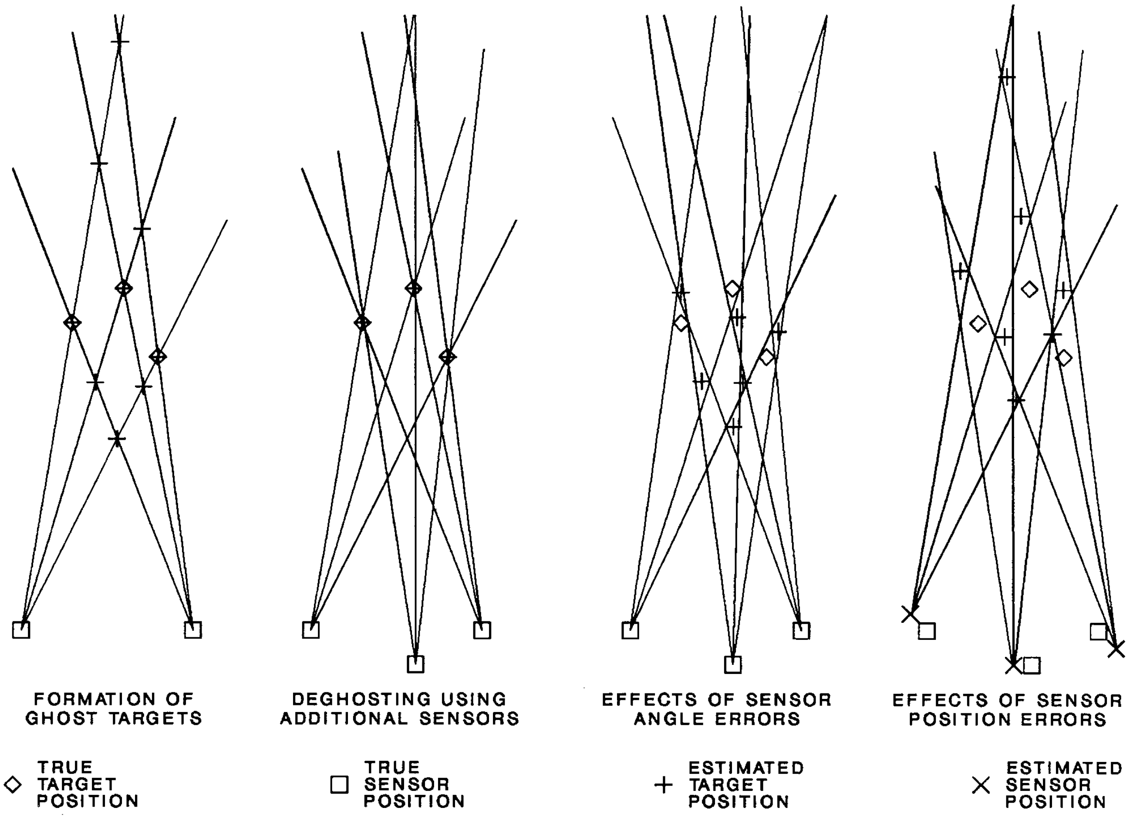
The track accuracy of fixed-coefficient filters degrades rapidly as "missed detections" become frequent. Kalman filters adjust their gains to accommodate missed detections with the least possible degradation in performance. Track accuracy of fixed-coefficient filters degrades if miscorrelation (frame-to-frame association of target detections to specific tracks) occurs. Frequent miscorrelation can cause track loss. The Kalman filter model does not include miscorrelation noise. Miscorrelation can lead to rapid divergence of track accuracy, further increased miscorrelation, and ultimate track loss. High missed detections favors Kalman filters. High miscorrelation favors fixed-coefficient filters.

The response time to adjust to target maneuvers is coefficient dependent, but predictable, in fixed-coefficient filters. Maneuvers will degrade track accuracy. However, if the maneuver does not carry the target out of the tracking window before the filter can adjust and if the target returns to the assumed class of dynamics after completion of the maneuver, then fixed-coefficient filters can follow the target through the maneuver. Target maneuvers are not included in the Kalman filter model. Violent maneuvers may produce miscorrelation and ultimate loss of track. If this does not occur, the Kalman filter will also follow the target through the maneuver. With either filter type it is good practice to run multiple filters with different dynamics models that can capture likely maneuvers. This is less of a penalty in fixed-coefficient filters than in Kalman filters.

Multi-Sensor, Multi-Platform Tracking

Many sensor systems are capable of making measurements of target angular position. This is useful in itself, but often more (the target range) is desired. A commonplace suggestion is to use two (or more) sensors and triangulate to find the target range. However, this is easier said than done. Consider Figure 15-9. In the first instance there are two sensors and three targets. Since only angle measurements are available, the three lines of sight from each of the two sensors give nine possible target positions (positions where two lines of sight cross) of which only three are real. The other positions are "ghost" targets. Not all combinations of positions are possible, but every ghost position is viable in one potential target set or another. The most common suggestion for "deghosting" is to add a third sensor (second instance in Figure 15-9). Now valid target positions require all three lines of site to occur at a point (so-called "triple crossings"). This would seem to be an acceptable technique, however, we have neglected the potential for sensor errors. The third instance in Figure 15-9 shows the effects of random errors in the measured target angles. The triple crossings disappear. One is tempted to assume that if three lines intersect forming a triangular region that is small enough, then this is the equivalent of a triple crossing point. However, some uncertainty would remain. In the figure it appears that triangular regions near the true target points exist, but there are several others just as small that would give gross errors. In the fourth instance,

Figure 15-9. Problems with using triangulation of angle-only measurements to estimate range.


FORMATION OF GHOST TARGETS
DEGHOSTING USING ADDITIONAL SENSORS
EFFECTS OF SENSOR ANGLE ERRORS
EFFECTS OF SENSOR POSITION ERRORS
◇ TRUE TARGET POSITION
□ TRUE SENSOR POSITION
+ ESTIMATED TARGET POSITION
× ESTIMATED SENSOR POSITION

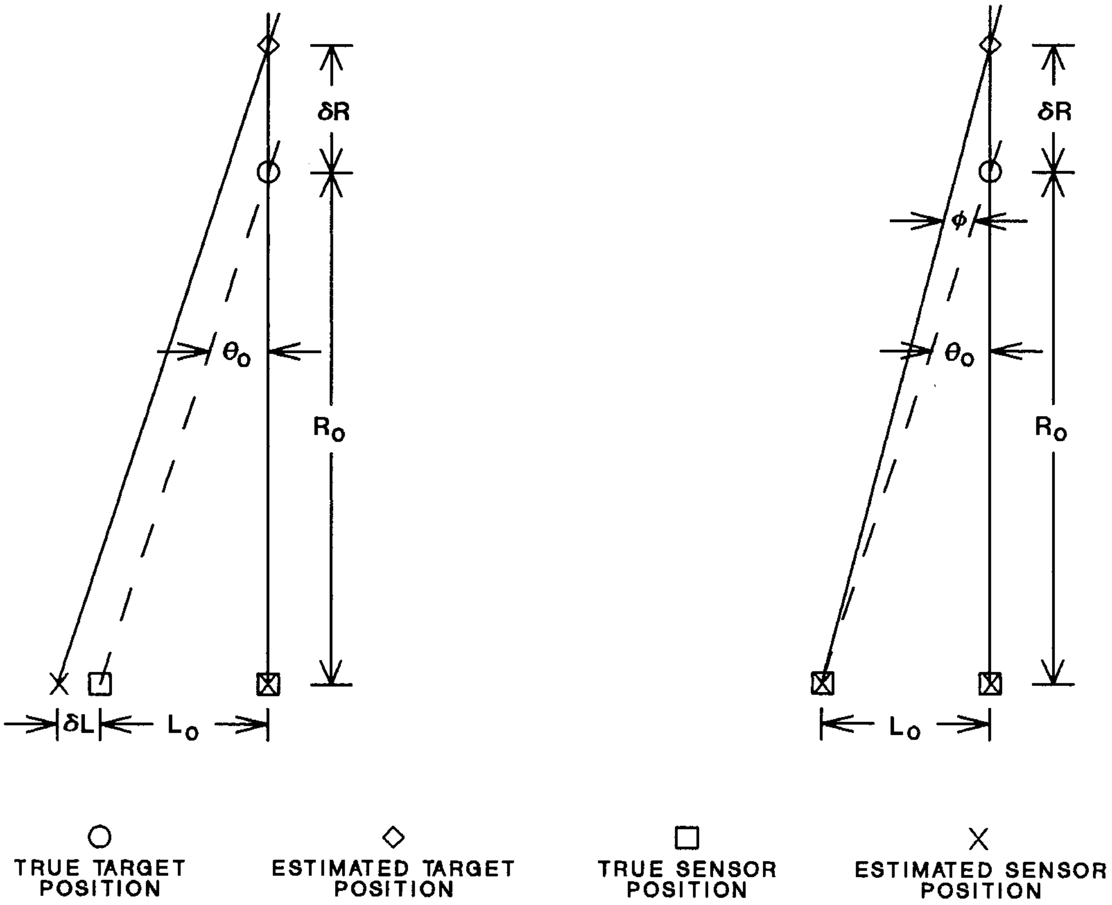
the sensors are assumed to have position measurement errors. When triangulation is attempted, the wrong baselines will be assumed. This also breaks the triple crossing degeneracies. If small triangular regions are assumed to indicate targets, then many ghosts appear. The combination of angular errors and position errors makes the problem even harder. Addition of still more sensors will ultimately solve the problem, but use of only two or three may not.

The magnitude of the effects of measurement errors can be determined from simple geometric analysis. Consider the first instance in Figure 15-10. This is a system in which both sensors make perfect angular measurements but the sensors make a small error δL in measuring their separation L_0. If R_0 is the true range to the target, the error in estimating range is easily determined to be

$$\delta R = \delta L\left(R_0 / L_0\right) = \delta L \cot\theta_0 \tag{15.58}$$

where θ_0 is the angular difference between the sensor angle measurements. If we pick reasonable

Figure 15-10. Geometry for evaluation of target position errors given errors in sensor position measurements and in sensor angle measurements.



values for the parameters, such as, L_0 =1 km, R_0 = 100 km, and δL = 10 m, then we find δR = 1 km. This is not too bad for many applications (target localization, targeting for carpet bombing, or cuing a weapon with a terminal seeker), but inadequate for others (indirect fire artillery).

Now consider the second instance in the figure. In this case, the baseline is assumed to be accurately known but the sensor angles are assumed to have a small relative angular error ϕ. From the given geometry it is easy to determine that

$$\begin{aligned}\delta R &= L_0 \cot(\theta_0 - \phi) - R_0 \\ &= L_0 \frac{1+\tan\theta_0 \tan\phi}{\tan\theta_0 - \tan\phi} - R_0 \quad . \qquad (15.59) \\ &= L_0 \frac{R_0 + L_0 \tan\phi}{L_0 - R_0 \tan\phi} - R_0\end{aligned}$$

Now assuming the same values as before, i.e., L_0 =1 km and R_0 = 100 km, and a modest value of the angular error, ϕ = 1 mrad, we calculate that δR = 11.1 km. This is too large for almost any purpose.

Some angle sensors (such as an infrared search and track system) can make angle measurements better than 1 mrad, although many others (radio frequency interferometers, radars in which the range function is being jammed) cannot. Let us consider a very realistic system. The Global Positioning System can provide position accuracy to roughly 10 m for military systems. Thus, any two-platform system should know its position well enough to get range to 1% accuracy, if angular errors were not present. Let us do an error budget for a system attempting to obtain 1 mrad overall angular accuracy (giving approximately 1% range error). First, each platform must determine the direction of true north relative to the platform's reference system (in order to communicate the angles to each other, a universal coordinate system must be used). There will be some error in this determination. Next, the sensor making the target angle measurements must be aligned with to respect to the platform reference system. There will be a significant error in this (most combat platforms are flexible and not perfectly elastic – they can warp significant amounts in the course of combat operations). Lastly, the sensor must estimate the position of the target with respect to its internal references. All together we have at least 6 angular error sources (3 from each platform). Assuming all are independent and that all are roughly equal (the optimistic assumption), then the allowable error per source is $1/\sqrt{6} = 0.41$ of the total allowable error. An 0.41 mrad error is difficult to achieve in any of the error sources let alone all of them. It must also be considered that even 1% range error is not adequate for missions such as artillery support. Targeting of sites to GPS accuracies to permit GPS guided weapons to be employed will require 0.1% errors at 10 km range decreasing to 0.001% at 1000 km range.

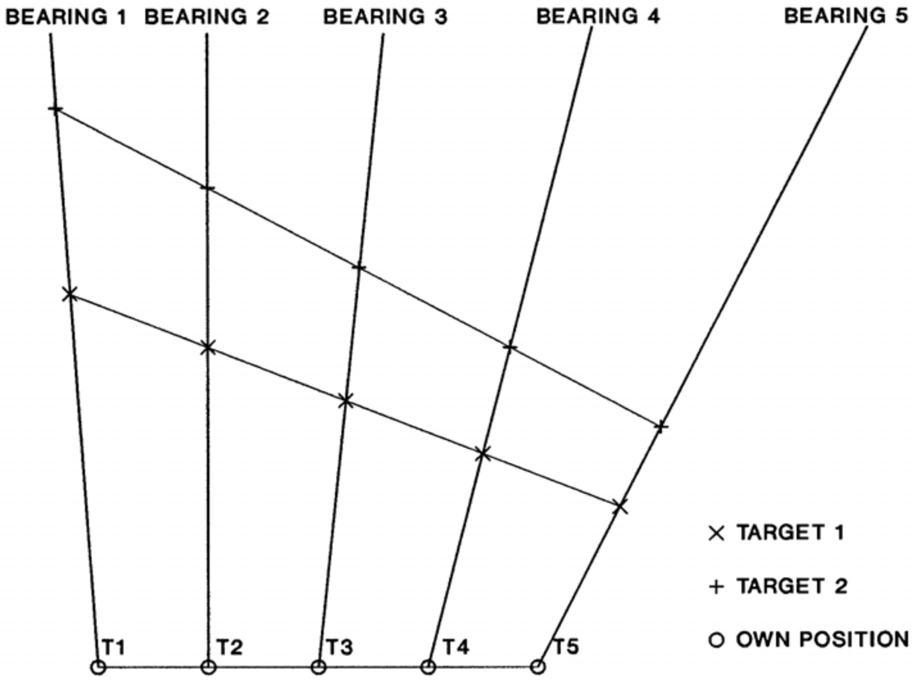
Angle-only tracking performance can be improved in several ways. The first is to use a larger baseline. As L_0 increases, the rms error δR decreases. The ideal geometry for triangulation is to have the baseline comparable to the range. In this ideal geometry the rms error is approximately the rms angular error times the range. If we use a 100 km baseline in our example, we might achieve something of the order of $\delta R/R = 0.001$ rather than the $\delta R/R = 0.11$ that was achieved. Unfortunately, there are many situations where the ideal geometry cannot be achieved. The situation can also be ameliorated by introducing multiple sensors and averaging over many measurements to reduce the contributions of errors. The rms range error will decrease as the square root of the number of independent measurements made (whether made from different locations by the same sensor or by other sensors). However, if sensor bias error is the dominant error source, then multiple measurements will not improve the accuracy, and multiple sensors will only improve the error by the square root of the number of sensors employed. As a consequence, it still often remains much easier to talk about using triangulation between multiple platforms than it is to actually implement it in a real system.

Target Motion Analysis

The preceding discussion of tracking and range estimation of targets from multiple platforms often elicits the question as to whether a target can be tracked using measurements made at different positions (and obviously different times). The answer to this question is a qualified yes and requires a discussion of the topic of **target motion analysis (TMA)**.[7], [8] TMA involves the study of angular bearings to a target over time taking account of the motion of the observing platform.

Consider an observing platform (denoted by O's) moving in a straight line at a constant velocity and a target (denoted by X's) also moving in a straight line at a constant velocity. Five bearing observations made at equal time intervals are shown in Figure 15-11. For the constant velocity target and platform, there is a smooth change in relative bearing over time. However, the angle only information is insufficient to determine the target range. A second target at a different range moving at a different constant velocity is shown that produces the exact same bearing measurements as the first target. In fact, there are an infinite number of possible constant-velocity paths that produce the same bearing measurements. There is one possible solution for each possible value of the initial target range. Distant targets must move faster and have a more acute closure angle than nearby targets, but they are all mathematically acceptable. If there is no other information available, such as an initial range, or a known velocity, there is no way to discriminate between solutions in the scenario presented in Figure 15-11.

Figure 15-11. Many possible target positions and velocities satisfy any set of bearings measured from a platform in straight line motion.



A solution to the problem can be obtained if one changes the scenario. In Figure 15-12, the target still continues to have uniform straight line motion, but the observing platform makes a maneuver midway through the observations. Before the maneuver, the bearing rates are consistent with many different target motions. After the maneuver, the new bearing rates are also consistent with many different target motions. However, the consistent solutions after the maneuver are not the same consistent solutions as before the maneuver. In fact, there is only one range and velocity that gives consistent bearing observations both before and after the platform maneuver. In Figure 15-12, postulated target motion 1 is consistent both before and after (target 1 is the actually target track). Postulated target motion 2 is consistent only before the maneuver. After the maneuver, that motion requires bearing observations that no longer agree with the true observations. Target track 2 can be eliminated.

A geometry useful in analyzing TMA is shown in Figure 15-13. The observation platform makes two bearing measurements α_1 and α_2 separated in distance by $V_O T$ where V_O is the known velocity of the platform and T is the known time between measurements. The target is initially located at an unknown range R_1 traveling at unknown speed V_T at an unknown angle γrelative to the bearing direction. Assume that the target range and speed are known. These are probably nothing but a guess, but we can assign values to the parameters, whether or not the parameters are correct. At speed V_T the target can move a distance $V_T T$ in the time between measurements.

Figure 15-12. Bearings measurements made along two linear segments resolve ambiguities.

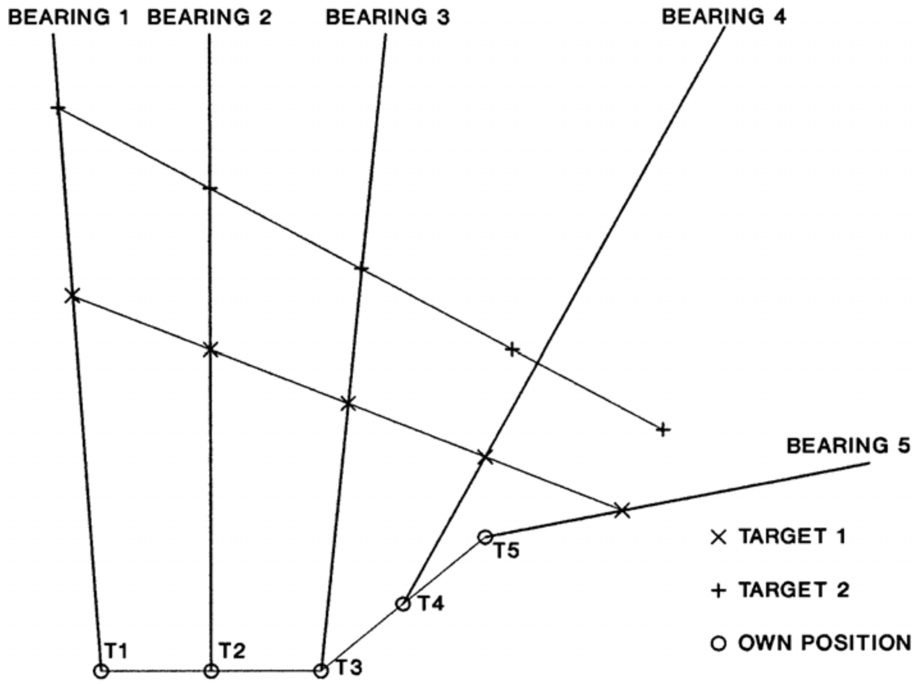
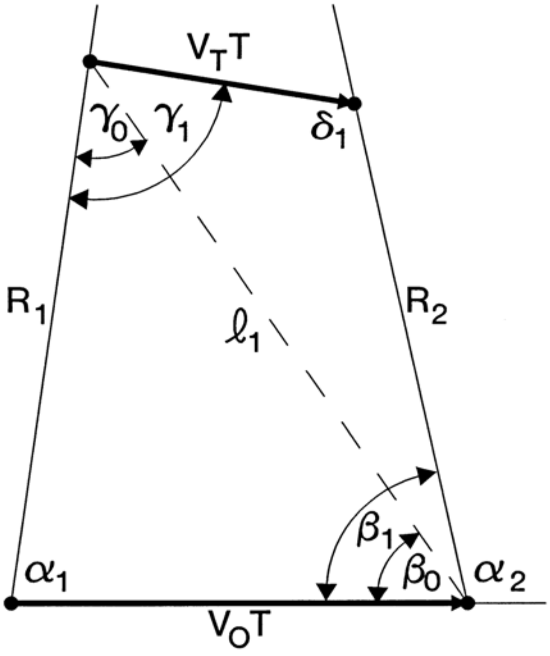


Figure 15-13. Geometry used in target motion analysis.



One side of the quadrilateral defined by the target and platform positions at the two measurement times is known precisely as are two of the four angles. Two more sides are assumed to be known. This information is sufficient to determine all of the other quadrilateral characteristics. Specifically, we are interest in determining what angles γ satisfy the observations.

It is convenient to determine the length ℓ of the diagonal connecting the initial target position to the final platform position. Using the law of cosines this is given by

$$\ell_1^2 = R_1^2 + (V_O T)^2 - 2R_1 V_O T \cos\alpha_1 \tag{15.60}$$

Similarly, the angle β_0 made between this diagonal and the platform motion side of the quadrilateral can also be found using the law of cosines

$$R_1^2 = \ell_1^2 + (V_O T)^2 - 2\ell_1 V_O T \cos\beta_0 \tag{15.61}$$

Solving for the angle yields

$$\beta_0 = \cos^{-1}\left(\frac{\ell_1^2 + (V_O T)^2 - R_1^2}{2\ell_1 V_O T}\right) \tag{15.62}$$

or

$$\beta_0 = \cos^{-1}\left(\frac{V_O T - R_1 \cos\alpha_1}{\left[R_1^2 + (V_O T)^2 - 2R_1 V_O T \cos\alpha_1 \right]^{1/2}} \right). \tag{15.63}$$

The third angle of this triangle γ_0 could be found in the same way, i.e.,

$$\gamma_0 = \cos^{-1}\left(\frac{R_1^2 + \ell_1^2 - (V_O T)^2}{2\ell_1 R_1} \right) \tag{15.64}$$

and

$$\gamma_0 = \cos^{-1}\left(\frac{R_1 - V_O T \cos\alpha_1}{\left[R_1^2 + (V_O T)^2 - 2R_1 V_O T \cos\alpha_1 \right]^{1/2}} \right), \tag{15.65}$$

or it could be determined from the fact that triangles have angles that must sum to π,

$$\gamma_0 = \pi - \alpha_1 - \beta_0 . \tag{15.66}$$

Given a solution for β_0, the other triangle defined by the diagonal can be addressed. From the law of sines we have

$$\frac{V_T T}{\sin(\beta_1 - \beta_0)} = \frac{\ell_1}{\sin\delta_1}, \tag{15.67}$$

which can be rearranged to yield

$$\sin\delta_1 = \frac{\ell_1 \sin(\beta_1 - \beta_0)}{V_T T} \tag{15.68}$$

and then solved to yield the third angle δ_1 of the quadrilateral

$$\delta_1 = \sin^{-1}\left(\frac{\ell_1 \sin(\beta_1 - \beta_0)}{V_T T} \right). \tag{15.69}$$

The desired angle γ_1 is found from the requirement that the four angles of a quadrilateral must sum to 2π. Thus,

$$\gamma_1 = 2\pi - \alpha_1 - \beta_1 - \delta_1 = \pi + \alpha_2 - \alpha_1 - \delta_1 \tag{15.70}$$

Substituting in the earlier expressions for β_0 and ℓ yields

$$\begin{aligned}\gamma_1 &= 2\pi - \alpha_1 - \beta_1 \\ &- \sin^{-1}\left(\left.\left(\begin{array}{l}\left[R_1^2 + \left(V_O T\right)^2 - 2R_1 V_O T\cos\alpha_1\right]^{1/2} \\ \cdot\sin\left(\beta_1 - \cos^{-1}\left(\dfrac{V_O T - R_1\cos\alpha_1}{\left[R_1^2 + \left(V_O T\right)^2 - 2R_1 V_O T\cos\alpha_1\right]^{1/2}}\right)\right)\end{array}\right)\right/ V_T T\right)\end{aligned} \tag{15.71}$$

Eq. (15.71) Contains only measured or assumed quantities. For an assumed range and speed, there are at most two mathematical solutions for γ_1. If we make a third bearing measurement, and repeat this analysis, one of the original two solutions will be found to inconsistent with the angles and ranges associated with the second computation.

With estimates for both γ_1 and γ_0 in hand, it is possible to predict the value of target range R_2 at the time of the second bearing measurement. Using the law of cosines again we have

$$R_2^2 = \ell_1^2 + \left(V_T T\right)^2 - 2\ell_1 V_T T\cos\left(\gamma_1 - \gamma_0\right). \tag{15.72}$$

Substitution of the expressions for γ_1, γ_0, and ℓ_1 into this expression (to produce an expression dependent only on measured or assumed values) is left as an exercise. The predicted value of R_2 can now be used as the starting point for analyzing the next bearing measurement. It should also be noted that γ_2 (the value of γ_1 that should result from analyzing the next pair of bearing measurements can also be estimated from the value of δ_1, i.e.,

$$\gamma_2 = \pi - \delta_1 . \tag{15.73}$$

The difference between the value of γ_2 predicted from the previous measurement analysis and the value calculated from the succeeding measurement analysis is a residual that can be used to determine how significantly measurement errors are affecting the "track".

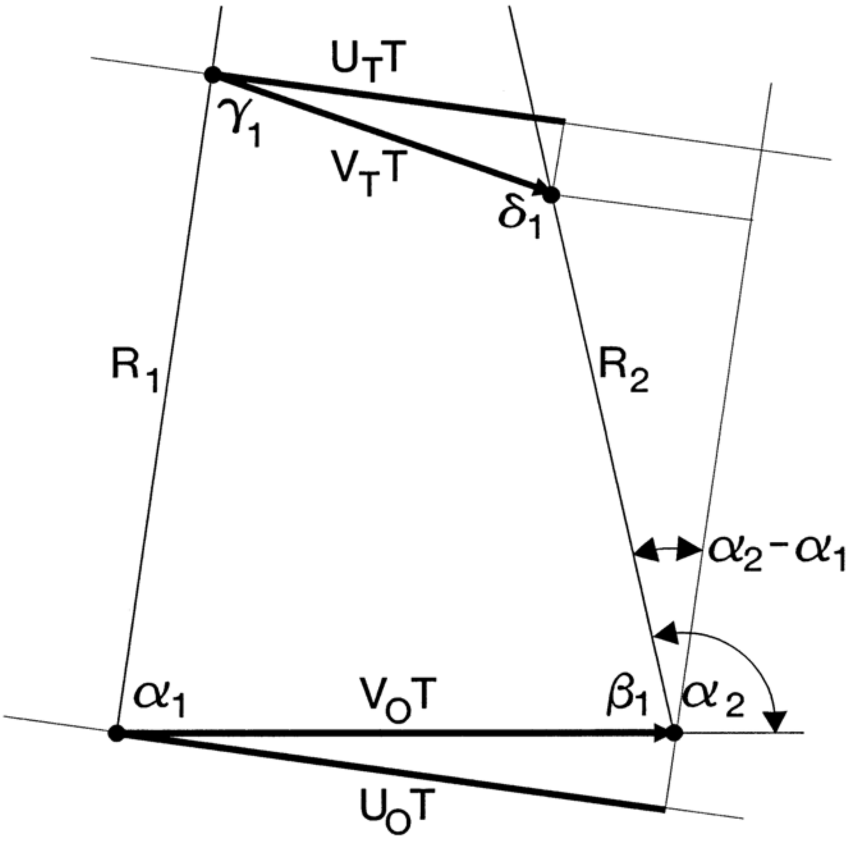
Three bearing measurements can be used to uniquely identify the target trajectory **IF** the range and speed are assumed to be known (and if the measurements are sufficiently accurate). If the bearing measurements have errors, then for each pair of measurements (and for each assumed set of range and speed values) the predicted motion angles (γ_1) will be different. Only by making a large number of measurements and smoothing with a filter will a "precise" angle be determined.

In practice, we do not know either the range or the speed. However, the speed is almost always limited to a moderately narrow range of values. Under normal conditions both ships and aircraft tend to move with speeds between the speed that gives them optimum fuel consumption (for maximum range) and their maximum sustained speed. Submarines hunting for targets tend to move at speed slightly below the maximum speed that gives them reasonably quiet signatures. These speed ranges are usually approximately known in advance. If the target type can be estimated, then the speed can be estimated (to perhaps a factor of two accuracy or possibly even better). Range might also be estimated. At initial detection of a target, it is reasonable to assume that the target is near the maximum detection range of the sensor. This range clearly depends on the target signal strength ("small" targets cannot be detected as far away as "large" targets). However, if the target type can be guessed, then the detection range can be estimated (often to better than factor of two accuracy). Other information may be available that can aid in refining these estimates. Sonar or passive rf signatures may identify the target type. They may even provide information on possible target motion. For, example, Doppler shift in an IFF (identification friend or foe) transponder return (some classes of IFF emitters tend to have relatively precise emission frequencies) or blade rate data from a sonar might give an improved estimate of velocity.

There is no single well-defined algorithm for performing a target motion analysis. This is an area of active research interest. We will assume that the bearings to the target have been measured over two extended straight line segments that are not collinear. One algorithmic approach (not necessarily the best nor the one used in most systems) to determining the real target position is to assume a target speed, estimate the motion angles (using the first set of straight line measurements) for several possible initial ranges. The process is repeated for the second set of straight line measurements. One range will give less angular error between the motion in the first set and the motion in the second set than will the other ranges. If these errors are monotonically decreasing as one end of the spread of values is approached, then additional values should be included until a minimum is obtained. Next this minimum-angular-error range is assumed and the motions for each segment are estimated for several velocities around the velocity originally selected. One velocity will yield less angular error than the others. Using this new velocity, the process is repeated assuming several range values (usually closer together than the original set of range values). Iterating back and forth between fixing the range and velocity should ultimately result in convergence to a single range and velocity that yields minimum error in the motion angles before and after the break in straight line measurements. NOTE: in the presence of unavoidable measurement errors, convergence means that the error will reduce to a small value after which it will not continue to get smaller. The iterations should be halted as soon as the error stops shrinking. It is not possible to obtain a zero-error solution when the initial measurements contain errors.

There are several techniques for estimating the target range from a very limited number of measurements. The estimates are not as accurate as the lengthy recursive analysis of an extended series of measurements, but they can be performed quickly and simply. One of these techniques is called **Ekelund ranging**. Consider the geometry shown in Figure 15-14. If we define the quantities U_O and U_T as the component of velocity (of the platform and target, respectively) perpendicular to the measured bearing, then from two bearing measurements made during straight line motion we find

Figure 15-14. The basics of geometry for analyzing Ekelund ranging.



$$R_2 \sin(\alpha_2 - \alpha_1) = U_O T - U_T T \tag{15.74}$$

where *T* is the time between the two bearing measurements α_1 and α_2 and the transverse velocity components are related to the total velocity components by

$$U_O = V_O \sin \alpha_1 \tag{15.75}$$

and

$$U_T = V_T \sin \gamma_1 .$$

Dividing both sides of Eq. (15.74) by the time difference *T* yields

$$\frac{R_2 \sin(\alpha_2 - \alpha_1)}{T} = U_O - U_T . \tag{15.76}$$

Now let the time difference become small, i.e., $T \rightarrow 0$. The difference between bearings will become small, so that *sin x ~ x*. The small bearing difference divided by a small time difference approximates the derivative of the bearing with respect to time. The range will not change appreciably over the measurement interval so we have $R_2 \sim R_1 \sim R$. Thus, Eq. (15.76) can be rewritten as

$$\lim_{T\rightarrow 0} \frac{R_2 \sin(\alpha_2 - \alpha_1)}{T} \approx \frac{R(\alpha_2 - \alpha_1)}{T} \approx \frac{R\Delta\alpha}{T} \approx R\frac{d\alpha}{dt} = U_O - U_T \quad (15.77)$$

The range times the bearing rate is equal to the difference in transverse velocity components.

Now consider bearing rate measurements immediately before and after a change in platform direction. Immediately before the direction change we have

$$R\frac{d\alpha}{dt} = U_O - U_T \, . \quad (15.78)$$

Immediately after the direction change, the range remains the same, the target transverse velocity remains the same, but the bearing rate and platform transverse velocity will be different. We have

$$R\frac{d\alpha'}{dt} = U_O' - U_T \ . \quad (15.79)$$

Eqs. (15.78) and (15.79) can be solved to yield the range *R*

$$R\frac{d\alpha}{dt} - R\frac{d\alpha'}{dt} = (U_O - U_T) - (U_O' - U_T) \quad (15.80)$$

or

$$R\left(\frac{d\alpha}{dt} - \frac{d\alpha'}{dt}\right) = U_O - U_O' \quad (15.81)$$

and finally

$$R = \frac{U_O - U_O'}{\frac{d\alpha}{dt} - \frac{d\alpha'}{dt}} \, . \quad (15.82)$$

Equation (15.82) is the Ekelund estimate of target range. It should be noted that this is estimate can have substantial errors because it has the difference of two small inaccurate quantities in the denominator. A bearing can be estimated with a certain limited precision. A bearing rate adds the uncertainties in the bearing measurements. A difference in bearing rate adds the uncertainties in two bearing rate measurements. This cascading of uncertainties leads to substantial error potential in the estimate of range. If the time is allowed to grow large enough to reduce errors in bearing rate, then errors associated with range changes and the approximation of the sine function will become significant. Range errors in excess of 10% are commonplace and may be much larger. Nevertheless, the Ekelund range is better than no range estimate at all and may often be sufficient.

Even without being able to perform the maneuver (or before initiating that maneuver) it is possible to obtain a limited amount of information from TMA. Figure 15-15 shows the true motion and the observed (relative) motion for four possible situations. In Figure 15-15a, the target is overtaking the platform. The bearing rate is positive and monotonically increasing as long as the range is decreasing. Once minimum range has been passed, the bearing rate will continue to be positive but it will begin to monotonically decrease. In Figure 15-15b, the target is being overtaken by the platform. The bearing rate is negative and monotonically increases until minimum separation occurs, after which it remains negative but monotonically decreases. Figure 15-15c show a target that is passing the platform. The bearing rate is negative and behaves exactly like the overtaken target case. However, the bearing rates tend to be noticeably higher than in the overtaken target. Nevertheless, it may be possible to confuse the two situations.

One case, however, is unequivocal. Figure 15-15d shows a target on a collision course with the platform. The bearing rate is zero. Zero bearing rate is a red flag for any tracking situation. The only straight-line, constant-velocity motions that produce zero bearing rate are collision courses or courses that were collision courses in the past. The latter results if the target originated from the platform or had a near miss with the platform. For example, a missile launched from the platform or a target that has overflown the platform will be outbound and exhibit zero bearing rate. It is rare that information allowing discrimination of inbound collision courses from outbound near-misses is completely absent.

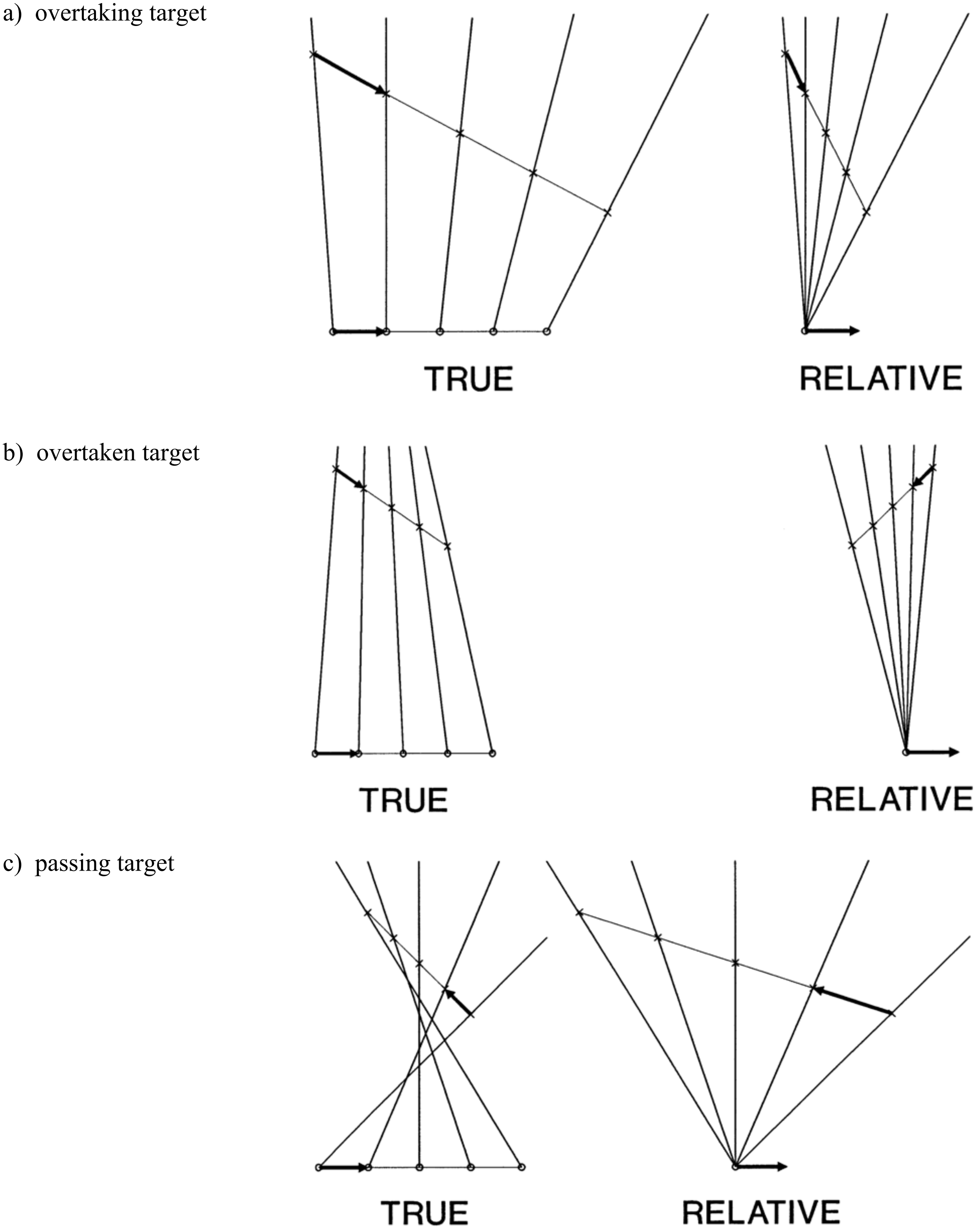
A skilled observer can use the bearing rate data to estimate the target motion. However, it is possible for even an unskilled observer to calculate the direction of relative motion. This quantity ζ is the target motion direction relative to the bearing of the platform motion. The geometry used to analyze this problem is shown in Figure 15-16. Applying the law of sines to the triangle forme from the first and second bearings yields

$$\frac{\sin(\pi-\alpha_1-\zeta)}{h}=\frac{\sin(\alpha_1+\zeta)}{h}=\frac{\sin(\alpha_1-\alpha_2)}{y_{12}}=\frac{\sin(\alpha_1-\alpha_2)}{V(t_2-t_1)}. \quad (15.83)$$

Applying the law of sines to the triangle formed from the second and third bearings yields

$$\frac{\sin(\pi-\alpha_3-\zeta)}{h}=\frac{\sin(\alpha_3+\zeta)}{h}=\frac{\sin(\alpha_2-\alpha_3)}{V(t_3-t_2)}. \quad (15.84)$$

Figure 15-15. Relative motions observed in different target motion situations.



a) overtaking target

b) overtaken target

c) passing target

Figure 15-15 (continued). Relative motions observed in different target motion situations.

d) collision course

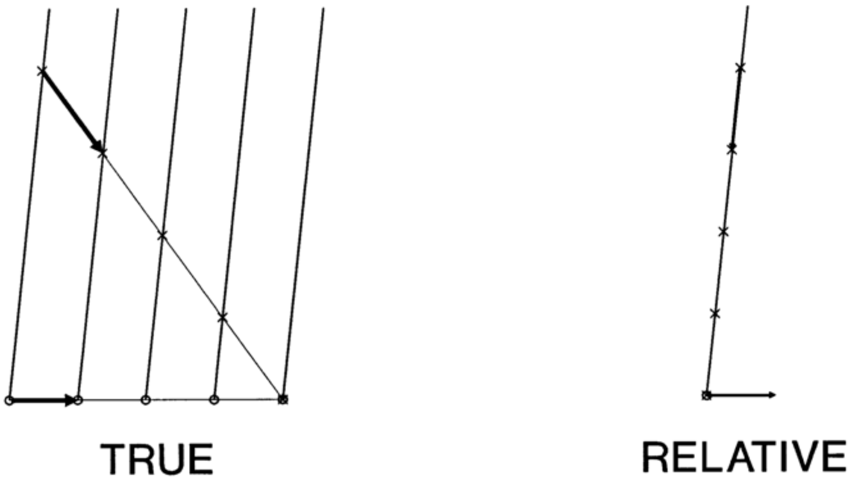
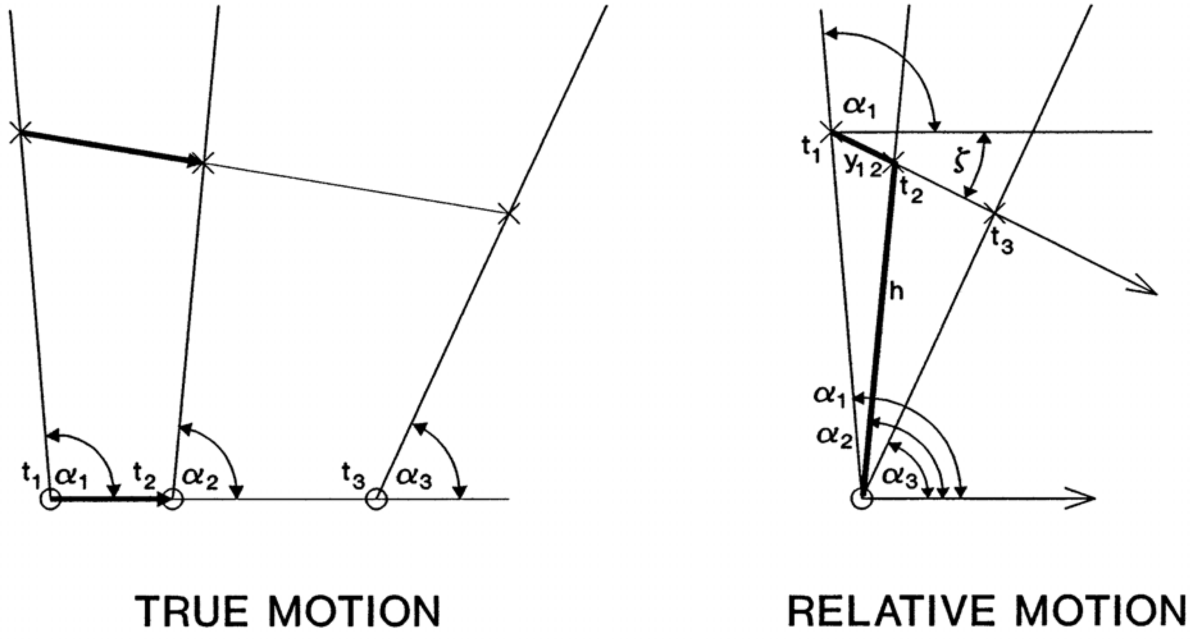


Figure 15-16. Geometry used to determine the direction of relative motion.



Solving both Equations (15.83) and (15.84) for *h/V* and equating the results yields

$$\frac{h}{V}=\frac{(t_2-t_1)\sin(\alpha_1+\zeta)}{\sin(\alpha_1-\alpha_2)}=\frac{(t_3-t_2)\sin(\alpha_3+\zeta)}{\sin(\alpha_2-\alpha_3)} \tag{15.85}$$

A little mathematical legerdemain allows us to obtain a more useful expression. That is, adding and subtracting α_1 from the arguments of both numerator and denominator on the right-hand side and regrouping the terms and then expanding the two *sin[(a)-(b)]* terms with four angles in them yields

$$\frac{(t_2-t_1)\sin(\alpha_1+\zeta)}{\sin(\alpha_1-\alpha_2)}=\frac{(t_3-t_2)\sin((\alpha_1+\zeta)-(\alpha_1-\alpha_3))}{\sin((\alpha_1-\alpha_3)-(\alpha_1-\alpha_2))}$$

$$=\frac{(t_3-t_2)(\sin(\alpha_1+\zeta)\cos(\alpha_1-\alpha_3)-\cos(\alpha_1+\zeta)\sin(\alpha_1-\alpha_3))}{\sin(\alpha_1-\alpha_3)\cos(\alpha_1-\alpha_2)-\cos(\alpha_1-\alpha_3)\sin(\alpha_1-\alpha_2)} \tag{15.86}$$

With a little more algebra we obtain

$$(t_2-t_1)\left(\frac{\cos(\alpha_1-\alpha_2)}{\sin(\alpha_1-\alpha_2)}\right)$$

$$=\frac{(t_3-t_2)\left(\left(\frac{\cos(\alpha_1-\alpha_3)}{\sin(\alpha_1-\alpha_3)}\right)-\left(\frac{\cos(\alpha_1+\zeta)}{\sin(\alpha_1+\zeta)}\right)\right)}{1-\left(\frac{\cos(\alpha_1-\alpha_3)}{\sin(\alpha_1-\alpha_3)}\right)\left(\frac{\sin(\alpha_1-\alpha_2)}{\cos(\alpha_1-\alpha_2)}\right)}, \tag{15.87}$$

which can be rewritten in terms of cotangents as

$$(t_2-t_1)(\cot(\alpha_1-\alpha_2))$$

$$=\frac{(t_3-t_2)(\cot(\alpha_1-\alpha_3)-\cot(\alpha_1+\zeta))}{1-(\cot(\alpha_1-\alpha_3))\left(\frac{1}{\cot(\alpha_1-\alpha_2)}\right)}. \tag{15.88}$$

Rearranging terms

$$
\begin{aligned}
&(t_2 - t_1)\left(\cot(\alpha_1 - \alpha_2) - \cot(\alpha_1 - \alpha_3)\right) \\
&\qquad = (t_3 - t_2)\left(\cot(\alpha_1 - \alpha_3) - \cot(\alpha_1 + \zeta)\right)
\end{aligned}
\tag{15.89}
$$

and simplifying yields

$$
\begin{aligned}
&(t_2 - t_1)\cot(\alpha_1 - \alpha_2) \\
&\qquad = (t_3 - t_1)\cot(\alpha_1 - \alpha_3) - (t_3 - t_2)\cot(\alpha_1 + \zeta) \cdot
\end{aligned}
\tag{15.90}
$$

This may finally be solved to yield the direction of relative motion.

$$
\zeta = -\alpha_1 + \cot^{-1}\left[\frac{(t_3 - t_1)\cot(\alpha_1 - \alpha_3) - (t_2 - t_1)\cot(\alpha_1 - \alpha_2)}{(t_3 - t_2)}\right] \tag{15.91}
$$

There is a final special case of TMA that can be useful. If the target is so distant that it cannot move a significant fraction of the range between the target and the observing platform, then the target may be assumed to be stationary. This is not an uncommon situation. For example, a jet aircraft at a range of 100 nmi can only move 5 nmi in any direction in the time it takes the observing platform to establish a 5 nmi triangulation baseline. A submarine detected in a convergence zone can only move 1 or 2 nmi in the time it takes for a destroyer or a helicopter to establish a 1-2 nmi baseline. The motion will result in errors but these errors are typically not significantly larger than the errors produced by triangulation using measurements with typical measurement errors. Figure 15-17 shows the geometry for analyzing **passive ranging** as single-platform triangulation of distant targets is called.

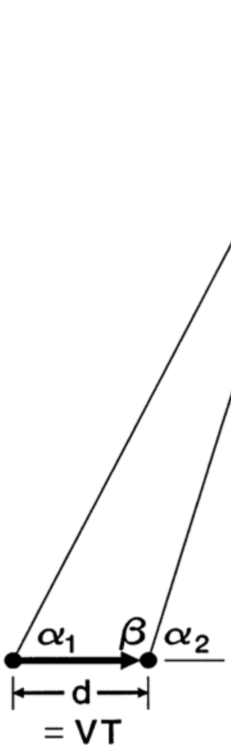
Applying the law of sines to the triangle shown in Figure 15-17 yields

$$
\frac{R}{\sin\alpha_1} = \frac{d}{\sin\gamma} = \frac{d}{\sin(\pi - \alpha_1 - \beta)} = \frac{d}{\sin(\alpha_1 + \beta)} . \tag{15.92}
$$

Rearranging terms in this equation yields the result

$$
R = \frac{d\sin\alpha_1}{\sin(\alpha_1 + \beta)} = \frac{d\sin\alpha_1}{\sin(\alpha_2 - \alpha_1)} = \frac{VT\sin\alpha_1}{\sin(\alpha_2 - \alpha_1)} \tag{15.93}
$$

Figure 15–17. Geometry for passive ranging of distant targets using triangulation.



where *d* is the baseline distance, *V* is the platform velocity (not the target velocity), *T* is the time used to synthesize the triangulation baseline, and α_1 and α_2 are the bearing angles relative to the platform motion. The effects of measurement errors can be estimated using the same relations developed for multi-platform tracking (Eqs. (15.58) and (15.59)). In the earlier section, these errors were estimated to be in the range of 1-10% of the target range. This is comparable to the error that would result if the target could move 1-10% of the target range.

References

[1] Skolnik, Merrill I., Introduction to Radar Systems 2nd Ed. (McGraw-Hill Book Co., New York NY, 1968) Chap.5.

[2] Tranchita, Charles J., Jakstas, Kazimieras, Palazzo, Robert G., and O'Connell, Joseph C., "Active Infrared Countermeasures" in David H. Pollock, (Ed.), Countermeasure Systems Vol. 7 of Infrared & Electro-Optical Systems Handbook (SPIE Press, Bellingham WA, 1993).

[3] Gerson, Gordon and Rue, Arthur K., "Tracking Systems", Chapter 22 in William L. Wolfe and George J. Zissis, (Eds.), The Infrared Handbook (Environmental Research Institute of Michigan, Ann Arbor MI, 1978).

[4] Legault, Richard, "Reticle and Image Analyses", Chapter 17 in William L. Wolfe and George J. Zissis, (Eds.), The Infrared Handbook (Environmental Research Institute of Michigan, Ann Arbor MI, 1978).

[5] Biberman, Lucien M., Reticles in Electro-Optical Devices (Pergamon Press, Oxford UK, 1966).

[6] Blackman, Samuel S., Multiple-Target Tracking with Radar Applications (Artech House, Norwood MA, 1986) pp. 19-47.

[7] VADM 'Doc', "TMA Basics" (undated). Available on the Internet at **http://members.nbci.com/_XMCM/688/dc_tma.html**.

[8] Wagner, Daniel H., Mylander, W. Charles, and Sanders, Thomas J., (Eds.), Naval Operations Analysis 3rd Ed. (Naval Institute Press, Annapolis MD, 1999).

Problems

15-1. Describe how monopulse tracking operates? In what types of target scenarios is monopulse tracking going to work best?

15-2. Why are conical scan, reticle, sequential lobing, etc. poor extended target trackers?

15-3. Describe the basic functioning and utility of a tracking filter.

15-4. In any tracking filters, what penalties do you pay for increased track smoothing?

15-5. You have the problem of tracking flies (one at a time -- a fly is nominally 1 cm in diameter and moves at speeds up to 5 m/s) at ranges up to 100 m. What techniques might be employed if 0.00001 rad tracking precision is required. If a fly can complete a maneuver in 1 msec and the tracking sensor is capable of only 0.0001 rad measurement precision, what filter gains and measurement intervals should be used to avoid serious track degradation.

15-6. What are the differences between the two types of fixed coefficient filters?

15-7. Compare the signals that a conical scan radar and a spin scan AM reticle seeker would see for targets at the center of the scan and for targets near the edge of the scan. A tracking system can employ either α-β or α-β-γ filters. For α=0.2 determine the optimum β and γ coefficients. Determine the response time and the hangoff errors for each. What condition would cause one or the other of these filters to be much more desirable than the other?

15-8. What simple output signal characteristic is analyzed to determine the tracking error signals in all of the following: conical scan, sequential lobing, and some spin-scan reticle tracking approaches?

15-9. What broad class of single target trackers works well against extended (resolved) targets?

15-10. Describe the track-while-scan tracking process.

15-11. You wish to track a target of 50 m length that has its predominant signature in the infrared. If the target range is roughly 1 m, what tracking options would you consider? If the target range is roughly 50 km, what tracking options would you consider? What tracking options would you consider in each case if several targets needed to be tracked simultaneously?

15-12. You wish to track by using triangulation. Your platforms have differential GPS, so the baseline accuracy can be as small as 0.1 m. If the baseline is 1 km and the target range is roughly 100 km, what important error sources remain? Estimate the magnitude of those errors and the magnitude of the resulting range estimation error.

15-13. Derive an expression for R_2 in Figure 15-13 in terms of only the assumed R_1 and $V_T T$ and the measured α_1, β_1, and $V_O T$.

15-14. A ship is making 10 knots. It begins a set of bearing measurements on an unknown target at time 0. After 5 minutes it makes a 90 degree turn to starboard and continues the measurements. The relative bearing measurements (direction of bow equals 0 degrees) are

TIME (minutes)	BEARING (degrees)
0	60.0
4	58.9
6	355.3
10	340.3

Estimate the range to the target at the turn.

15-15. A target is assumed to be distant relative to a platform moving with a velocity of 300 m/s. The platform observes the target at a relative bearing of 30° and then observes the target 10 seconds later at a bearing of 31°. Roughly how far away is the target?

APPENDIX A

UNITS, PHYSICAL CONSTANTS, AND USEFUL CONVERSION FACTORS

Units

This book conforms to the use of the International System of Units (SI) wherever practical [1]. This metric system is based on seven base units and two supplementary angular units as shown in Table A-1.

Table A-1. SI base and supplementary units.

QUANTITY	NAME	SYMBOL
length	meter	m
mass	kilogram	kg
time	second	s
electric current	ampere	A
thermodynamic temperature	kelvin	K
amount of substance	mole	mol
luminous intensity	candela	cd
plane angle	radian	rad
solid angle	steradian	sr

Many additional units are commonly accepted for use with SI. These are shown in Table A-2.

Table A-2. Units in use with SI.

QUANTITY	NAME	SYMBOL	DEFINITION
time	minute	min	1 min = 60 s
	hour	h	1 h = 3600 s
	day	d	1 d = 86400 s
plane angle	degree	°	1° = (π/180) rad
	minute	′	1′ = (1/60)°
	second	″	1″ = (1/60)′
volume	liter	L	1 L = 0.001 m^3
mass	metric ton, tonne	t	1 t = 1000 kg

From the base units in Tables A-1 and A-2 are derived all other units of the SI system. Some of the derived units have specially-named non-SI counterparts or are used so often in certain applications that they have been given special names and symbols. These derived units are listed in Table A-3.

Table A-3. SI derived units.

QUANTITY	NAME	SYMBOL	EQUIVALENT
frequency	hertz	Hz	s^{-1}
force	newton	N	$kg\text{-}m/s^2$
pressure, stress	pascal	Pa	N/m^2
work, energy, heat	joule	J	N-m, $kg\text{-}m^2/s^2$
power	watt	W	J/s
electric charge	coulomb	C	A-s
electric potential, emf	volt	V	J/C, W/A
resistance	ohm	Ω	V/A
conductance	siemens	S	A/V, Ω^{-1}
inductance	henry	H	Wb/A
capacitance	farad	F	C/V
magnetic flux	weber	Wb	V-s, N-m/A
magnetic flux density	tesla	T	Wb/m^2, N/A-m
Celsius temperature	degree Celsius	°C	K - 273.15
luminous flux	lumen	lm	cd-sr
illuminance	lux	lx	lm/m^2
radioactivity	becquerel	Bq	s^{-1}
absorbed dose	gray	Gy	$m^2\text{-}s^{-2}$, J/kg
dose equivalent	sievert	Sv	J/kg

Several non-SI units are strongly entrenched in their respective applications and have no roughly comparable SI counterparts. As a consequence their use is provisionally accepted. A few of these units cannot be defined precisely in terms of SI units; their values must be experimentally determined. The special non-SI units that are provisionally acceptable are listed in Table A-4.

Table A-4. Non-SI units commonly accepted for use.

QUANTITY	NAME	SYMBOL	EQUIVALENT
cross section	barn	b	10^{-28} m^2
pressure	bar	bar	10^5 Pa
radioactivity	Curie	Ci	3.7 x 10^{10} Bq
exposure to xrays	roentgen	R	2.58 x 10^{-4} C/kg
absorbed radiation	rad	rad	0.01 Gy
length	Cngstrom	C	10^{-10} m
energy	electron volt	eV	1.602176462 x 10^{-19} J
mass	atomic mass unit	u	1.66053873 x 10^{-27} kg

The prefixes shown in Table A-5 are used in SI to create units larger or smaller than the basic ones.

Table A-5. SI prefixes.

FACTOR	PREFIX	SYMBOL	FACTOR	PREFIX	SYMBOL
10^{24}	yotta	Y	10^{-1}	deci	d
10^{21}	zetta	Z	10^{-2}	centi	c
10^{18}	exa	E	10^{-3}	milli	m
10^{15}	peta	P	10^{-6}	micro	μ (often u)
10^{12}	tera	T	10^{-9}	nano	n
10^9	giga	G	10^{-12}	pico	p
10^6	mega	M	10^{-15}	femto	f
10^3	kilo	k	10^{-18}	atto	a
10^2	hecto	h	10^{-21}	zepto	z
10^1	deka	da	10^{-24}	yocto	y

Thus, for example, 1 kilometer equals 10^3 meters, while 1 picosecond equals 10^{-12} seconds.

Non-SI Systems of Units

The only claim that the SI system of units has to be the universal system of units is that the international scientific community has accepted them as standard. Many other systems of units could just as easily have been selected. The "English system of units" based on seconds, inches/feet/yards/miles, and ounces/pounds/tons is no less fundamental nor less useful than the metric system. The metric system claim to fame is based almost entirely on the supposed (and in fact almost entirely cultural) superiority of the decimal system. Other cultures have used number bases other than 10. The advent of computers has demonstrated that binary units might be even more fundamental than decimal units.

Because of this many other systems of units have been independently derived by the founders of a field of endeavor and have continued to be widely used until recently. All of the original derivations in classical physics and many if not most of the applications occurred before the adoption of the SI system of units. Thus, anyone attempting to find information in a field is as likely to encounter the use of one of the non-standard systems of units as they are to encounter the use of SI units. The first few sections of this appendix are intended as **Rosetta Stones** with respect to systems of units.

Many of the systems of units arose from the study of different aspects of electricity and magnetism.[2] At least five major unit systems arose here: Electrostatic units (esu – a form of cgs units – centimeter/gram/second units), Electromagnetic units (emu – another form of cgs units), Gaussian units (the dominant form of cgs units, Heaviside-Lorentz units, and Rationalized MKSA (meter/kilogram/second/ampere – identical with SI units as far as electrical and magnetic units are concerned and is essentially the direct predecessor of SI units). The various unit systems are defined in terms of how they relate to fundamental electrical and magnetic quantities and to Maxwell's equations. For example, different units may have a different proportionality between the electric field (*E*) and the charge (*q*) at distance (*r*)

$$E = k_1 \frac{q}{r} \tag{A.1}$$

They may have different proportionality between the Ampere force per unit length (*dF/dl*) and the currents (*I* and *I'*) in two long wires separated by distance (*d*)

$$\frac{dF}{dl} = 2k_2 \frac{I\,I'}{d} \tag{A.2}$$

They may have different proportionality between electromotive force and time rate of change of magnetic induction as expressed in the Maxwell equation

$$\nabla \times \vec{E} = -k_3 \frac{\partial \vec{B}}{\partial t} \tag{A.3}$$

They may have different proportionality between magnetic induction (B) and a current (I) at distance (r)

$$B = 2k_2k_4\frac{I}{r} \tag{A.4}$$

Using these proportionalities, Maxwell's equations take the form

$$\nabla \cdot \vec{E} = 4\pi\, k_1\, \rho \tag{A.5}$$

$$\nabla \times \vec{E} = -\, k_3\, \frac{\partial \vec{B}}{\partial t} \tag{A.6}$$

$$\nabla \cdot \vec{B} = 0 \tag{A.7}$$

$$\nabla \times \vec{B} = 4\pi\, k_2 \alpha\, \vec{J} + \frac{k_2 \alpha}{k_1}\frac{\partial \vec{E}}{\partial t} \tag{A.8}$$

Table A-6 lists the values of each of the four proportionality constants for the five major electromagnetic unit systems. Table A-7 lists the units and the associated conversion factors to the equivalent SI units for the electrostatic, electromagnetic, Gaussian, and SI units.

Table A-6. Proportionality constants defining the different electromagnetic unit systems.

UNIT SYSTEM	k_1	k_2	k_3	k_4	k_5	ε_0	μ_0
Electrostatic (esu)	1	$1/c^2$	1	1	4π	1	$1/c^2$
Electromagnetic (emu)	c^2	1	1	1	4π	$1/c^2$	1
Gaussian	1	$1/c^2$	$1/c$	c	4π	1	1
Heaviside-Lorentz	$1/4\pi$	$1/4\pi c^2$	$1/c$	c	1	1	1
Rationalized MKSA (SI)	$1/4\pi\varepsilon_0$	$\mu_0/4p$	1	1	1	ε_0	μ_0

The four proportionality constants are not completely independent, it is possible to show that

$$k_1 / k_2 = c^2 \tag{A.9}$$

$$k_1 / k_2 k_3 k_4 = c^2 \tag{A.10}$$

and

$$k_3 = 1 / k_4 \tag{A.11}$$

Additional useful equations include the Lorentz force relation

$$\vec{F} / q = \vec{E} + k_3 \vec{v} \times \vec{B} \tag{A.12}$$

constitutive relations

$$\vec{D} = \varepsilon_0 \vec{E} + k_5 \vec{P} \tag{A.13}$$

and

$$\vec{H} = \frac{1}{\mu_0} \vec{B} - k_5 \vec{M} \tag{A.14}$$

Using the constants ε_0, μ_0, and k_5 from Table A-6, the form of the constitutive relations in each system of units can be determined.

There remain many important equations in the literature that are not easily transformed using the results above. Table A-8 lists conversion factors for electromagnetic variables in Gaussian and SI units, the two most common unit systems. To convert equations from Gaussian to SI units, the quantities in the center column are replaced by the modified quantities in the right-hand column, everywhere that they arise.

Table A-7. Electromagnetic Systems of Units. In equations below, c = 2.997925 x 10^8 m/s.

		CGS UNIT SYSTEMS		MKS UNITS
QUANTITY	Electrostatic	Electromagnetic	Gaussian	SI (MKSA)
Length	Centimeter	Centimeter	Centimeter	Meter (m)
Mass	Gram	Gram	Gram	Kilogram (kg)
Time	Second	Second	Second	Second (s)
Force	Dyne (10^{-5} N)	Dyne (10^{-5} N)	Dyne (10^{-5} N)	Newton (N)
Energy	Erg (10^{-7} J)	Erg (10^{-7} J)	Erg (10^{-7} J)	Joule (J)
Charge	Statcoulomb (0.1/c C)	Abcoulomb (10 A)	Statcoulomb (0.1/c C)	Coulomb (C)
Current	Statampere (0.1/c A)	Abampere (10 A)	Statampere (0.1/c A)	Ampere (A)
Electric Field Field	Statvolt/cm (c/10^4 V/m)	Abvolt/cm (10^{-6} V/m)	Statvolt/cm (c/10^4 V/m)	Volt/meter
Electric Potential	Statvolt (c/10^6 V)	Abvolt (10^{-8} V)	Statvolt (c/10^6 V)	Volt (V)
Electric Polarization	Statcoul/cm^2 ($10^3/4\pi$c C/m^2)	Abcoul/cm^2 ($10^{-4}/4\pi$ C/m$^{2)}$	Statcoul/cm^2 ($10^3/4\pi$c C/m^2)	C/m^2
Electric Displacement	Statcoul/cm^2 ($10^3/4\pi$c C/m^2)	Abcoul/cm^2 ($10^{-4}/4\pi$ C/m^2)	Statcoul/cm^2 ($10^3/4\pi$c C/m^2)	C/m^2
Resistance	Statohm (c^2/10^5 Ω)	Abohm (10^{-9} Ω)	Second/cm (c^2/10^5 Ω)	Ohm (Ω)
Capacitance	Statfarad (10^5/c^2 F)	Abfarad (10^9 F)	cm (10^5/c^2 F)	Farad (F)
Magnetic Flux	Statvolt-second (c/10 ^{6} Wb)	Maxwell (10^{-8} Wb)	Maxwell (10^{-8} Wb)	Weber (Wb)
Magnetic Induction	Statvolt-sec/cm^2 (c/10^2 Wb/m$^{2)}$	Gauss (10^{-4} T)	Gauss (10^{-4} T)	Tesla (Wb/m$^{2)}$
Magnetic Field	Statamp-turn/cm ($10^3/4\pi$c A-turn/m)	Oersted ($10^3/4\pi$ A-turn/m)	Oersted ($10^3/4\pi$ A-turn/m)	Amp-turn/m
Magnetization	Statvolt-sec/cm^2 (c/4πx10^2 Wb/m^2)	Gauss ($10^{-4}/4\pi$ Wb/m^2)	Gauss ($10^{-4}/4\pi$ Wb/m$^{2)}$	Wb/m^2
Inductance	Stathenry (c^2/10^5 Hy)	Abhenry (10^{-9} Hy)	Stathenry? (c^2/10^5 Hy)	Henry (Hy)
Permittivity	ε ($\varepsilon_0=1$)	εc^2 ($\varepsilon_0=1/c^2$)	ε ($\varepsilon_0=1$)	$\varepsilon/\varepsilon_0$
Permeability	μc^2 ($\mu_0=1/c^2$)	μ ($\mu_0=1$)	μ ($\mu_0=1$)	μ/μ_0

Table A-8. Conversion factors for variables in Gaussian and SI units systems.

VARIABLE	GAUSSIAN	SI
VELOCITY OF LIGHT	c	$\left(\mu_0\varepsilon_0\right)^{-1/2}$
ELECTRIC FIELD	$\vec{E}$	$\left(4\pi\varepsilon_0\right)^{1/2}\vec{E}$
ELECTRIC POTENTIAL	Φ	$\left(4\pi\varepsilon_0\right)^{1/2}\Phi$
VOLTAGE	V	$\left(4\pi\varepsilon_0\right)^{1/2}V$
ELECTRIC DISPLACEMENT	$\vec{D}$	$\left(4\pi\varepsilon_0\right)^{1/2}\vec{D}$
CHARGE	q	$q/\left(4\pi\varepsilon_0\right)^{1/2}$
CHARGE DENSITY	ρ	$\rho/\left(4\pi\varepsilon_0\right)^{1/2}$
CURRENT	I	$I/\left(4\pi\varepsilon_0\right)^{1/2}$
CURRENT DENSITY	$\vec{J}$	$\vec{J}/\left(4\pi\varepsilon_0\right)^{1/2}$
ELECTRIC POLARIZATION	$\vec{P}$	$\vec{P}/\left(4\pi\varepsilon_0\right)^{1/2}$
MAGNETIC INDUCTION	$\vec{B}$	$\left(4\pi/\mu_0\right)^{1/2}\vec{B}$
MAGNETIC FIELD	$\vec{H}$	$\left(4\pi\mu_0\right)^{1/2}\vec{H}$
MAGNETIZATION	$\vec{M}$	$\left(\mu_0/4\pi\right)^{1/2}\vec{M}$
CONDUCTIVITY	σ	$\sigma/\left(4\pi\varepsilon_0\right)$
DIELECTRIC CONSTANT (PERMITTIVITY)	ε	$\varepsilon/\varepsilon_0$
PERMEABILITY	μ	μ/μ_0
RESISTANCE	R	$\left(4\pi\varepsilon_0\right)R$
IMPEDANCE	Z	$\left(4\pi\varepsilon_0\right)Z$
INDUCTANCE	L	$\left(4\pi\varepsilon_0\right)L$
CAPACITANCE	C	$C/\left(4\pi\varepsilon_0\right)$

Natural Units

In specialized fields (like relativistic quantum theory or atomic physics) there are often certain quantities that are "natural" choices for units. For example, consider a warehouse. In this warehouse all materials are stored on pallets. Shipments involve receiving or sending of complete pallets. As inventories shift, complete pallets are moved from location to location to maintain a logical and efficient storage system. In such a warehouse, the logical unit is a "pallet". When a worker says he just relocated 37 Commodity X from site A to site D, the "natural unit" of "pallet" can be assumed without its explicit inclusion. This type of thinking has occurred numerous times in various fields of physics.

In relativistic quantum theory [3] it is natural to use velocities in units of c (the speed of light) and actions (energy times time) in units of $\hbar$ (Planck's constant divided by 2π). Because of this, it is not uncommon for relativistic quantum theorists to set $c = \hbar = 1$ in equations. Use of these **natural units** leads to considerable simplification of the equations. All computations can be performed in natural units and at the very end, if it is desired to express the results in SI units, powers of $\hbar$ and c can be added back into the final equation to make the dimensions correct (see the section on dimensional analysis later in this appendix).

In atomic physics [4], it is common to use **atomic units**, a set of natural units introduced by Hartree as a modification of the Gaussian system of units.[5] In atomic units, $\mathrm{e} = \hbar = m_e = 1$ in all equations. In atomic units, the unit of charge must be the electronic charge and the unit of mass must be the electron mass. The unit of length will become the Bohr radius

$$a_0 = \hbar^2 / m_e \mathrm{e}^2$$

The unit of energy is twice the ionization potential of the hydrogen atom = $m_e\mathrm{e}^4/\hbar^2$. The unit of velocity is the electron velocity in the first Bohr orbit = $\mathrm{e}^2/\hbar = \alpha c$ = the fine structure constant times the speed of light. The unit of time = $\hbar^3/m_e\mathrm{e}^4$ and so on. Once again, at the end of all calculations, powers of $\hbar$, m_e, and e can be added back in based on dimensional analysis.

Other natural unit systems may be encountered. People working on unified field theories or quantum gravity might use units in which $G = \hbar = c = 1$. The unit of mass would then become the Planck mass, the unit of length would become the Planck length, and the unit of time would become the Planck time (see Table A- for definition of the Planck units). It is even conceivable that someone might set $\mathrm{e} = \hbar = m_e = c = 1$. The author remembers hearing a discussion of this possibility as a student, but does not remember any details and has never seen it successfully used. It is probably apocryphal, because $c = 1$ would require the unit of velocity to be the speed of light, while $\mathrm{e} = \hbar = m_e = 1$ requires the unit of velocity to be αc, an inconsistency. Setting too many fundamental constants equal to unity will result in such inconsistencies.

Decibels and Logarithmic Units

The **level** of a field quantity F (such as electric field strength, voltage, pressure, etc.) is defined by the relation[6]

$$L_F = \ln\left(F / F_0\right) \tag{A.1a}$$

where F_0 is a reference amplitude. The level of a power quantity P (such as power, intensity, etc.) is defined by the relation

$$L_P = 0.5\ln\left(P / P_0\right) \tag{A.1b}$$

where P_0 is a reference power. Since power quantities are usually the squares of associated field quantities, i.e., $P = F^2$, then the levels are related by

$$L_P = 0.5\ln\left(P / P_0\right) = 0.5\ln\left(F^2 / F_0^2\right) = \ln\left(F / F_0\right) = L_F\,. \tag{A.2}$$

The definitions have been arranged such that the levels of the field quantities and their associated powers are equal. When expressed as a level, there is no need to distinguish between fields and powers.

One **neper (Np)** is the level of a field quantity when $F/F_0 = e$, that is, when $\ln(F/F_0) = 1$. Thus, we see that the fundamental unit of level for field quantities is nepers. The word neper is derived from John Napier's (the inventor of the logarithm) spelling of his own name. By examination of Eqs. (A.1) and (A.2) we see that one neper must be the level of a power quantity when $P/P_0 = e^2$, that is, when $0.5 \ln(P/P_0) = 1$. The level equations Eq. (A.1) have units of nepers

$$L_F = \ln\left(F / F_0\right) \quad \text{Np} \tag{A.3a}$$

and

$$L_P = 0.5\ln\left(P / P_0\right) \quad \text{Np}\,. \tag{A.3b}$$

One **bel (B)** is the level of a field quantity when $F/F_0 = 10^{1/2}$, that is, when $2 \log(F/F_0) = 1$. In the same vein, one bel is the level of power quantity when $P/P_0 = 10$, that is, when $\log(P/P_0) = 1$.

In units of bels, the level equations become

$$L_F = 2\log\left(F / F_0\right) \quad \text{B} \tag{A.4a}$$

and

$$L_P = \log\left(P / P_0\right) \quad \text{B}\,. \tag{A.4b}$$

Note that the bel is not simply the base-10 logarithm equivalent of the base-*e* unit called the neper. The neper is naturally defined with respect to the field quantities, while the bel is naturally defined with respect to the power quantities. Both bel and neper differ from most other units in the respect that they are **logarithmic units** as opposed to **arithmetic** units.

Logarithmic units are convenient when measured values can vary by many orders of magnitude. As we will see below, the lowest sound pressure detectable by human hearing is roughly 1 μPa (= 0 dB); hearing damage can begin to occur at extended exposures to sound pressures even lower than 0.1 Pa = 10^5 μPa = 100 dB; death will occur at sound levels of around 100 kPa = 10^{11} μPa = 220 dB. Given that interesting sound pressure levels vary over at least 11 orders of magnitude when described in the linear units of Pascals, a logarithmic unit such as the decibel has the potential benefit of only varying from 0 dB to 220 dB).

In the acoustics community, the decibel (dB) is the universally used and accepted unit of sound intensity *I* or sound pressure *p*. Technically, both *I* and *p* are levels (i.e., *I* = sound intensity level and *p* = sound pressure level). However, in common usage the word level is often omitted. One decibel is one-tenth of a bel. The use of decibel rather than bel undoubtedly arose from practical considerations. Early measurement accuracies were likely of the order of 1 decibel. Thus, data could be reported in integer decibels. The use of bels would have lead to many fractional numbers and more complex record-keeping.

The acoustic units are thus defined by the relations [7]

$$I(\text{in dB}) = 10\log\left[I / I_0\right] \tag{A.5a}$$

or

$$p(\text{in dB}) = 20\log\left[p(\text{in }\mu\text{Pa}) / p_0\right] \tag{A.5b}$$

where I_0 is a reference intensity that is produced by a reference sound pressure p_0. Acoustic intensity is proportional to the square of acoustic pressure. In atmospheric acoustics, the reference pressure is p_0 = 20 μPa; in underwater acoustics, the reference pressure is p_0 = 1 μPa. The 20 μPa reference level was chosen because it is the lowest sound pressure that can be perceived by the average young adult in a perfectly quiet environment. All audible sounds will have positive levels. Sounds with negative levels will never be audible. Since audibility is seldom a consideration in underwater sound measurements (humans were not designed to hear well underwater) and given that the acoustic impedance of sea water is considerably different than air, it was reasonable to assume a more obvious standard (i.e., some power of ten relative to the SI unit of pressure, the Pascal).

Other applications have found logarithmic units to be valuable, such as the Richter scale of earthquake magnitudes. However, rather than developing a new and unique name, many of these other applications have adopted the term decibels. In this more general application,

$$X(\text{in dB}) = 10\log\left[X / X_0\right] \tag{A.6}$$

where X_0 is a reference level. Carrier-to-Noise Ratio = Carrier Power/Noise Power = CNR is almost universally given in units of decibels (the noise power acts as a convenient reference level). Note that if two quantities are related by one being the square of the other, the squared quantity (e.g., power or sound intensity) has a multiplier 10 in Eq. (A.5a) while the other quantity (e.g., voltage or sound pressure) has a multiplier of 20. Refer back to Eq. (A.4a & b)

In some instances X_0 is set equal to 1 (dimensionless). In this case then the decibel value of X carries an explicit unit. For example, in microwave transmission systems it is convenient to discuss microwave power level in units of dBW, that is,

$$P(\text{in dBW}) = 10\log\left[P(\text{in W})\right]. \tag{A.7}$$

The decibel units may become fairly complex as in the case of electromagnetic intensity

$$\Phi(\text{in dBJ / m}^2\text{ - s}) = 10\log\left[P(\text{in J / m}^2\text{ - s})\right]. \tag{A.8}$$

The special unit **dBi** is sometimes used in discussions of antenna performance. In this case the "i" refers to "isotropic". An antenna with a gain in a particular direction of -20 dBi has a gain that is 20 dB below that of an isotropic radiator (equally intensity emitted in all directions) at the same wavelength.

The choice between using decibel (logarithmic) or linear units may depend on something as simple as which type of units is easier to graph. For example, exponential curves plot as straight lines when one of the units is logarithmic. One could plot the exponential using a logarithmic scale (semi-log paper) or one could plot the exponential on linear scaled paper with logarithmic units. If the plot covers more than 5 orders of magnitude the decibel plot will be much less cluttered with grid lines than the equivalent semi-log plot.

When using logarithmic units, one must be careful not to improperly mix logarithmic and linear quantities. For example, the expression for the carrier-to-noise ratio of a radar may be written in the linear form

$$CNR = a\ P\ D^4\ \sigma\ /\ R^4 \tag{A.9}$$

where P is in W, D is in m, σ is in m^2, R is in m, and a is a collection of constants and factors that make CNR dimensionless. Alternately, it may be written in logarithmic form

$$10\log\left[CNR\right] = 10\log\left[a\ P\ D^4\ \sigma\ /\ R^4\right] \tag{A.10}$$

or

$$\begin{aligned} CNR(\text{in dB}) = 10\log a + P(\text{in dBW}) + 40\log\left(D \text{ in m}\right) \\ + \sigma(\text{in dBsm}) - 40\log\left(R \text{ in m}\right) \end{aligned} \tag{A.11}$$

where several of the units have been expressed in logarithmic form (decibel form). In particular the radar cross section is expressed in dBsm (decibel square meters -- the radar cross section in decibels relative to 1 square meter). *CNR (in dB)* must be converted to *CNR (linear)* before it may be used in equation (A.9). *P* and σ must be similarly converted. It is common for some parameters used in a calculation to be expressed logarithmically and others to be expressed linearly. It is essentially that one form or the other must be chosen before computations are made, and all parameters must be converted to the chosen form. Failure to do this is an exceedingly common, easily understandable, and **absolutely inexcusable** error.

Given the difficulties that logarithmic units present, one may wonder why we bother with them. The answer is tradition. In the not so distant past (roughly 1975), before computers and scientific calculators became commonplace, the logarithmic form was often more easily calculated (sums and differences are intrinsically easier than multiplications and divisions). Much of our current use of decibel units is a holdover from the much earlier "necessity" to use the logarithmic equations rather than the linear equations. In some fields, key equations (those often used to perform computations) are still given in logarithmic form and commonly used graphs are plotted in logarithmic units, and all of the experienced practitioners are still fluent in the use of the logarithmic forms. Since junior personnel often emulate successful older practitioners, the junior people keep the tradition alive.

Furthermore, there is often a need to do quick and dirty calculations. These are usually referred to as "back-of-the-envelope" calculations. Unfortunately, the designer sometimes does not have even an envelope handy. Calculations using the add-and-subtract logarithmic form of the equation and quick decibel approximations to the numbers can often be performed quickly in one's head, and may yield answers of acceptable accuracy. If one remembers that

1	= 0 dB
1.25	= 1 dB
1.6	= 2 dB
2	= 3 dB
2.5	= 4 dB
e	= 4.343 dB
π	= 5 dB
4	= 6 dB
5	= 7 dB
6.25	= 8 dB
8	= 9 dB
10	= 10 dB

to good accuracy, then mental computations are considerably simplified. Although the need to have logarithmic units has almost disappeared, it is assured that system designers will continue to encounter them for a long time to come.

Units of Attenuation

Many physical quantities lose strength (attenuate) as they pass through material media. In many instances, the attenuation as a function of distance traveled (*R*) takes an exponential form

$$X\left(R\right) = X\left(0\right)e^{-\alpha R} . \tag{A.12}$$

where α is the attenuation coefficient. Rearranging equation (A.8) and taking the natural logarithm of both sides yields

$$\alpha R = -\ln\left[X\left(R\right) / X\left(0\right)\right] . \tag{A.13}$$

This bears a resemblance to the definition of a level. However, we should first remark that *X* is more often a power quantity rather than a field quantity. If *X* is a power quantity, then we observe

$$\begin{aligned} \alpha R &= -\ln\left[X\left(R\right) / X\left(0\right)\right] = -2\left(0.5\ln\left[X\left(R\right) / X\left(0\right)\right]\right) \\ &= -2\left(\left[X\left(R\right) / X\left(0\right)\right] \text{ in Np}\right) \end{aligned} . \tag{A.14}$$

In fact the quantity *0.5* αR is assigned the dimensionless unit of neper. One neper implies an attenuation in *X* of a factor of *1/e*^{2}. Since *R* has units of length, the attenuation coefficient α must have units of nepers/unit length. Multiplying α by the distance traveled *R* give the total number of nepers (powers of *1/e*^{2}) of attenuation the quantity *X* has undergone. In practice, the term neper is almost never used anymore and attenuation coefficients are given units of inverse distance (e.g., α might have units of km^{-1}). The neper and the hidden factor of two are quietly assumed. Since the attenuation coefficient always multiplies a distance and always goes into an exponential, this usually causes no problems.

Unfortunately, in fields where decibels are commonly used, attenuation coefficients are commonly given in decibels per unit length (or total decibels over a fixed length). This leads to another common cause of error. You can't mix decibels and nepers in the same equation. Consider the example of atmospheric transmission of electromagnetic radiation intensity as defined by the exponential equation

$$T = e^{-\alpha R} = I\left(R\right) / I\left(0\right) . \tag{A.14}$$

In decibel form

$$T(\text{in dB}) = -A(\text{in dB / km}) \cdot R(\text{in km}) = 10\log\left[e^{-\alpha R}\right] \tag{A.15}$$

where we have called A the attenuation coefficient in decibel form. In neper form

$$T(\text{in Np}) = 0.5\alpha(\text{in km}^{-1})\cdot R(\text{in km}) = -0.5\ln\left[e^{-\alpha R}\right]. \tag{A.16}$$

If we wish to express α in Np/km then we must divide it by 2.

$$\alpha(\text{in Np - km}^{-1}) = 0.5\alpha(\text{in km}^{-1}) \tag{A.17}$$

Note that if the transmission equation is given as a ratio of field quantities, then the factor of two disappears and *α(in Np-km^{-1})* and *α(in km^{-1})* are identical.

Since the transmission must be the same ratio regardless of how it is calculated we must have the equality

$$10^{-0.1A(\text{in dB-km}^{-1})\cdot R(\text{in km})} = e^{-\alpha(\text{in km}^{-1})\cdot R(\text{in km})} \tag{A.18}$$

Taking the natural logarithm of both sides of this equation yields

$$A(\text{in dB - km}^{-1}) = \frac{\alpha(\text{in km}^{-1})}{0.1\ln 10} = \frac{\alpha(\text{in km}^{-1})}{0.2303} = 4.343\alpha(\text{in km}^{-1}) \tag{A.19}$$

for comparing the decibel and per kilometer forms of the attenuation coefficient. Since attenuation coefficients are used as exponents, failure to convert to the proper form for an equation will lead to significant errors. It is very common for the total attenuation to be expressed as $e^{-\alpha R}$ in a computation, yet have α be expressed in dB/km. In this example, α must be converted to km^{-1} form before computing the exponential.

Dimensional Analysis

Dimensional analysis is an extremely simple but powerful tool for eliminating errors creating information. The fundamental principles of dimensional analysis are that (a) the net dimensions of one side of an equation must equal the net dimensions of the other side of that equation, and (b) the dimensions of any specific physical quantity (regardless of how it is expressed) must have a unique, yet universally specific form when reduced to fundamental units of mass, length, time, and charge. Principle (b) deserves more elaboration. Energy can be expressed or calculated in many ways: kinetic energy ½ mv^2, gravitational potential energy mgh, rotational energy ½ $I\omega^2$, thermal energy kT, or electrical potential energy qV. However, energy is energy and must ultimately have units of mass x length^2 / time^2. All five of the preceding expressions when expressed in units of mass, length, time, and charge, indeed reduce to the fundamental form of mass x length^2 / time^2. The fundamental forms of a number of physical quantities and a few common physical constants are summarized in Table A-7.[8] It should be noted that there are a few older systems of units that have different dimensionality for the fundamental expression. The Gaussian set of units used in electromagnetism is one of these. The primary differences occur only in electrical or magnetic quantities. Energy will always ultimately end up with units of $\mathsf{ml^2/t^2}$.

As mentioned above, application of these principles can reduce errors. Consider this problem. You find an expression for a quantity you need in an old textbook. You are not sure this book consistently uses SI units. The formula is the Klein-Nishina formula for Compton scattering.

$$\frac{d\sigma}{d\Omega}\left(\text{in cm}^2\ /\ \text{sr}\right)=\frac{e^4}{2m^2c^4}\left(\frac{\nu'}{\nu}\right)^2\left(\frac{\nu}{\nu'}+\frac{\nu'}{\nu}-\sin^2\theta\right) \qquad \text{(A.20)}$$

The differential cross section supposedly is in units of cm^2/sr. You do a quick dimensional check after eliminating the dimensionless frequency terms in brackets.

$$\frac{d\sigma}{d\Omega}\propto\frac{e^4}{2m^2c^4}\Rightarrow\frac{\mathsf{q^4t^4}}{\mathsf{m^2l^4}} \qquad \text{(A.21)}$$

This expression has units of charge in it and is clearly not reducible to units of length squared. In SI units many quantities with charge that are not obviously electrical expressions (like Ohm's law) have the permittivity of free space $4\pi\varepsilon_0$ somewhere in the expression. Our expression does not. Looking in Table we find that permittivity has units of t^2 q^2 / m l^3. If we divide by $(4\pi\varepsilon_0)^2$ we can eliminate all of the charges.

$$\frac{d\sigma}{d\Omega}\propto\frac{e^4}{2m^2c^4\varepsilon_0{}^2}\Rightarrow\frac{\mathsf{q^4t^4}}{\mathsf{m^2l^4}\left(\mathsf{t^2q^2\ /\ ml^3}\right)^2}=\frac{\mathsf{q^4m^2l^6t^4}}{\mathsf{q^4m^2l^4t^4}}=\mathsf{l^2} \qquad \text{(A.22)}$$

The dimensionally correct expression becomes

$$\frac{d\sigma}{d\Omega}\left(\text{in cm}^2 / \text{sr}\right) = \frac{e^4}{32\pi^2 \varepsilon_0^2 m^2 c^4}\left(\frac{\nu'}{\nu}\right)^2 \left(\frac{\nu}{\nu'} + \frac{\nu'}{\nu} - \sin^2\theta\right). \qquad \text{(A.23)}$$

The reader may be interested to note that this is a real-life example, and at one time the author did need to perform a dimensional analysis to verify the proper form of the equation, as there was an error in the source.

It is much more common to catch simple mistakes in exponents, such as expressing distance traveled by an accelerating object as

$$d = \frac{1}{2}at\,. \qquad \text{(A.24)}$$

A quick dimensional check reveals

$$\mathsf{l} = \left(\frac{\mathsf{l}}{\mathsf{t}^2}\right)(\mathsf{t}) = \left(\frac{\mathsf{l}}{\mathsf{t}}\right) \qquad \text{(A.25)}$$

which is obviously incorrect by one power of time. The incorrect equation can be rapidly altered to read correctly

$$d = \frac{1}{2}at^2\,. \qquad \text{(A.26)}$$

Even more complicated exponent mistakes can be caught and diagnosed. For example, we may express the electron plasma frequency as

$$\omega_e = \frac{ne^2}{\varepsilon_0 m} \qquad \text{(A.27)}$$

where *n* is the electron density (l^{-3}). Dimensional analysis shows

$$\frac{1}{\mathsf{t}} = \frac{\left(1/\mathsf{l}^3\right)\left(\mathsf{q}^2\right)}{\left(\mathsf{t}^2\mathsf{q}^2/\mathsf{ml}^3\right)(\mathsf{m})} = \frac{1}{\mathsf{t}^2} \qquad \text{(A.28)}$$

The term on the right-hand side has a dimension which is the square of the term on the left-hand side. In this case it is easy to see that the correct expression for electron plasma frequency should be

$$\omega_e = \left(\frac{ne^2}{\varepsilon_0 m}\right)^{1/2} \qquad \text{(A.29)}$$

These examples should serve to convince the reader of the power and utility of dimensional analysis. It is a tool that is all too infrequently used, to the detriment of everyone.

Table A-7. Dimensions of Physical Quantities. The basic dimensions are mass (m), length (l), time (t), charge (q), and temperature (K).

Physical Quantity	Symbol	Dimensions	Physical Quantity	Symbol	Dimensions
Acceleration	a	l / t^2	Magnetic Flux	Φ	$m\ l^2 / t\ q$
Angular Momentum	J	$m\ l^2 / t$	Magnetic Induction	B	m / t q
Angular Velocity	ω	1 / t	Magnetic Moment	μ	$l^2\ q / t$
Area	A	l^2	Magnetization	M	q / l t
Boltzmann Constant	k	$m\ l^2 / t^2\ T$	Magnetomotive Force	$\mathcal{M}$	q / t
Capacitance	C	$t^2\ q^2 / m\ l^2$	Mass	m, M	m
Charge	q	q	Moment of Inertia	μ	$m\ l^2$
Charge Density	ρ	q / l^3	Momentum	p, P	m l / t
Conductance	t	$q^2 / m\ l^2$	Momentum Density	-	$m / l^2\ t$
Conductivity	σ	$t\ q^2 / m\ l^3$	Permeability	μ	$m\ l / q^2$
Current	I	q / t	Permittivity	ε	$t^2\ q^2 / m\ l^3$
Current Density	J	$q / l^2\ t$	Planck Constant	h, $\hbar$	$m\ l^2 / t$
Density	ρ	m / l^3	Polarization	P	q / l^2
Displacement, Range	d, x, R	l	Power	P	$m\ l^2 / t^3$
Electric Displacement	D	q / l^2	Power Density	-	$m / l\ t^3$
Electric Field	E	$m\ l / t^2\ q$	Pressure	p, P	$m / l\ t^2$
Electric Potential,	V, Φ	$m\ l^2 / t^2\ q$	Reluctance	$\mathcal{R}$	$q^2 / m\ l^2$
Electromotive Force	$\mathcal{E}$	$m\ l^2 / t^2\ q$	Resistance	R	$m\ l^2 / t\ q^2$
Energy	U, W	$m\ l^2 / t^2$	Resistivity	η, ρ	$m\ l^3 / t\ q^2$
Energy Density	w, ε	$m / l\ t^2$	Temperature	T	T
Force	F	$m\ l / t^2$	Thermal Conductivity	κ	$m\ l / t^3$
Frequency	f, ν, ω	1 / t	Time	t	t
Gas Constant	R	$m\ l^2 / t^2$ T mole	Torque	τ	$m\ l^2 / t^2$
Gravitational Constant	G	$l^3 / m\ t^2$	Vector Potential	A	m l / t q
Impedance	Z	$m\ l^2 / t\ q^2$	Velocity, Speed	v	l / t
Impulse	-	m l / t	Viscosity	η, μ	m / l t
Intensity	I	m / t^3	Volume	V	l^3
Inductance	L	$m\ l^2 / q^2$	Vorticity	ζ	1 / t
Length	l	l	Work	W	$m\ l^2 / t^2$
Magnetic Intensity	H	q / l t			

Useful Physical Constants

The derivations and physical arguments presented in the text make frequent use of numerical estimates. To facilitate these and other such estimates, in Table A-6 we present a list of fundamental physical constants (in SI units). The values listed are from the 2002 values [11] recommended by the Committee on Data for Science and Technology of the International Council of Scientific Unions (CODATA). They are the best self-consistent set of values available and have replaced those from the last CODATA least squares adjustments completed in 1986 [9] and 1998 [10]. A number of other useful constants extracted from the CRC Handbook of Chemistry and Physics [12] and Lang's Astrophysical Data [13] are also included. In most cases, the last digit given for each experimentally determined constant is uncertain.

Table A-6. Useful physical constants.

PHYSICAL CONSTANT	**SYMBOL**	**VALUE**	**UNIT**
pi	π	3.1415926535897932	---
base of natural (Naperian) logarithms	e	2.7182818284590452	---
Euler's constant	γ	0.57721566490153286	---
speed of light	c	2.99792458×10^{8}	m/s
permittivity of free space	ε_0	$8.854187817 \times 10^{-12}$	F/m
permeability of free space	μ_0	$4\pi \times 10^{-7}$	$N\text{-}s^2/C^2$
impedance of free space ($[\mu_0/\varepsilon_0]^{1/2}$)	Z_0	376.730313461	Ω
Planck constant	h	$6.6260693 \times 10^{-34}$	J-s
" " ($h/2\pi$)	$\hbar$	$1.05457168 \times 10^{-34}$	J-s
elementary charge	e	$1.60217653 \times 10^{-19}$	C
Boltzmann constant	k	$1.3806505 \times 10^{-23}$	J/K
electron mass	m_e	$9.1093826 \times 10^{-31}$	kg
" mass-energy	$m_e c^2$	0.510998918	MeV
" magnetic moment	μ_e	$-9.28476412 \times 10^{-24}$	$J\text{-}T^{-1}$
muon mass	m_μ	$1.88353140 \times 10^{-28}$	kg
" mass-energy	$m_\mu c^2$	105.6583692	MeV
" magnetic moment	μ_μ	$-4.49044799 \times 10^{-26}$	$J\text{-}T^{-1}$
tau mass	m_τ	3.16777×10^{-27}	kg
" mass-energy	$m_\tau c^2$	1776.99	MeV
proton mass	m_p	$1.67262171 \times 10^{-27}$	kg
" mass-energy	$m_p c^2$	938.272029	MeV
" magnetic moment	μ_p	$1.41060671 \times 10^{-26}$	$J\text{-}T^{-1}$
neutron mass	m_n	$1.67492728 \times 10^{-27}$	kg
" mass-energy	$m_n c^2$	939.565360	MeV
" magnetic moment	μ_n	$-0.96623645 \times 10^{-26}$	$J\text{-}T^{-1}$

Table A-6. (cont.) Useful physical constants.

PHYSICAL CONSTANT	SYMBOL	VALUE	UNIT
deuteron mass	m_d	3.34358335 x 10^{-27}	kg
" mass-energy	m_dc^2	1875.61282	MeV
" magnetic moment	μ_d	0.43307342 x 10^{-26}	J-T^{-1}
helion mass	m_h	5.00641214 x 10^{-27}	kg
" mass-energy	m_hc^2	2808.39142	MeV
" magnetic moment	μ_h	-1.074553024 x 10^{-26}	J-T^{-1}
alpha mass	m_α	6.6446565 x 10^{-27}	kg
" mass-energy	$m_\alpha c^2$	3727.37917	MeV
hydrogen atom mass	m_H	1.6735325 x 10^{-27}	kg
atomic mass unit	u	1.66053886 x 10^{-27}	kg
Avogadro number	N_A	6.0221415 x 10^{23}	mol^{-1}
molar mass of hydrogen atoms	M(H)	1.007825034 x 10^{-3}	kg
fine structure constant ($e^2/2\varepsilon_0hc$)	α	7.297352568 x 10^{-3}	—
Rydberg constant ($m_ee^4/8\varepsilon_0{}^2h^3c$)	R_∞	1.0973731568525 x 10^7	m^{-1}
" "		3.289841960360 x 10^{15}	Hz
Bohr radius ($\varepsilon_0h^2/\pi m_ee^2$)	a_0	5.291772108 x 10^{-11}	m
Compton wavelength of electron (h/m_ec)	λ_C	2.426310238 x 10^{-12}	m
" " " " ($h/2\pi m_ec$)	$\bar{\lambda}_C$	3.861592678 x 10^{-13}	m
classical electron radius ($e^2/4\pi\varepsilon_0m_ec^2$)	r_e	2.817940325 x 10^{-15}	m
Thomson cross section ($8\pi r_e{}^2/3$)	σ_e	0.665245873 x 10^{-28}	m^2
Bohr magneton ($eh/4\pi m_e$)	μ_B	9.27400949 x 10^{-24}	J/T
nuclear magneton ($eh/4\pi m_p$)	μ_N	5.05078343 x 10^{-27}	J/T
magnetic flux quantum ($h/2e$)	Φ_0	2.06783372 x 10^{-15}	Wb
quantum of circulation ($h/2m_e$)		3.636947550 x 10^{-4}	m^2-s^{-1}
Stefan-Boltzmann constant ($2\pi^5k^4/15h^3c^2$)	σ	5.670400 x 10^{-8}	W/m^2-K^4
first radiation constant ($2\pi hc^2$)	c_1	3.74177138 x 10^{-16}	W-m^2
second radiation constant (hc/k)	c_2	0.014387752	m-K
Wien displacement law constant, $\lambda_{max}T$	b	2.8977685 x 10^{-3}	m-K
ideal gas constant	R	8.314472	J/mol-K
standard atmosphere	atm	1.013250000 x 10^5	Pa
molar volume of ideal gas (273.15K, 1atm)	V_m	22413.996	cm^3/mol
density of air at STP (273.15K, 1atm)	ρ_a	1.285	kg/m^3
Loschmidt constant (N_a/V_m)	n_0	2.6867773 x 10^{25}	m^{-3}
Faraday constant	F	96485.3383	C/mol
Newtonian constant of gravitation	G	6.6742 x 10^{-11}	m^3/kg-s^2
standard acceleration of gravity	g	9.80665	m/s^2
Planck mass ($[hc/2\pi G]^{1/2}$)	m_P	2.17645 x 10^{-8}	kg
Planck length ($[hG/2\pi c^3]^{1/2}$)	l_P	1.61624 x 10^{-35}	m
Planck time ($[hG/2\pi c^5]^{1/2}$)	t_P	5.39121 x 10^{-44}	s
Fermi coupling constant	$G_F/(\hbar c)^3$	1.16639 x 10^{-5}	GeV^{-2}
weak mixing angle	$\sin^2\theta_W$	0.22215	—

Table A-6. (cont.) Useful physical constants.

PHYSICAL CONSTANT	SYMBOL	VALUE	UNIT
Josephson constant	K_{J-90}	483597.879	$GHz-V^{-1}$
Von Klitzing constant	R_{K-90}	25812.807449	Ω
astronomical unit	AU	$1.4959787061 \times 10^{8}$	km
solar mass	M_s	1.9891×10^{30}	kg
earth mass	M_e	5.9742×10^{24}	kg
lunar mass	M_m	7.3483×10^{22}	kg
solar radius	R_s	6.9599×10^{8}	m
earth radius, average	R_e	6.3710×10^{6}	m
" " , polar		6.356755×10^{6}	m
" " , equatorial		6.378140×10^{6}	m
lunar radius	R_m	1.7380×10^{6}	m
earth-sun distance, average	a_e	$1.4959787061 \times 10^{8}$	km
" " , aphelion	Q	1.5207×10^{8}	km
" " , perihelion	q	1.4707×10^{8}	km
earth-moon distance, average	a_m	3.8440×10^{5}	km
" " , apogee	Q_m	4.0674×10^{5}	km
" " , perigee	q_m	3.5641×10^{5}	km
solar luminosity	L_s	3.826×10^{26}	J/s
solar radiation at top of atmosphere	I_0	1360.45	W/m^2
velocity of sound in dry air @ 0°C	c_s	331.36	m/s
viscosity of air at 288.15 K	μ	1.7894×10^{-5}	kg/m-s
density of water @ 3.98°C	ρ_W	1000.000000	kg/m^3
heat of fusion of water @ 0°C	ΔH_{fus}	333.51	J/kg
heat of vaporization of water @ 100°C	ΔH_{vap}	2257.48	J/kg
heat capacity of solid water @ 298 K	C_p^0	37.11	J/K-mol
heat capacity of liquid water @ 298 K	C_p^0	75.35	J/K-mol
heat capacity of water vapor @ 298 K	C_p^0	33.60	J/K-mol

Standard Temperature and Pressure (STP) is T = 273.15K and p = 1atm.

Selected Conversion Factors

Many units which are not part of the SI unit system may be encountered in discussions of sensors and weapons systems. The following list of selected conversion factors is presented to facilitate comparison and assimilation of results from different sources. The list is not organized according to any significant degree.

1 atm = 760 Torr = 101325.0 Pa = 29.9212598 in Hg = 14.695941 psi
= 1.01325 x 10^6 dyne/cm^2 = 1013.25 mbar = 33.90 ft H_2O @ 4°C

1 C = 0.1 nm = 10^{-10} m 1 Fermi = 10^{-15} m

1 fathom = 6 ft = 1.82880000 m

1 mil = 25.4 μm = 0.001 in = 0.0254 mm

1 m = 39.37007874 in = 1.093613298 yd

1 mi = 1609.3440000 m = 5280 ft = 1760.0000 yd

1 nmi = 1852.0000 m = 6076.1155 ft = 2025.3718 yd

1 furlong = 40 rods = 220 yards = 660 feet = 0.125 mi (statute) = 201.168 m

1 league (statute) = 0.86897625 leagues (naut. Int.) = 4.828032 km = 3 mi (statute)

1 light year (ly) = 9.460530 x 10^{15} m = 63239.74 AU

1 parsec (pc) = 3.085678 x 10^{16} m = 3.261633 ly

1 year ~ 365.25 days = 8766 hr = 525960 min = 31,557,600 sec ~ π x 10^7 sec

1 kt = 1 nmi/hr = 0.51444444 m/s = 1.150779451 mph

1 Liter = 1.056688 qt = 0.001 m^3

1 gal (US) = 4 qt = 128 fl.oz. = 3.7854 liters = 3785.4 cm^3

1 ounce (US fluid) = 29. 5727 ml

1 ounce (US fluid) = 8 drams = 2 tablespoons (tbsp) = 6 teaspoons (tsp)

1 US barrel of petroleum (bbl) = 42 gal = 5.6146 ft^3 = 158.983 liters

1 hogshead = 63 gal = 238.47427 l

1 bushel (US) = 32 dry qt = 37.2367 liq qt = 35.2381 liters

1 long ton (LT)= 2240 lbs = 1.12 ton = 1016.05 kg = 1.01605 metric ton (mT =tonne)

1 pound (lb) = 16 oz (Avoirdupois) = 12 oz (Troy) = 256 drams = 7000 grains

1 pound (lb) = 453. 59237 g

1 ounce (Avoirdupois) = 28.349523 g

1 carat = 0.2 g = 3.08647 grains = 0.00705479 oz (Avoirdupois)

1 Newton = 100,000 dynes

1 dyne = 10^{-5} N

1 foot-pound = 1.35582 J

1erg = 10^{-7} J

1 rad = 57.295780 °

1° = 17.45329237 mrad

1 μrad = 0.20626480 ″

1″ = 4.848137 μrad

1 cal (thermochemical) = 4.1840000 J = 4.1840000 x 10^{7} ergs
= 2.611447677 x 10^{19} eV

1 cal (Int'l. Steam Table) = 4.1868000 J = 4.1868000 x 10^{7} ergs
= 2.613195300 x 10^{19} eV

1 Btu (thermochemical) = 1054.350 J = 251.9957 cal (thermochemical)

1 Btu (Int'l Steam Table) = 1055.05585 J = 251.9957 cal (Int'l Steam Table)

1 eV = 1.60217653 x 10^{-19} J = 2.41798940 x 10^{14} Hz = 8065.54446 cm^{-1}

1 cm^{-1} = 2.99792458 x 10^{10} Hz = 1.23984190 x 10^{-4} eV
= 1.98644560 x 10^{-23} J = 4.74771893 x 10^{-27} kcal

1 eV/molecule = 23.0605492 kcal/mol = 9.64853377 x 10^{4} J/mol

1 cm^{-1}/molecule = 2.85914352 x 10^{-3} kcal/mol = 11.9626565 J/mol

1 kcal/mol = 6.94769460 x 10^{-21} J/molecule
= 349.755090 cm^{-1}/molecule = 0.0433641016 eV/molecule

kT (Boltzmann const. x Temperature) = 1 eV, when T = 1.160450476 x 10^{4} K
kT = 1 cm^{-1}, when T = 1.43877512 K

1 MeV = 1.78266180 x 10^{-30} kg

1 megaton of TNT (MT) = 1 x 10^{15} cal = 4.184 x 10^{15} J = 2.61144757 x 10^{28} MeV

1 Uranium or Plutonium fission = 200 MeV total energy released
= 180 MeV prompt energy released

1 Deuterium-Tritium fusion = 17.59 MeV energy released

1 MT = 1.4508 x 10^{26} fissions = 240.9 moles = 56.61 kg U-235 fissioned

1 kg U-235 fissioned = 17.665 kT

1 MT = 1.48462 x 10^{27} fusions = 2465 moles = 12.33 kg d-t fused

1 kg d-t fused = 81.10 kT

1 MT = 0.046553 kg matter-antimatter completely annihilated

1 kg Matter-Antimatter annihilation = 8.98755179 x 10^{16} J = 21.4808 MT

For photons: λ x E = 1.23984190 μm-eV

1 gauss = 0.0001 T = 10^5 gamma

1 hp = 0.7457 kW = 550 foot-pounds/sec = 0.706243 Btu/sec

1 ft-candle = 10.76391 lux 1 Lambert = 3183.099 cd/m^2

1 ft-L (foot-Lambert) = 3.426259 cd/m^2

1 acre = 43560 ft^2 = 0.0015625 sq.mi. = 4046.8564 m^2 = 0.40468564 hectare (ha)

1 township = 36 sections = 36 sq. mi. = 23,040 acres = 93.23957 km^2

1 acre-foot = 43560 ft^3 = 1233.482 m^3 = 325,851.43 gal

°F = (9/5)°C + 32 °C = (5/9)(°F - 32)

1 flick = 1 W/cm^2-sr-μm

1 ppmv = 10^{-6} P/RT = 4.4615 x 10^{-5} (273.15/T)(P/101325)
= 1.20272 x 10^{-7} (P/T) mole/m^3

1 ppmv = 10^{-6} PM/RT = 1.20272 x 10^{-4} MP/T mg/m^3

1 μgal = 10^{-8} m-s^{-2} ~ 10^{-9} "g"

1 Eötvös = 1 E = 10^{-9} s^{-2} = 0.1 μgal/m

Useful Mathematical Expressions [14], [15]

Differentials

$$d(uv) = u\,dv + v\,du \qquad d(u^n) = n\,u^{n-1}du$$

$$d\left(\frac{u}{v}\right) = \frac{v\,du - u\,dv}{v^2} \qquad \frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

$$d(e^u) = e^u du \qquad d(a^u) = a^u \ln a \;\; du$$

$$d(\ln u) = \frac{du}{u} \qquad d(\log_a u) = \frac{du}{u \ln a}$$

$$d(\sin u) = \cos u\; du \qquad d(\cos u) = -\sin u\; du$$

$$d(\tan u) = \sec^2 u\; du \qquad d(\cot u) = -\csc^2 u\; du$$

$$d(\sec u) = \sec u\; \tan u\; du \qquad d(\csc u) = -\csc u\; \cot u\; du$$

$$d(\sin^{-1} u) = \frac{du}{\sqrt{1-u^2}} \qquad d(\cos^{-1} u) = -\frac{du}{\sqrt{1-u^2}}$$

$$d(\tan^{-1} u) = \frac{du}{1+u^2} \qquad d\left(\int_a^u f(t)\,dt\right) = f(u)\,du$$

$$d(\sinh u) = \cosh u\; du \qquad d(\cosh u) = \sinh u\; du$$

$$d(\sinh^{-1} u) = \frac{du}{\sqrt{u^2+1}} \qquad d(\cosh^{-1} u) = \frac{du}{\sqrt{u^2-1}}$$

$$\frac{df(t,u,v,\cdots,z)}{ds} = \frac{\partial f}{\partial t}\frac{\partial t}{\partial s} + \frac{\partial f}{\partial u}\frac{\partial u}{\partial s} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial s} + \cdots + \frac{\partial f}{\partial z}\frac{\partial z}{\partial s}$$

Integrals

$$\int u\,dv = uv - \int v\,du \qquad \int du\,u^n = \frac{u^{n+1}}{n+1}$$

$$\int du\,e^u = e^u \qquad \int du\,a^u = a^u\,/\ln a$$

$$\int du\cos u = \sin u \qquad \int du\sin u = -\cos u$$

$$\int du\tan u = -\ln(\cos u)$$

$$\int du\sin^2 u \;=\; -\frac{1}{2}\cos u\;\sin u \;+\; \frac{1}{2}u \;=\; \frac{1}{2}u - \frac{1}{4}\sin 2u$$

$$\int du\sin^n u \;=\; -\frac{1}{n}\cos u\;\sin^{n-1}u \;+\; \frac{n-1}{n}\int du\sin^{n-2}u$$

$$\int du\cos^2 u \;=\; \frac{1}{2}\cos u\;\sin u \;+\; \frac{1}{2}u \;=\; \frac{1}{2}u + \frac{1}{4}\sin 2u$$

$$\int du\cos^n u \;=\; \frac{1}{n}\cos^{n-1}u\;\sin u \;+\; \frac{n-1}{n}\int du\cos^{n-2}u$$

$$\int du\frac{1}{a^2+u^2} = \frac{1}{a}\tan^{-1}\left(\frac{u}{a}\right)$$

$$\int du\frac{1}{a^2-u^2} = \frac{1}{a}\tanh^{-1}\left(\frac{u}{a}\right) = \frac{1}{2a}\ln\left(\frac{a+u}{a-u}\right) \qquad \left(a^2 > u^2\right)$$

$$\int du\frac{1}{u^2-a^2} = -\frac{1}{a}\coth^{-1}\left(\frac{u}{a}\right) = \frac{1}{2a}\ln\left(\frac{u+a}{u-a}\right) \qquad \left(u^2 > a^2\right)$$

$$\int du\frac{1}{\sqrt{a^2-x^2}} = \sin^{-1}\left(\frac{u}{|a|}\right) = -\cos^{-1}\left(\frac{u}{|a|}\right) \qquad \left(a^2 > u^2\right)$$

Definite integrals

$$\int_0^\infty du\, e^{-au} = \frac{1}{a}$$

$$\int_0^\infty du\, u^{n-1}\, e^{-u} = \Gamma(n) = (n-1)!$$

$$\int_0^\infty du\, u^n\, e^{-au} = \frac{\Gamma(n+1)}{a^{n+1}} = \frac{n!}{a^{n+1}}$$

$$\int_0^\infty du\, e^{-a^2u^2} = \frac{\sqrt{\pi}}{2a}$$

$$\int_0^\infty du\, u\, e^{-a^2u^2} = \frac{1}{2a^2}$$

$$\int_0^\infty du\, u^2\, e^{-a^2u^2} = \frac{\sqrt{\pi}}{4a^3}$$

$$\int_0^\infty du\, u^{2n}\, e^{-a^2u^2} = \frac{1\cdot 3\cdot 5\cdots(2n-1)\sqrt{\pi}}{2^{n+1}a^{2n+1}}$$

$$\int_0^\infty du\, \frac{u^3}{e^u-1} = \frac{\pi^4}{15}$$

$$\int_{-\infty}^\infty du\, \frac{1}{(1+a^2u^2)(1+b^2u^2)} = \frac{\pi}{a+b}$$

$$\int_0^{\pi/2} du\, \sin^n u = \begin{cases} \dfrac{1\cdot 3\cdot 5\cdot 7\cdots(n-1)}{2\cdot 4\cdot 6\cdot 8\cdots n}\dfrac{\pi}{2} & n = \text{even integer } (n \neq 0) \\[2ex] \dfrac{2\cdot 4\cdot 6\cdot 8\cdots(n-1)}{1\cdot 3\cdot 5\cdot 7\cdots n} & n = \text{odd integer } (n \neq 1) \end{cases}$$

Quadratic equation

$$\text{If} \quad ax^2 + bx + c = 0, \quad \text{then} \quad x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Cubic equation

If $\qquad y^3 + py^2 + qy + r = 0$

then let $\qquad a = \left(3q - p^2\right)/3 \qquad$ and $\qquad b = \left(2p^3 - 9pq + 27r\right)/27$

Also let $\qquad A = \sqrt[3]{\frac{-b}{2} + \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}} \qquad$ and $\qquad B = \sqrt[3]{\frac{-b}{2} - \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}}$

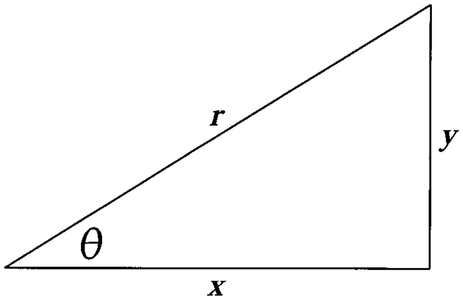
Then the solutions are

$$x = \begin{cases} A + B, \\ -\frac{A+B}{2} + \frac{A-B}{2}\sqrt{3}, \\ -\frac{A+B}{2} - \frac{A-B}{2}\sqrt{3} \end{cases}$$

Functions of right triangles

Pythagorean theorem

$$x^2 + y^2 = r^2$$



$\sin\theta = y/r$

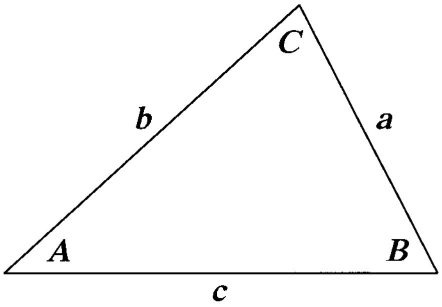
$\cos\theta = x/r$

$\sec\theta = r/x = 1/\cos\theta \qquad\qquad \csc\theta = r/y = 1/\sin\theta$

$$\tan\theta = x / y = \frac{\sin\theta}{\cos\theta}$$

$$\cot\theta = y / x = 1/\tan\theta$$

Laws of sines and cosines



$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}$$

= diameter of circumscribed circle
(See below)

$$a^2 = b^2 + c^2 - 2bc\cos A$$

$$b^2 = a^2 + c^2 - 2ac\cos B$$

$$c^2 = a^2 + b^2 - 2ab\cos C$$

$$a = b\cos C + c\cos B$$

$$b = a\cos C + c\cos A$$

$$c = a\cos B + b\cos A$$

$$\sin A = \frac{2}{bc}\sqrt{s(s-a)(s-b)(s-c)} \qquad \text{where} \quad s = (a+b+c)/2$$

Trigonometric identities

$$\sin^2\theta + \cos^2\theta = 1 \qquad \sec^2\theta - \tan^2\theta = 1 \qquad \csc^2\theta - \cot^2\theta = 1$$

$$\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta$$

$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta$$

$$\tan(\alpha \pm \beta) = \frac{\tan\alpha \pm \tan\beta}{1 \mp \tan\alpha\tan\beta}$$

$$\sin 2\theta = 2\sin\theta\cos\theta$$

$$\cos 2\theta = \cos^2\theta - \sin^2\theta = 2\cos^2\theta - 1 = 1 - 2\sin^2\theta$$

$$\tan 2\theta = \frac{2\tan\theta}{1-\tan^2\theta}$$

$$\sin 3\theta = 3\sin\theta - 4\sin^3\theta$$

$$\cos 3\theta = 4\cos^3\theta - 3\cos\theta$$

$$\tan 3\theta = \frac{3\tan\theta - \tan^3\theta}{1-3\tan^2\theta}$$

$$\sin 4\theta = 8\sin\theta\cos^3\theta \; - \; 4\sin\theta\cos\theta$$

$$\cos 4\theta = 8\cos^4\theta \; - \; 8\cos^2\theta \; + \; 1$$

$$\sin\alpha + \sin\beta = 2\sin 0.5(\alpha+\beta)\cdot\cos 0.5(\alpha-\beta)$$

$$\sin\alpha - \sin\beta = 2\cos 0.5(\alpha+\beta)\cdot\sin 0.5(\alpha-\beta)$$

$$\cos\alpha + \cos\beta = 2\cos 0.5(\alpha+\beta)\cdot\cos 0.5(\alpha-\beta)$$

$$\cos\alpha - \cos\beta = -2\sin 0.5(\alpha+\beta)\cdot\sin 0.5(\alpha-\beta)$$

$$\tan\alpha + \tan\beta = \frac{\sin(\alpha+\beta)}{\cos\alpha\cos\beta}$$

$$\tan\alpha - \tan\beta = \frac{\sin(\alpha-\beta)}{\cos\alpha\cos\beta}$$

$$\sin\alpha\sin\beta = 0.5[\cos(\alpha+\beta) - \cos(\alpha-\beta)]$$

$$\cos\alpha\cos\beta = 0.5[\cos(\alpha+\beta) + \cos(\alpha-\beta)]$$

$$\sin\alpha\cos\beta = 0.5[\sin(\alpha+\beta)+\sin(\alpha-\beta)]$$

$$\cos\alpha\sin\beta = 0.5[\sin(\alpha+\beta)-\sin(\alpha-\beta)]$$

$$\sinh x = \frac{e^x - e^{-x}}{2} \qquad \cosh x = \frac{e^x + e^{-x}}{2}$$

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}} \qquad \coth x = \frac{1}{\tanh x}$$

$$\operatorname{sech} x = \frac{1}{\cosh x} \qquad \operatorname{csch} x = \frac{1}{\sinh x}$$

$$e^{\pm i\theta} = \cos\theta \pm i\sin\theta$$

Taylor series expansion

$$f(x+x_0) = f(x_0) + f'(x_0)x + f''(x_0)\frac{x^2}{2!} + f'''(x_0)\frac{x^3}{3!} + \cdots$$

Expansions

$$(1\pm x)^{-1} = 1 \mp x + x^2 \mp x^3 + \cdots$$

$$(1+x)^{1/2} = 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} - \cdots$$

$$(1+x)^{-1/2} = 1 - \frac{x}{2} + \frac{3x^2}{8} - \frac{5x^3}{16} + \cdots$$

$$(1\pm x)^{-1} = 1 \mp x + x^2 \mp x^3 + x^4 \mp x^5 + \cdots$$

$$(1\pm x)^{-2} = 1 \mp 2x + 3x^2 \mp 4x^3 + 5x^4 \mp 6x^5 + \cdots$$

$$\left(1 \pm x\right)^n = 1 \pm nx + \frac{n\left(n-1\right)}{2!}x^2 \pm \frac{n\left(n-1\right)\left(n-2\right)}{3!}x^3 + \cdots$$

$$\left(1 \pm x\right)^{-n} = 1 \mp nx + \frac{n\left(n+1\right)}{2!}x^2 \mp \frac{n\left(n+1\right)\left(n+2\right)}{3!}x^3 + \cdots$$

$$\left(x \pm y\right)^n = x^n \mp \frac{n}{1!}x^{n-1}y + \frac{n\left(n-1\right)}{2!}x^{n-2}y^2 \mp \cdots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

$$a^x = 1 + x\ln a + \frac{\left(x\ln a\right)^2}{2!} + \frac{\left(x\ln a\right)^3}{3!} + \cdots$$

$$\ln x = \frac{x-1}{x} + \frac{1}{2}\left(\frac{x-1}{x}\right)^2 + \frac{1}{3}\left(\frac{x-1}{x}\right)^3 + \cdots \qquad x > \frac{1}{2}$$

$$\ln x = \left(x-1\right) - \frac{1}{2}\left(x-1\right)^2 + \frac{1}{3}\left(x-1\right)^3 - \cdots \qquad 0 < x \le 2$$

$$\ln x = 2\left[\left(\frac{x-1}{x+1}\right) + \frac{1}{3}\left(\frac{x-1}{x+1}\right)^3 + \frac{1}{5}\left(\frac{x-1}{x+1}\right)^5 + \cdots\right] \qquad x > 0$$

$$\ln\left(a+x\right) = \ln a + 2\left[\left(\frac{x}{2a+x}\right) + \frac{1}{3}\left(\frac{x}{2a+x}\right)^3 + \frac{1}{5}\left(\frac{x}{2a+x}\right)^5 + \cdots\right]$$

where $a > 0, \quad -a < x < +\infty$

$$n! = n \times \left(n-1\right) \times \cdots \times 2 \times 1 \cong e^{-n}n^n\sqrt{2\pi n} \qquad \text{for large n}$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots$$

$$\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \frac{62x^9}{2835} + \cdots$$

$$\sin^{-1} x = x + \frac{1}{2\cdot 3}x^3 + \frac{1\cdot 3}{2\cdot 4\cdot 5}x^5 + \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6\cdot 7}x^7 + \cdots$$

$$\cos^{-1} x = \frac{\pi}{2} - \left[x + \frac{1}{2\cdot 3}x^3 + \frac{1\cdot 3}{2\cdot 4\cdot 5}x^5 + \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6\cdot 7}x^7 + \cdots\right]$$

$$\tan^{-1} x = \begin{cases} x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \cdots & x^2 < 1 \\ \frac{\pi}{2} - \frac{1}{x} + \frac{1}{3x^3} - \frac{1}{5x^5} + \frac{1}{7x^7} - \cdots & x > 1 \\ -\frac{\pi}{2} - \frac{1}{x} + \frac{1}{3x^3} - \frac{1}{5x^5} + \frac{1}{7x^7} - \cdots & x < -1 \end{cases}$$

$$\sinh x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots$$

$$\cosh x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots$$

$$\tanh x = x - \frac{x^3}{3} + \frac{x^5}{15} - \frac{17x^7}{315} + \frac{62x^9}{2835} - \cdots \qquad \left(x^2 < \pi^2/4\right)$$

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4}$$

$$\sum_{k=1}^{n} k^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$$

Vector operators

$$\nabla\Phi = \frac{\partial\Phi}{\partial x}\hat{\mathrm{x}} + \frac{\partial\Phi}{\partial y}\hat{\mathrm{y}} + \frac{\partial\Phi}{\partial z}\hat{\mathrm{z}} \qquad \text{Gradient operator}$$

$$\nabla\cdot\vec{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} \qquad \text{Divergence operator}$$

$$\nabla\times\vec{A} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right)\hat{\mathrm{x}} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right)\hat{\mathrm{y}} + \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)\hat{\mathrm{z}} \qquad \text{Curl operator}$$

$$\nabla^2 = \nabla\cdot\nabla = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) \qquad \text{Laplacian operator} \qquad \nabla\cdot\nabla\Phi = \nabla^2\Phi$$

$$\nabla^2(\Psi\Phi) = \Psi\nabla^2\Phi + 2\nabla\Psi\cdot\nabla\Phi + \Phi\nabla^2\Psi$$

$$\nabla\left(\nabla\cdot\vec{A}\right) = \nabla^2\vec{A} + \nabla\times\nabla\times\vec{A} \qquad\qquad \nabla\times\nabla\times\vec{A} = \nabla\left(\nabla\cdot\vec{A}\right) - \nabla^2\vec{A}$$

$$\nabla\times\nabla\Phi = 0 \qquad\qquad \nabla\cdot\left(\nabla\times\vec{A}\right) = 0$$

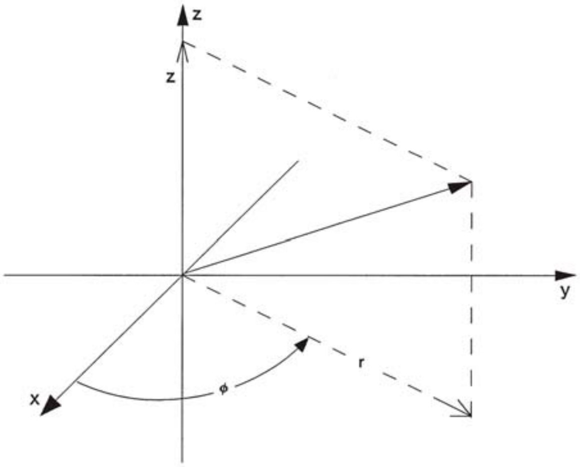
$$\nabla(\vec{u}\cdot\vec{v}) = (\vec{v}\cdot\nabla)\vec{u} + (\vec{u}\cdot\nabla)\vec{v} + \vec{v}\times\nabla\times\vec{u} + \vec{u}\times\nabla\times\vec{v}$$

Cylindrical coordinates

$$x = r\cos\phi$$

$$y = r\sin\phi$$

$$z = z$$



$$\nabla\Phi = \frac{\partial\Phi}{\partial r}\,\hat{r} + \frac{1}{r}\frac{\partial\Phi}{\partial\phi}\,\hat{\phi} + \frac{\partial\Phi}{\partial z}\,\hat{z}$$

$$\nabla^2\Phi = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\Phi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\Phi}{\partial\phi^2} + \frac{\partial^2\Phi}{\partial z^2}$$

$$\nabla\cdot\vec{A} = \frac{1}{r}\frac{\partial(rA_r)}{\partial r} + \frac{1}{r}\frac{\partial A_\phi}{\partial\phi} + \frac{\partial A_z}{\partial z}$$

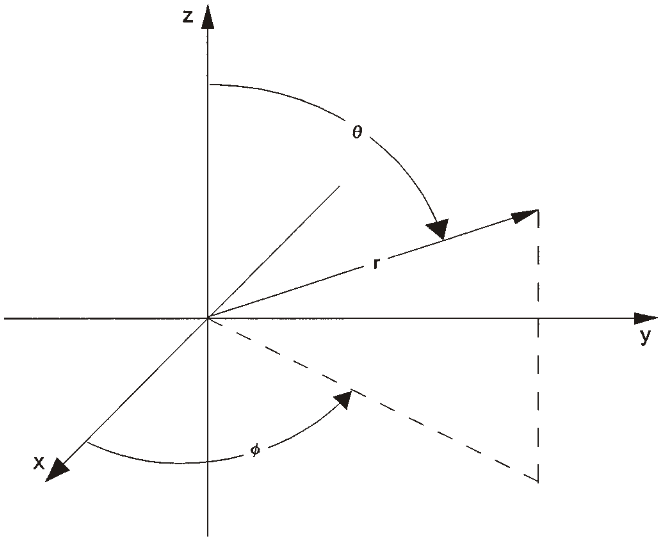
$$\nabla\times\vec{A} = \begin{vmatrix} \dfrac{\hat{r}}{r} & \hat{\phi} & \dfrac{\hat{z}}{r} \\ \dfrac{\partial}{\partial r} & \dfrac{\partial}{\partial\phi} & \dfrac{\partial}{\partial z} \\ A_r & rA_\phi & A_z \end{vmatrix}$$

Spherical coordinates

$$x = r\cos\phi\sin\theta$$

$$y = r\sin\phi\sin\theta$$

$$z = r\cos\theta$$



$$\nabla\Phi = \frac{\partial\Phi}{\partial r}\hat{r} + \frac{1}{r}\frac{\partial\Phi}{\partial\theta}\hat{\theta} + \frac{1}{r\sin\theta}\frac{\partial\Phi}{\partial\phi}\hat{\phi}$$

$$\nabla^2\Phi = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Phi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Phi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\Phi}{\partial\phi^2}$$

$$\nabla\cdot\vec{A} = \frac{1}{r^2}\frac{\partial\left(r^2A_r\right)}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial\left(\sin\theta\, A_\theta\right)}{\partial\theta} + \frac{1}{r\sin\theta}\frac{\partial A_\phi}{\partial\phi}$$

$$\nabla\times\vec{A} = \begin{vmatrix} \dfrac{\hat{r}}{r^2\sin\theta} & \dfrac{\hat{\theta}}{r\sin\theta} & \dfrac{\hat{\phi}}{r} \\ \dfrac{\partial}{\partial r} & \dfrac{\partial}{\partial\theta} & \dfrac{\partial}{\partial\phi} \\ A_r & rA_\theta & r\sin\theta\, A_\phi \end{vmatrix}$$

Miscellaneous Formulae

Circumference of circle (radius *a*)	$s = 2\pi a$
Area of circle (radius *a*)	$A = \pi a^2$
Area of an ellipse (axes *a*, *b*)	$A = \pi ab$
Surface area of sphere (radius *a*)	$A = 4\pi a^2$
Volume of sphere (radius *a*)	$V = 4\pi a^3 / 3$
Length of chord of circle (radius *a*, angular subtense *θ*, chord height *h*)	$s = 2a\sin(\theta / 2)$ $s = \left(4h(2a-h)\right)^{1/2}$
Area of curved surface of a right cone (height *h*, base radius *a*)	$A = \pi a\sqrt{a^2 + h^2}$
Volume of a right cone (height *h*, base radius *a*)	$V = \pi a^2 h / 3$
Radius of circle inscribed in triangle	$r = \dfrac{\sqrt{s(s-a)(s-b)(s-c)}}{s}$
Radius of circumscribed circle	$R = \dfrac{abc}{4\sqrt{s(s-a)(s-b)(s-c)}}$
(Triangle side lengths *a*, *b*, and *c*)	$s = (a + b + c) / 2$ (In 2 equations above)

References

[1] Nelson, Robert A., "Guide for Metric Practice", *Phys. Today*, 46 (8) Pt.2, BG15-BG16 (1993).

[2] Jackson, John D., Classical Electrodynamics 2nd Ed. (John Wiley and Sons, New York NY, 1975) pp. 811-821.

[3] Sakurai, J. J., Advanced Quantum Mechanics (Addison-Wesley, Reading MA, 1967) pp.179-181.

[4] Bethe, Hans A. and Salpeter, Edwin E., Quantum Mechanics of One- and Two-Electron Atoms (Plenum Press, New York NY, 1977) pp. 2-4.

[5] Hartree, D. R., *Proc. Cambridge Phil. Soc.*, 24, 89 (1928).

[6] National Institute of Science and Technology, Guide for the Use of the International System of Units (SI), NIST Special Publication 811 (U. S. Government Printing Office, Washington DC, 1995).

[7] Beranek, Leo L., Acoustics (McGraw-Hill Book Co., New York NY, 1954) pp. 12-13.

[8] Resnick, Robert and Halliday, David, Physics Part I (John Wiley and Sons, New York NY, 1967) Appendix F.

[9] Cohen, E. Richard, and Taylor, Barry N., The 1986 Adjustment of the Fundamental Physical Constants, CODATA Bulletin 63 (Pergamon Press, Elmsford NY, 1986).

[10] Mohr, Peter J., and Taylor, Barry N., "CODATA recommended values of the fundamental physical constants: 1998", *Rev. Mod. Phys.*, 72 (2), 351-495 (2000).

[11] Mohr, Peter J., and Taylor, Barry N., "The Fundamental Physical Constants", *Physics Today*, (2004). Available on the Internet at http://www.physicstoday.org/guide/fundconst.pdf .

[12] Weast, Robert C. (Ed.), CRC Handbook of Chemistry and Physics, 66th ed. (CRC Press, Boca Raton FL, 1985).

[13] Lang, Kenneth R., Astrophysical Data: Planets and Stars (Springer-Verlag, New York NY, 1992).

[14] Beyer, William T. (Ed.), CRC Standard Mathematical Tables 28th Ed. (CRC Press, Boca Raton FL, 1988).

[15] Abramowitz M. and Stegun, I. A., Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables National Bureau of Standards Applied Mathematics Series 55 (U. S. Government Printing Office, Washington DC, 1972).

APPENDIX B

FINITE DIFFERENCE AND FINITE ELEMENT TECHNIQUES

Finite Difference Approximation of Derivatives

Many systems of differential equations cannot be analyzed to yield closed-form solutions. Many numerical analysis techniques exist which allow the behavior of such systems to be modeled on a computer [1]-[5]. Approximation of derivatives by ratios of finite differences of the variables is one of the more easily applied.

The basis of the **finite difference** approach is found in the definition of a derivative. Given a function *x* of variable *t*, the derivative of *x* with respect to *t* is defined as [6]

$$\frac{dx}{dt} \equiv \lim_{\Delta t \to 0} \frac{x(t+\Delta t) - x(t)}{\Delta t} \tag{B.1}$$

Consider a sequence of values $t_n = n\Delta t$, then for any value *i*, there is a value $x_i = x(t_i)$. Each x_i is a finite element of extent Δt of the function $x(t)$. The finite difference ratio

$$\frac{x_{i+1} - x_i}{t_{i+1} - t_i} = \frac{x(t_{i+1}) - x(t_i)}{t_{i+1} - t_i} = \frac{x(t_i + \Delta t) - x(t_i)}{\Delta t} \tag{B.2}$$

becomes the derivative when the increment Δt becomes infinitesimally small. For nonzero Δt, Eq. (B.2) is an approximation to the derivative. The smaller Δt is, the better the approximation. Furthermore, the mean value theorem states that at least one point t_C, such that $t_i < t_C < t_{i+1}$, has a derivative equal to the value of Eq.(B.2), for any Δt.

A derivative approximation

$$\frac{dx}{dt} \approx \frac{x_{i+1} - x_i}{t_{i+1} - t_i} \tag{B.3}$$

is the best approximation to the true dx/dt at $t = t_i + 0.5\Delta t$, not at $t = t_i$ or $t = t_{i+1}$. In using the finite difference approach to solving differential equations, all functions and derivatives must be evaluated at the same point. Thus it is more useful to define the first derivative as

$$\frac{dx}{dt} \approx \frac{x_{i+1} - x_{i-1}}{2\Delta t} \tag{B.4}$$

which is evaluated at t = t_i rather than halfway between t_i and t_{i+1}.

The second derivative may be defined in terms of a difference of first derivatives

$$\begin{aligned} \frac{d^2x}{dt^2} &\approx \frac{\left.\frac{dx}{dt}\right|_{t_i+0.5\Delta t} - \left.\frac{dx}{dt}\right|_{t_i-0.5\Delta t}}{\Delta t} \approx \frac{\frac{x_{i+1}-x_i}{\Delta t} - \frac{x_i - x_{i-1}}{\Delta t}}{\Delta t} \\ &\approx \frac{x_{i+1} - 2x_i + x_{i-1}}{(\Delta t)^2} \end{aligned} \tag{B.5}$$

Eq.(B.5) is properly evaluated at t = t_i.

Finite Difference Solution of Differential Equations

Now consider the following differential equation

$$A(x,t)\frac{d^2x}{dt^2}+B(x,t)\frac{dx}{dt}+C(x,t)=0 \tag{B.6}$$

The finite difference solution of Eq.(B.6) proceeds by substituting for the derivatives to obtain

$$A(x_i,t_i)\frac{x_{i+1}-2x_i+x_{i-1}}{(\Delta t)^2}+B(x_i,t_i)\frac{x_{i+1}-x_{i-1}}{\Delta t}+C(x_i,t_i)=0 \tag{B.7}$$

If x_i and x_{i-1} are known, then it becomes a simple matter to solve for the unknown x_{i+1}. When x_{i+1} is known, it becomes the new x_i; the old x_i becomes the new x_{i-1} and the process of solving for x_{i+1} repeats. This recursive process can be repeated indefinitely. Once the complete set of x_i is known for the set of t_i, the solution is in hand.

The problem with the recursive solution Eq.(B.7) is that x_{i-1} and x_i must be known to determine x_{i+1}. Fortunately, many interesting problems begin with known initial values

$$x(t_0)=D \tag{B.8}$$

and

$$\left.\frac{dx}{dt}\right|_{t_0}=E \tag{B.9}$$

If we shift the t-axis such that i=0 coincides with t = t_0, then

$$x(t_0)=D=x_0 \tag{B.10}$$

If Δt is small enough, the function x(t) can be approximated by straight line segment between any two neighboring points. In this limit the derivative can be assumed to be symmetric about its center point

$$\frac{dx}{dt}\approx\frac{x_{i+1}-x_{i-1}}{2\Delta t}\approx\frac{x_i-x_{i-1}}{\Delta t} \tag{B.11}$$

For i=0, we can solve for x_{-1}

$$x_{-1} = x_0 - \Delta t \left. \frac{dx}{dt} \right|_{t_0} = D - E\,\Delta t \tag{B.12}$$

With the results (B.10) and (B.12) the recursive solution process may begin.

Finite Element Computation of Definite Integrals

Problems with continuous variation of some parameter can often be solved by breaking problem into finite chunks of space and/or time, treating each **finite element** as a homogeneous entity, and evaluating the interactions between discrete elements to determine the behavior of the whole. As the number of elements gets larger (and their finite extent gets smaller), the discrete result approaches the continuous result. In its simplest form one may use this concept to compute definite integrals.

A definite integral (in one variable) can be considered to represent the area under a segment of a continuous curve. If the curve is broken into finite elements and the area under each element is estimated, then the total area is the sum of the areas under the finite elements. The simplest estimate of area under a segment of curve is value of the curve at the midpoint times the width of the segment. This is a rectangular approximation. Other more sophisticated estimates can be used. These are more computationally efficient, but in the limit (and assuming perfect computers) will not give better results. Since we are merely establishing a general technique, it is not only satisfactory but preferable to stick with the simplest approximations. In this case, it can be shown that the definite integral of a function can be computed using the expression

$$\begin{aligned}\int_a^b dx\ f(x) &\equiv \lim_{N\to\infty}\sum_{i=1}^{N}\left(\frac{b-a}{N}\right) f\left(a+\left(i-\tfrac{1}{2}\right)\left(\frac{b-a}{N}\right)\right) \\ &= \lim_{N\to\infty}\sum_{i=1}^{N}\Delta x\ f(x_i)\end{aligned} \qquad \text{(B.13)}$$

As N gets larger, Δx gets smaller, and the approximation to the true value of the definite integral gets more accurate. In the limit as $N \to\infty$, the identity will be valid. Eq. (B.13) is the finite element statement of the **rectangular rule** for numerical integration. It is readily generalized to multiple functions and multiple dimensions if necessary.

References

[1] G. Reece, Microcomputer Modeling by Finite Differences (Macmillan, London, UK, 1986).

[2] S. E. Koonin, Computational Physics (Addison-Wesley Publishing Co., Reading, MA, 1986).

[3] A. Constantinides, Applied Numerical Methods with Personal Computers (McGraw-Hill Book Co., New York, NY, 1987).

[4] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, Numerical Recipes: The Art of Scientific Computing, Revised Edition (Cambridge University Press, Cambridge, UK, 1989).

[5] W. T. Vetterling, S. A. Teukolsky, W. H. Press, and B. P. Flannery, Numerical Recipes Example Book (FORTRAN) (Cambridge University Press, Cambridge, UK, 1985).

[6] G. B. Thomas, Jr. and R. L. Finney, Calculus and Analytic Geometry, 7th Ed. (Addison-Wesley Publishing Co., Reading, MA, 1988).

APPENDIX C

PROBABILITY AND STATISTICS

Probability and statistics are tools that are essential in the analysis of combat systems. In this appendix we present a brief review of the critical aspects of these tools. We begin with combinatorial analysis – the ways in which groups of objects can be ordered and arranged.

Combinatorial analysis

Combinatorial analysis involves the different possibilities of arranging objects. The objects to be arranged are called **elements** and the combination of these elements including their order is called an **arrangement**. Every possible arrangement of a given set of elements in any order is called a **permutation** of those elements. For example, given elements *a*, *b*, and *c* then the possible permutations are:

abc, *cab*, *bca*, *acb*, *bac*, *cba*

There are *n!* permutations of distinct elements. If some elements (*j*, *k*, ..., and *m*) are identical, and the numbers of the identical elements are p_j, p_k, ..., and p_m, then the number of permutations is given by

$$N = \frac{n!}{p_j!\,p_k!\cdots p_m!}\,. \tag{C.1}$$

Two elements in an arrangement are said to form an **inversion** if their ordering is opposite to their natural (or original?) order. For example, we think of the order *a b c d e* as natural. If we have an arrangement *b a c e d*, then the pairs of elements *b* and *a* and *e* and *d* form inversions. A permutation is said to be an **even permutation** if the number of inversions is even, and an **odd permutation** if the number of inversions is odd.

The number of permutations of *k* different elements selected from *n* possible elements without repetition of any selected element is given by

$$^{n}P_{k} = \frac{n!}{(n-k)!}\,. \tag{C.2}$$

The number of permutations of *k* elements selected from *n* possible elements with any degree of repetition allowed is given by

$$ {}^{n}P_{k}^{(r)} = n^{k} \tag{C.3} $$

In a **combination** the order in which the elements occur is not taken into account. Thus, *ab* and *ba* represent permutations of the combination of *a* and *b*. The number of combinations of *k* elements taken from *n* elements without repetition is

$$ {}^{n}C_{k} = \binom{n}{k} = \frac{n!}{(n-k)!\,k!}\,. \tag{C.4} $$

The numbers represented by the ${}^{n}C_{k}$ are sometimes called the binomial coefficients because the binomial expansion can be written as

$$ \begin{aligned} (x+y)^{n} &= x^{n} + \frac{n!}{(n-1)!\,1!}x^{n-1}y + \cdots + \frac{n!}{(n-m)!\,m!}x^{n-m}y^{m} + \cdots \\ &\quad + \frac{n!}{1!(n-1)!}xy^{n-1} + y^{n} \\ &= \binom{n}{0}x^{n} + \binom{n}{1}x^{n-1}y + \cdots + \binom{n}{m}x^{n-m}y^{m} + \cdots + \binom{n}{n}y^{n} \\ &= \sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^{k} \end{aligned} \tag{C.5} $$

The number of combinations of *k* elements taken from *n* possible elements with any degree of repetition is given by

$$ {}^{n}C_{k}^{(r)} = \binom{n+k-1}{k} \tag{C.6} $$

Probability

Classical probability theory concerns itself with trials and events. A **trial** is an observation or experiment that is characterized by a set of conditions to be satisfied and by being repeatable. An **event** is the result of a trial that can occur but does not necessarily occur. A **certain event** *S* is one which always occurs. An **impossible event** ∅ is one which never occurs. Any trial may yield a number of events simultaneously. For example, if the outcomes of a trial are integers, then each integer is an event, odd integers and even integers are events, prime integers are events, etc. Obviously, if the outcome of the trial is 7, then odd integer, and prime integer are also outcomes, and therefore events. Events are said to be **mutually exclusive**, if as the result of a trial only one of them can occur. In the integer example, odd integer and even integer are mutually exclusive events. An event is the **sum** *C* **of multiple events** *A, B, ...* if in any trial any of the events *A, B,* or ... occur. The sum is denoted by

$$C = A \cup B \cup \cdots \tag{C.7}$$

An event is **product** *C* **of multiple events** *A, B, ...* if in any trial all of the events *A, B,* and ... occur. The product is denoted by

$$C = A \cap B \cap \cdots \tag{C.8}$$

The classical definition of probability is: "*The probability of an event A equals the ratio of the favorable outcomes to the total number of outcomes, provided they are equally likely.*" Thus, if a **trial** can result in *n* equally likely **events** and if *m* of these are favorable to the occurrence of an event *E*, then the probability *P(E)* of the event *E* occurring is

$$P(E) = \frac{m}{n} = \frac{\text{number of favorable events}}{\text{number of possible events}} \tag{C.9}$$

The probability of the sum of a number of mutually exclusive events (i.e., the probability that any of the events occurs) is equal to the sum of the probabilities of these events.

$$P(E_1 + E_2 + \cdots + E_k) = P(E_1) + P(E_2) + \cdots + P(E_k) \tag{C.10}$$

An **unconditional probability** depends only on the set of conditions to be satisfied by the trial. A **conditional probability** depends on at least one further condition. This additional condition is often the prior occurrence of some other event. The conditional probability that an event *E* occurs given that the event *F* has already occurred is denoted by

$$P(E|F) = \text{Probability}(E \text{ occurs given } F \text{ has occurred}) \tag{C.11}$$

The probability *P(EF)* for the simultaneous occurrence of two events *E* and *F* is the product of the probability *P(E)* of the first event *E* and the conditional probability *P(F|E)* that the second event *F* will occur given that *E* has already occurred.

$$P(EF) = P(E) \cdot P(F|E) \tag{C.12}$$

If a set of events F_i, where *i* = 1, 2, ..., n, forms a **complete system**, then one and only one of these events must always result in any trial. If a set of events F_i forms a complete system and *E* is a further event, then the event $E \bullet F_i$ for any *i* is mutually exclusive with any other event $E \bullet F_j$ where *j* differs from *i*. Thus, the **total probability** (unconditional) of the event *E* is given by

$$P(E) = \sum_{i=1}^{n} P(EF_i) = \sum_{i=1}^{n} P(F_i) P(E|F_i) \tag{C.13}$$

Two events are **independent** of each other, if the occurrence (or non-occurrence) of one event does not affect the occurrence or non-occurrence of the second event. That is,

$$P(E|F) = P(E) \quad \text{and} \quad P(F|E) = P(F). \tag{C.14}$$

If a set of events E_i where *i* = 1, ..., *n* are all independent, then the probability that all events of the set will occur in a trial is given by

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1) \cdot P(E_2) \cdots P(E_n) = \prod_{i=1}^{n} P(E_i) \tag{C.15}$$

A useful result in probability theory is **Bayes' theorem**. This theorem relates the conditional probability of one event occurring if a second event occurs to the inverse conditional probability (the probability of the "second" event occurring given that the "first" event occurs). That is,

$$P(F_i|E) = \frac{P(E|F_i) P(F_i)}{\sum_{i=1}^{n} P(E|F_i) P(F_i)} = \frac{P(E|F_i) P(F_i)}{P(E)} \tag{C.16}$$

Another useful result is the probability that an event occurs at least once in *N* independent trials. If *P*(1) is the probability that the event occurs in one trial, then

$$P_M(1) = 1 - P(1) \tag{C.17}$$

is the probability that the event does not occur, i.e., the probability of a "miss". The probability of *N* consecutive "misses" is

$$P_M(N) = \left[P_M(1)\right]^N = \left[1 - P(1)\right]^N \tag{C.18}$$

The probability that at least one "hit" occurs is one minus the probability of getting all "misses". Thus, the probability of at least one "hit" in N trials is given by

$$P(N) = 1 - P_M(N) = 1 - \left[1 - P(1)\right]^N \ . \tag{C.19}$$

Random Variables and Functions of Random Variables

A **random variable** *X* is a variable which, in repeated trials, assumes different values *x*, each of which represents a **random event**. A event can be considered random if its probability of occurrence is not conditional on the outcome of any earlier trial. Random variables are **discrete** if *X* can assume only a finite number (or at most a countably infinite number) of values. A countably infinite quantity is one in which given any finite length interval of the variable's range the number of possible values is finite. For example the set of integers is countably infinite (in any finite range of numbers from y_1 to y_2 there is a finite and countable number of integers). Random variable are **continuous** if *X* can assume all possible values in a finite or infinite interval.

For discrete random variables, the random events x_i are treated as **discontinuities**. Only the finite values x_i have meaning. The probability of each random event $P(X = x_i)$ is treated as the magnitude of the discontinuity. From a functional perspective we can assign a delta function to each value x_i and a finite probability $P(x_i)$ to each delta function. That is,

$$P(x) = P(x_i)\delta(x_i) \tag{C.20}$$

The sum over all possible values of *x* must of course equal unity.

$$\int_{-\infty}^{\infty} dx\ P(x) = \sum_{i=1}^{n} P(x_i)\delta(x_i) = \sum_{\text{all } i} P(X = x_i) = 1 \tag{C.21}$$

For continuous random variables, the probability of getting precisely a value *x* is zero, because there are an infinite number of precise values that could be selected. From our earlier result, if there is one event which gives the precise value *x* and an infinite number of events that give values other than *x* then the probability will be 1/∞ or zero. We must therefore speak of the probability that the value of *X* lies within a small range about *x*. We call the function that describes this probability a **probability density function** *f(x)*. More specifically we can ascribe the probability that *X* takes on a value between *x - 0.5dx* and *x + 0.5dx*, where *dx* is a very small interval, to the function *f(x)*

$$P(x - 0.5dx \le X \le x + 0.5dx) = f(x)dx \tag{C.22}$$

Note that a more rigorous mathematical treatment of probability might replace the second ≤ sign with a < sign. However, since $P(x) = 0$ for any exact value, there is no error in extending the expression to include both ends of the interval.

It is common to define *f(x) dx* using a unit interval. In this case the *dx* is usually dropped. The probability density function *f(x)* then has a value of probability per unit interval (of *X*). When multiplied by an interval, the probability density function yields the probability that the function takes a value within that interval. The definite integral of the probability density function over a larger interval yields the probability that the variable lies within that larger interval, i.e.,

$$\int_a^b dx\ f(x) = \text{Probability that a} \geq x \geq \text{b} \tag{C.23}$$

Since *f(x)* represents a probability per unit interval, if *f(x)* is integrated over the entire range of possible values, the total probability must be unity.

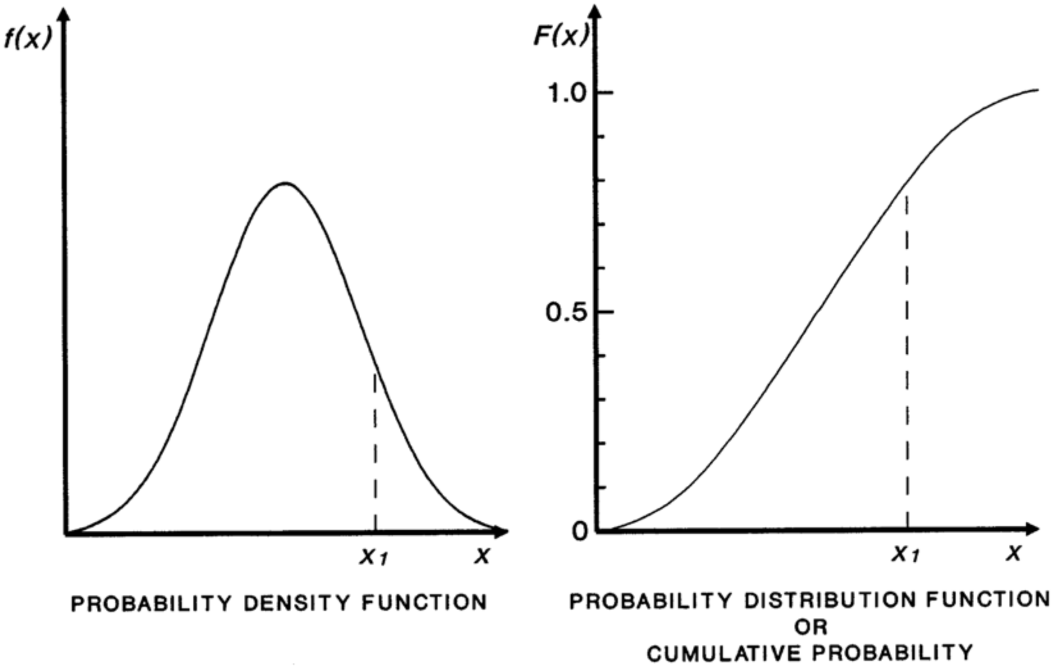
$$\int_{-\infty}^{\infty} dx\ f(x) = 1 \tag{C.24}$$

A second useful function is the definite integral of *f(x)* from -∞ to some finite value *x*. This integral denoted by *F(x)* in many texts is known as the **distribution function** or the **cumulative probability function**.

$$F(x) \equiv \int_{-\infty}^{x} dt\ f(t) = \text{Probability that } X \leq x \tag{C.25}$$

That is, *F(x)* is the probability that the random variable *X* takes on some value that is less than or equal to *x*. The relationship between *f(x)* and *F(x)* is shown in Figure C-1.

Figure C-1. Relationship between a probability density function and the corresponding distribution function.



Combining Eqs. (C.24) and (C.25), we find

$$F(\infty) = \int_{-\infty}^{\infty} dt\ f(t) = \text{Probability that } X \leq \infty = 1 \tag{C.26}$$

That is, the probability of getting any possible value must be unity. As a consequence, we can quickly determine that

$$1 - F(x) = \int_{x}^{\infty} dt\ f(t) = \text{Probability that } X \geq x\,. \tag{C.27}$$

Since $F(a)$ is the probability that X is less than or equal to a and $F(b)$ is the probability that X is less than or equal to b, then we also find

$$F(b) - F(a) = \int_{a}^{b} dx\ f(x) = \text{Probability that } a \leq X \leq b\,. \tag{C.28}$$

The probability density function is sometimes written in the form $f_x(x)$ where the subscript denotes either the type of distribution or the random variable to which that distribution refers. The letters p and P are also sometimes used instead of f and F when describing density and distribution functions.

Certain properties of density and distribution functions occasionally prove useful. From Eq.(C.25) we see that the distribution function can be obtained by integrating the density function. Conversely, the density function can be obtained by differentiating the distribution function.

$$f(x) = \frac{dF(x)}{dx} \tag{C.29}$$

The distribution function must have limits of

$$F(-\infty) = 0 \qquad \text{and} \qquad F(\infty) = 1 \tag{C.30}$$

The distribution function must also be a monotonic increasing function of increasing x

$$F(x_2) \geq F(x_1) \qquad \text{for} \qquad x_2 > x_1 \tag{C.31}$$

If y is a function $g(x)$ where x is a random variable with density function $f_x(x)$, then the density function of y is $f_y(y)$ and is given by the expression

$$f_y(y) = \frac{f_x(x_1)}{\left|dg(x_1)/dx\right|} + \frac{f_x(x_2)}{\left|dg(x_2)/dx\right|} + \cdots + \frac{f_x(x_n)}{\left|dg(x_n)/dx\right|} \tag{C.32}$$

where the x_i are the roots of the equation $y = g(x) = g(x_i)$.

A cumulative probability curve is a plot of the "*Probability that Variable X has a value less than or equal to x*" as a function of "*x*", that is, a plot of the probability distribution function. When *X* is at its minimum value,

$$P(X_{min}) = F(X_{min}) = 0. \tag{C.33}$$

When X is at its maximum value,

$$P(X_{max}) = F(X_{max}) = 1. \tag{C.34}$$

When X is at its median value,

$$P(X_{median}) = F(X_{median}) = 0.5. \tag{C.35}$$

In practice there are a number of probability density and probability distribution functions that are commonly encountered. In Table C-1 we list a number of useful density functions and their corresponding distribution functions (when these are calculable – many density functions do not have simple corresponding distribution functions). A number of special functions are used in Table C-1. These are defined below:

Error Function
$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x dy\, e^{-y^2} \tag{C.36}$$

Heaviside Step Function
$$U(x;a) = \begin{cases} 0 & x < a \\ 1 & x \geq a \end{cases} \tag{C.37}$$

Gamma Function
$$\Gamma(x) = \int_0^\infty dy\, y^{x-1} e^{-y} \tag{C.38}$$

Modified Bessel Function
$$I_0(x) = 1 + \frac{x^2}{2^2} + \frac{x^4}{2^2 \cdot 4^2} + \frac{x^6}{2^2 \cdot 4^2 \cdot 6^2} + \cdots \tag{C.39}$$

Delta Function
There are several definitions of the delta function. These include:

Table C-1. Common probability functions.

DENSITY FUNCTION	DISTRIBUTION FUNCTION	
Normal (or Gaussian) with mean = μ and variance = σ^2		(C.40)
$f_N(x) = \left(2\pi\sigma^2\right)^{-1/2}\ e^{-(x-\mu)^2/2\sigma^2}$	$F_N(x) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)$	
Dirac		(C.41)
$f_{DIRAC}(x) = \delta(x;\, x_{AV})$	$F_{DIRAC}(x) = U(x;\, x_{AV})$	
Uniform (Rectangular pulse) with mean = x_0 + *0.5* Δ*x* and variance = Δ*x/12*		(C.42)
$f_U(x;\, x_0, \Delta x) = \begin{cases} 0 & x < x_0;\ x > x_0 + \Delta x \\ \dfrac{1}{\Delta x} & x_0 \le x \le x_0 + \Delta x \end{cases}$	$F_U(x) = \begin{cases} 0 & x < x_0 \\ \dfrac{x - x_0}{\Delta x} & x_0 \le x \le x_0 + \Delta x \\ 1 & x > x_0 + \Delta x \end{cases}$	
Binomial (n+1 points, p+q=1) with mean = *np* and variance = *np(1-p)*		(C.43)
$f_B(x) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \delta(x-k)$	$P_B(x = k) = \binom{n}{k} p^k q^{n-k}$	
Poisson with mean = *a* and variance = *a*		(C.44)
$f_{POISSON}(x) =$ Not evaluated.	$P_P(x = k) = \frac{a^k e^{-a}}{k!}$	
Gamma -- *f(x)* is maximum for *x=b/c*		(C.45)
$f_\Gamma(x) = \frac{c^{b+1}}{\Gamma(b+1)} x^b e^{-cx} U(x;\, 0)$	$F_\Gamma(x) =$ Not evaluated.	
Rayleigh Amplitude		(C.46)
$f_R(x) = \frac{x}{a^2} e^{-x^2/2a^2} U(x;\, 0)$	$F_R(x) =$ Not evaluated.	
Rayleigh Power with mean = *a* and variance = *a*^{2}		(C.47)
$f_{RAYLEIGH}(x) = \frac{1}{a} e^{-x/a} U(x;\, 0)$	$F_{RAYLEIGH}(x) = \left(1 - e^{-x/a}\right) U(x;\, 0)$	

Table C-1 (continued)**.** Common probability functions.

DENSITY FUNCTION DISTRIBUTION FUNCTION

Rice with mean = μ and m^2 = ratio of constant power to average of random power (C.48)

$$f_{RICE}(x) = \frac{1+m^2}{\mu} e^{-m^2-\left(1+m^2\right)\left(\frac{x}{\mu}\right)} I_0\left(2m\sqrt{1+m^2\left(\frac{x}{\mu}\right)}\right)$$

Rice (m=1) (C.49)

$$f_{RICE}(x) = \frac{4x}{a^2} e^{-2x/a} U(x;0) \qquad F_{RICE}(x) = \text{ Not evaluated.}$$

Beta -- *f(x)* is maximum for *x=b/(b+c)* (C.50)

$$f_B(x) = \frac{\Gamma(b+c+2)}{\Gamma(b+1)\Gamma(c+1)} x^b (1-x)^c \quad 0 \le x \le 1 \qquad F_B(x) = \text{ Not evaluated.}$$

Cauchy (C.51)

$$f_{CAUCHY}(x) = \frac{a/\pi}{a^2+x^2} \qquad F_{CAUCHY}(x) = \text{ Not evaluated.}$$

Laplace (C.52)

$$f_{LAPLACE}(x) = \frac{a}{2} e^{-a|x|} \qquad F_{LAPLACE}(x) = \text{ Not evaluated.}$$

Maxwell (C.53)

$$f_{MAXWELL}(x) = \frac{\sqrt{2}}{a^3\sqrt{\pi}} x^2 e^{-x^2/2a^2} U(x;0) \qquad F_{MAXWELL}(x) = \text{ Not evaluated.}$$

Weibull with median value = x_m (C.54)

$$f_W\left(\frac{x}{x_m}\right) = b\ln 2\left(\frac{x}{x_m}\right)^{b-1} e^{-(\ln 2)(x/x_m)^b} U(x;0) \quad F_W\left(\frac{x}{x_m}\right) = 1 - e^{-(\ln 2)(x/x_m)^b}$$

Log-normal with mean (of *ln* x) = μ and variance (of *ln* x) = S^2 (C.55)

$$f_{LN}(x) = \frac{1}{xS\sqrt{2\pi}} e^{-(\ln x-\mu)^2/2S^2} \qquad F_{LN}(x) = \text{ Not evaluated.}$$

$$\delta(x;\mu) = \lim_{\sigma\to 0}\left[\left(2\pi\sigma^2\right)^{-1/2}\; e^{-(x-\mu)^2/2\sigma^2}\right]$$
$$= \begin{cases} 0 & x \neq \mu \\ \infty & x = \mu \end{cases} \tag{C.56}$$

or

$$\delta(x;\mu) = \lim_{\Delta x\to 0}\left[f_U(x;\mu,\Delta x) = \begin{cases} 0 & x<\mu;\; x>\mu+\Delta x \\ \dfrac{1}{\Delta x} & \mu \le x \le \mu+\Delta x \end{cases}\right]$$
$$= \begin{cases} 0 & x \neq \mu \\ \infty & x = \mu \end{cases} \tag{C.57}$$

The integral of any delta function over all values of *x* is 1, showing that the delta function is a valid probability density function.

A common problem encountered in military applications is the dispersion of "shots" about an aimpoint, or an analog. Let the aimpoint be the origin of coordinates. The probability that a shot will hit the target at a lateral position $r = (x^2 + y^2)^{1/2}$ is given by a Gaussian probability density function

$$p(r) = p(x,y) = \frac{1}{2\pi\sigma^2} e^{-\left((x-\eta)^2+(y-\xi)^2\right)/2\sigma^2} \tag{C.58}$$

where $\rho = (\eta, \xi)$ is the **bias** of the mean and σ is the **dispersion** (i.e., the standard deviation) about the mean. The normalization to this distribution is $(2\pi\sigma^2)^{-1}$ as compared to the normalization $(2\pi\sigma^2)^{-1/2}$ as shown in Eq.(C.40). This results because we have a two-dimensional distribution (*x, y*) rather a one-dimensional distribution (*x*). Let us set the bias to zero. Now the probability distribution is

$$p(r) = p(x,y) = \frac{1}{2\pi\sigma^2} e^{-\left(x^2+y^2\right)/2\sigma^2} = \frac{1}{2\pi\sigma^2} e^{-r^2/2\sigma^2} \tag{C.59}$$

The probability ($P(r\le a)$) that an individual shot lies within radius *a* of the aimpoint can be determined from the integral

$$P\left(r \leq a\right) = \int_0^a dr\; 2\pi r\; p\left(r\right) = \frac{1}{2\pi\sigma^2}\int_0^a dr\; 2\pi r\; e^{-r^2/2\sigma^2}$$

$$= 1 - e^{-a^2/2\sigma^2} \tag{C.60}$$

The **circular error probable** is the radius obtained using Eq.(C.58) for a probability of 50%. That is,

$$P\left(r \leq CEP\right) = 0.5 = 1 - e^{-CEP^2/2\sigma^2} \tag{C.61}$$

or

$$CEP = \left(2\ln 2\right)^{1/2}\sigma = 1.17741\sigma\,. \tag{C.62}$$

Statistics

Statistics are used to describe random functions when the probability density function is unknown or only approximately known. One of the most useful statistical quantities is the mean or the expected value of the function. The expected value of a random variable *X* can be determined from the probability density function by

$$E(x) = \int_{-\infty}^{\infty} dx \; x \; f(x) \, . \tag{C.63}$$

for a continuous random variable. If the random variable is discrete, the expected value is

$$E(x) = \sum_{\text{all } x_i} x_i f(x_i) = \sum_{\text{all } x_i} x_i p_i \tag{C.64}$$

where p_i is the probability that the value x_i would be obtained.

The expected value of any function *g(x)* of that random variable can similarly be found from

$$E(g(x)) = \int_{-\infty}^{\infty} dx \; g(x) \; f(x) \tag{C.65}$$

The expected value is the average value that would be obtained from a large number of measurements. As such it can be determined from the probability density function or it can be found by averaging a large number (*N*) of random measurements x_j.

$$E(x) = \frac{1}{N} \sum_{j=1}^{N} x_j \tag{C.66}$$

The mean does little to describe the variability of the individual measurements. More general statistical measures are the moments of the distribution. Moments may be determined about the origin or about the mean. The mth moment about the origin of the distribution *f(x)* is defined by the relations

$$\nu_m = E(x^m) = \sum_{\text{all } x_i} x_i^{\;m} f(x_i) \tag{C.67}$$

for discrete random variables and

$$\nu_m = E(x^m) = \int_{-\infty}^{\infty} dx \; x^m \; f(x) \tag{C.68}$$

for continuous random processes. The mean μ is given by the first moment about the origin.

$$\mu = E(x) = \nu_1 \tag{C.69}$$

The m^th^ moment about the mean is defined by the relations

$$\mu_m = E\left((x-\mu)^m\right) = \sum_{\text{all } x_i} (x_i - \mu)^m f(x_i) \tag{C.70}$$

for discrete random functions and

$$\mu_m = E\left((x-\mu)^m\right) = \int_{-\infty}^{\infty} dx\, (x-\mu)^m \; f(x) \tag{C.71}$$

for continuous random processes. Obviously $\mu_1 = 0$.

The following quantities computed from the various moments are usually used to describe the statistics of a quantity. These are

Mean $= \mu = \nu_1$ (C.72)

Variance (= Var(x)) $= \sigma^2 = \mu_2 = \nu_2 - \nu_1^2$ (C.73)

Standard Deviation $= \sigma = \mu_2^{1/2}$ (C.74)

Skewness $= \alpha_3/2 = \mu_3/2\sigma^3$ (C.75)

Kurtosis $= (\alpha_4 - 3)/2 = ((\mu_4/\mu_2^2) - 3)/2$ (C.76)

The mean describes the “center” of the distribution. The variance describes its “width. The skewness describes the asymmetry in the distribution. The kurtosis is essentially a measure of the flatness of the peak and the verticality of the sides of the distribution (usually referenced to a Gaussian distribution).

The mean of a collection of measurements is given by Eq. (C.64). The variance of that same collection of measurements is given by

$$\mathrm{Var}(x) = \frac{1}{N}\left[\sum_{j=1}^{N} x_j^2 - \frac{1}{N}\left[\sum_{j=1}^{N} x_j\right]^2\right]. \tag{C.77}$$

The **median** value of a distribution or of a collection of measurements is that value above which one finds half of the values and below which one finds the other half. The median value x_{median} of a distribution is that value of x for which

$$F(x_{median}) = 0.5 \tag{C.78}$$

or

$$\int_{-\infty}^{x_{median}} dx\ f(x) = \int_{x_{median}}^{\infty} dx\ f(x) \tag{C.79}$$

If the N elements of a collection of measurements are ordered by increasing value, i.e., $x_m < x_{m+1}$ for all m, then the median value is $x_{N/2}$. The **mode** of a distribution or collection of measurements is that value of x for which the density function is a maximum, or the value most commonly encountered. The mode is the value that is most likely to be picked in a single trial.

Variance and Covariance of Vector Random Functions

In the preceding section we described a number of statistics of simple random variables. Many problems involve more than a single random variable. For example, position involves three variables (x, y, and z) each of which may be random and may or may not vary independently.

The variance of a random variable *X* has been previously identified as being given by the relation

$$\mathrm{Var}(X) = \mathrm{E}\left\{(X-\mathrm{E}(X))^2\right\} = \mathrm{E}\left\{(X-\mathrm{E}(X))(X-\mathrm{E}(X))\right\}. \tag{C.80}$$

The covariance of two jointly distributed random variables is a generalization of the variance and is defined by

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left\{(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right\}. \tag{C.81}$$

The covariance basically describes how one random variable varies with respect to the other random variable. Note that

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X). \tag{C.82}$$

It is often of interest to know the degree to which variations in one random variable track variations in another random variable. The correlation coefficient *r* provides such a description. The correlation coefficient is defined by

$$r = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}. \tag{C.83}$$

If two random variables move in perfect synchronization with each other, the correlation coefficient approaches unity, $r = 1$, and the variables are said to be correlated. If the changes in one variable show absolutely no definable relationship to the changes in the second variable, then the correlation coefficient approaches zero, $r = 0$, and the variables are said to be uncorrelated. If one variable changes exactly the opposite to changes in the second variable, then the correlation coefficient approaches minus one, $r = -1$, and the variables are said to be anti-correlated. Two random variables which have variations with limited relationship to each other, that is, $-1 < r < 0$ and $0 < r < 1$, then the variables are partially correlated. The closer the correlation coefficient approaches +1 (-1), the more highly correlated (anti-correlated) are the two random variables.

If the random quantity is a vector, then we frequently find that we need information on how one component of that vector varies relative to other components. We summarize all of this information in the expectation value and covariance of that vector. For a random vector

$$\vec{X} = \{X_1, X_2, \cdots, X_N\}, \tag{C.84}$$

the expectation value is just the vector composed of the expectation value of each component

$$\mathrm{E}\{\vec{X}\} = \{\mathrm{E}(X_1), \mathrm{E}(X_2), \cdots, \mathrm{E}(X_N)\}. \tag{C.85}$$

The covariance of the random vector is an *N* x *N* matrix defined by

$$\begin{aligned}\sum_{i,j} &= \mathrm{Cov}(X_i, X_j) \\ &= \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_N) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_1, X_N) & \mathrm{Cov}(X_2, X_N) & \cdots & \mathrm{Var}(X_N) \end{bmatrix}\end{aligned} \tag{C.86}$$

References

[1] Anonymous, "Probability Theory and Statistics" in The VNR Concise Encyclopedia of Mathematics, W. Gellert, H. Kustner, M. Hellwich, & H. Kastner, Eds. (Van Nostrand Reinhold, New York NY, 1977) pp. 575-607.

[2] Papoulis, Athanasios, Probability, Random Variables, and Stochastic Processes, (McGraw-Hill Book Co., New York NY, 1965).

[3] Selby, Samuel M. and Girling, Brian (Eds.), Standard Mathematical Tables 14th Edition (Chemical Rubber Co., Cleveland OH, 1965).

INDEX