1989

Line simplification and digital cartographic databases

Hoyt Walker
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

Walker, Hoyt, "Line simplification and digital cartographic databases" (1989). *Master's Theses*. 3234.
DOI: <https://doi.org/10.31979/etd.vpw4-acfs>
https://scholarworks.sjsu.edu/etd_theses/3234

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 1339660

Line simplification and digital cartographic databases

Walker, Hoyt, M.A.

San Jose State University, 1989

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

Line Simplification and Digital Cartographic Databases

A Thesis

Presented to

The Faculty of the Department of Geography

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

By

Hoyt Walker

November, 1989

APPROVED FOR THE DEPARTMENT OF GEOGRAPHY

Dr. David Schwarz

Dr. Duilio Peruzzi

Leonard Gaydos

APPROVED FOR THE UNIVERSITY

Abstract

Line Simplification and Digital Cartographic Databases

by Hoyt Walker

This thesis examines one component of the cartographic generalization process, line simplification, and its implementation with computer technology. The effectiveness of line simplification techniques in producing useful maps from digital map data is studied. Many techniques have been proposed to simplify lines; several of these have been implemented here. The approaches have been compared on their ability to simplify lines from two cartographic databases over both modest and extreme changes in scale. The comparisons are based on the simplification of individual lines, on the production of complete maps, and on mathematical measures of performance. One promising technique is enhanced to overcome certain limitations which are apparent during extreme scale change. The effect of map projections on the performance of such techniques is considered. While much additional work is needed, this study indicates that line simplification techniques can extend the range of scales over which certain cartographic databases can be utilized.

Table of Contents

Chapter I — Introduction	1
Chapter II — Cartographic Generalization	6
Generalization and Information Theory	7
Limitations of Generalized Communication	13
Traditional Cartographic Generalization	15
<i>Selection</i>	15
<i>Simplification</i>	18
<i>Classification</i>	21
<i>Symbolization</i>	24
<i>Induction</i>	27
Difficulties in Traditional Generalization	28
Chapter III — Computer Use in Cartography	29
The Promise of Computerization	30
Current Applications	34
Limitations of Automation	37
Chapter IV — Cartographic Databases	43
Digital Map Representation	43
<i>Geometry</i>	44
<i>Attributes</i>	45
<i>Topology</i>	46
Cartographic Data Structures	48
<i>Vector Data Structures</i>	48
<i>Raster Data Structures</i>	55
<i>Comparison of Raster and Vector Structures</i>	58
<i>Overlays and Tiles</i>	59
Data Capture	60
<i>Tablet Digitizing</i>	61
<i>Raster Scan Digitizing</i>	63
Design Considerations for Cartographic Databases	65
<i>Matching Structure to Function</i>	65
<i>Resolution and Size</i>	68
Chapter V — Line Simplification Algorithms	74
Classification of Algorithms	74
Description of Algorithms	77
<i>Point Selection</i>	77
<i>Low Pass Filtering</i>	78
<i>Angle Detection</i>	80
<i>Tolerance Bands</i>	84

<i>Point Relaxation</i>	87
<i>Fractal Techniques</i>	91
Algorithm Summary	99
Chapter VI — Methodology	102
Cartographic Databases Utilized	102
<i>NCARCDB</i>	103
<i>WDB II</i>	103
Common Data Format	104
Evaluation of Simplification Algorithms	109
<i>Subjective Evaluation</i>	109
<i>Mathematical Evaluation</i>	111
Chapter VII — Results	118
Database Formats	118
Normal and Tangential Tolerances	121
Qualitative Characteristics with Limited Scale Change	123
<i>Individual Chain Evaluation</i>	123
<i>Global Evaluation</i>	136
Qualitative Characteristics with Extreme Scale Change	146
<i>Individual Chain Evaluation</i>	146
<i>Global Evaluation</i>	151
Quantitative Evaluation	165
Chapter VIII — Technique Enhancement and	
Map Projection Effects	168
Enhancement of the Banding Technique	169
Map Projection Considerations	184
Chapter IX — Conclusions	189
References	199
Appendix A — Coordinate Systems and Vectors	208
Appendix B — Computer Storage Terminology	215
Appendix C — Implementation Details	219

Acknowledgments

Numerous people are directly and indirectly involved in the completion of a work of this size and several of these people need to be recognized. First, I have had the pleasure of having Dr. David Schwarz as my thesis advisor. His encouragement and suggestions have been greatly appreciated throughout the process. I must also thank the other members of my committee, Dr. Duilio Peruzzi and Leonard Gaydos, for their helpful comments. The input from all the members of the committee has greatly improved the quality of the completed thesis. In addition, I would like to thank the entire faculty of the Geography Department at San Jose State University for an enlightening and inspiring introduction to the formal study of geography.

Several of my colleagues at Lawrence Livermore National Laboratory have also played a role in this research. First, I must thank Daniel Rodriguez for his usual fine job of editing early drafts of several of the chapters in this thesis. His comments greatly accelerated the process of producing a finished document. I must thank my supervisor in Scientific Software Division, Sandra Taylor, for her support of my efforts. Especially, I must thank Dr. Thomas Sullivan, both personally and in his role as head of the Atmospheric Release Advisory Capability, for his encouragement and for allowing the use of the various facilities of the project for this work. As institutions, I must express appreciation for Scientific Software Division, the Atmospheric and Geophysical Sciences Division, and Lawrence Livermore National Laboratory as a whole for providing an environment in which I could discover and pursue a fascinating new career direction.

Finally, I would like to thank my good friend, Barbara Carlson, for her interest, support, and understanding during my work at San Jose State University.

Chapter I

Introduction

Maps play an important role in many fields of endeavor. They are an indispensable tool for dealing with questions of location and spatial variation. Any object that is on or near the surface of the earth can be mapped to show where it is with respect to other objects also on the earth. In addition, any phenomenon that varies from place to place on the earth can be portrayed on a map to enhance the understanding of that variation. Cartography, the art and science of making maps, is most closely allied with geography which incorporates the study of spatial variation and its underlying causes. The application of computers to cartography is a relatively recent development which has had a profound influence throughout the field. Computers have an obvious utility in completing some of the more mundane tasks in map-making, such as the calculation of map projection coordinates. However, the influence of the computer is not limited to the reproduction of tedious manual operations at a higher speed. As more sophisticated research and development occurs in computer cartography, the fundamental problems that face traditional cartography must be readdressed in light of the strengths and weaknesses of this new tool, the computer.

The issue of cartographic generalization is one that is central to traditional cartography. No map, even a hypothetical map at a scale of 1:1, can show all the detail that exists on the surface of the earth within the mapped area. As a result, the cartographer must select and present the information in such a way that the

message of the map can be accurately interpreted by the map reader. Cartographic generalization refers to the processes of selection, classification, simplification, symbolization, and induction which the cartographer performs in making a meaningful map. These processes are all functions of the scale of the map. For example, a cartographer may select all streets to be shown on a city map but might only select interstate highways for a map of the entire United States. Traditionally, this process of scale-dependent generalization has been much of the art which has existed in map-making. The individual cartographer has relied on personal experience, knowledge of basic geographical processes, and general cartographic principles to produce a map which reflects his or her own view of the information being portrayed. As a result, two different cartographers, or one cartographer at different times, may produce different generalizations for a specific map based on identical information. The principles that have developed for traditional generalization have assumed the intelligence and experience of a human cartographer. The application of computers to the generalization process requires a reevaluation of previous cartographic principles in light of the fact that computers are devices that can be easily programmed to do mechanical tasks but are very difficult to program to perform actions associated with human intelligence such as pattern recognition. While this reexamination is motivated by the need to adapt to the limitations of a new tool, the results may benefit all of cartography by forcing a clarification of the principles of generalization and a more fundamental understanding of the process.

Cartographic generalization is a very complex issue. To study a complicated process, it is often helpful to break the process down and study its more basic components. Of the five processes comprising cartographic generalization mentioned above, only one, simplification, will be considered in significant detail here. Such

study can also be aided by classifying the objects of the process (maps) and by examining a particular subset of these objects. A common classification of maps distinguishes between *base maps* and *thematic maps*. Base maps are usually designed to answer a wide variety of questions dealing with location; United States Geological Survey (USGS) 1:24,000 topographic maps are good examples of complete and comprehensive base maps. Thematic maps are meant to answer questions about a specific topic or theme. A thematic map often incorporates a base map to provide a background which orients the map reader, thereby allowing any spatial variation portrayed to be interpreted easily. This study will largely confine itself to problems associated with base maps.

Computer mapping systems rely on cartographic databases, i.e., digital representations of map information in a form amenable to computer processing, to provide base map information. In the current study, the base map data will include major lakeshores, coastlines, and political boundaries. The acquisition, processing and maintenance of such databases can be expensive. Therefore, efficient handling of the cartographic data is required for applications involving a wide range of map scales. One approach is to acquire a number of different databases, each appropriate for a relatively narrow range of scales. However, the expense of dealing with numerous externally developed databases, along with possible difficulties when changing scale near the limit of a given database's scale range, indicate that this solution is not ideal. Another approach is to derive maps for all scales from one underlying database having a large enough scale to include the maximum detail required by the system. This minimizes the problem of handling many different data sets and can ensure consistency at various scales. However, substantial generalization is required to derive a small-scale map from a much larger scale database. Numerous

algorithms exist to perform automatic line simplification (Zoraster, Davis, & Hugus, 1984), which is an essential component of the necessary generalization. This study applies some of these algorithms to the simplification of cartographic databases composed of hundreds of different lines. Thus, line simplification techniques are examined in the context of their use with databases which approach the complexity required for operational cartographic systems.

There are two main goals for this work. The first is to perform a comprehensive comparison of several of the various techniques for line simplification. The second goal is to quantify the degree to which the useful range of scales of a cartographic database can be extended by the application of these techniques. The results of this work provide an indication of the feasibility of deriving maps at widely varying scales from one database. To pursue these goals, small-scale map information is derived from databases intended for use at larger scales. Preliminary work, covering modest changes in scale, utilizes the National Center for Atmospheric Research cartographic database (NCARCDB), which was originally created for use at very small scales. Examination of the behavior of the various techniques over extreme changes in scale makes use of a much larger source of digital map data. This database is the Central Intelligence Agency (CIA) World Data Bank II (WDB II) that was designed for use at scales from 1:1,000,000 to 1:3,000,000. Simplified map data is derived for several smaller scales (e.g., 1:10,000,000; 1:20,000,000) using six generalization techniques. The performance of these techniques is evaluated both subjectively and mathematically.

A number of subjects are covered to provide a background for the work that was performed. Chapter II covers some aspects of the theory of cartography and

generalization and some of the features of traditional generalization practice. Chapter III gives a summary of the current state of computer application in cartography and some of the difficulties that exist in this area. Chapter IV is a discussion of cartographic databases and their structure. In addition, the topic of data capture as the means of producing a cartographic database is considered. Chapter V reviews a number of line simplification algorithms. The details of the methodology used in this study are described in Chapter VI and Chapter VII covers the results this work. Chapter VIII examines some minor enhancements to one promising algorithm and discusses some aspects of the use of map projections in line simplification. Conclusions are presented in Chapter IX.

Chapter II

Cartographic Generalization

A map can be defined as “a device for storing and communicating information about the physical and social phenomenon that are distributed over the earth’s surface” (Campbell, 1984, p. 52). They are sometimes thought of as being pictures, though abstract pictures, of a portion of the earth’s surface from above. Thus, the sheet of paper on which the “picture” resides is the device which stores information about the earth; the picture may communicate some of this information to a map reader. Although the model of a map as a picture is helpful in some ways, it does have limitations. It oversimplifies some of the important processes that occur in the creation of a map. A more general model of cartography can be developed by considering the transformations involved. Perkal first suggested that “cartographic transformations, employed to represent the surface of the earth on maps, are of two simple types: map projections and generalization” (1958). Map projections are

means of geometrically or mathematically transforming locations on the earth to positions on a flat surface such as a piece of paper (Snyder, 1982, 1987). While this type of transformation can have substantial errors, these errors can be precisely described (Laskowski, 1989; Maling, 1972).

Generalization is also a transformation. In this case, real data about the earth is changed into a form that can be represented clearly on the map. This transformation is much less precisely defined than the map projection transformation, and the errors that result from this process are not easily described. In addition to

these two central transformations, all the processes of map making, from surveying to printing, can be viewed as transformations with associated errors and distortions which may or may not be quantifiable (Tobler, 1976).

Generalization and Information Theory

In treating the map-as-a-picture model, a question arises regarding the utility of mapmaking versus aerial photography (and remotely sensed imagery) as a means of presenting spatial information. It can be argued that an aerial photograph contains more information than a map and, therefore, it is a preferable means of communicating spatial information. If the data content is measured in bits using the concepts of information theory, it can be shown that this is true (Knöpfli, 1982). Consequently, if maximizing the data that appears on a piece of paper is the goal, then an aerial photograph can be considered superior. However, a more productive way of viewing maps and aerial photographs is to consider them as channels of communication, i.e., a means of transmitting messages from a source (the map maker or photographer) to a receiver (the person interpreting the patterns of the photograph or map). Communication channels fall roughly into two types: noiseless and noisy. In a noiseless channel, the message presented by the source is transmitted in the same form to the receiver. Thus, information is moved from point to point without error. On the other hand, a noisy channel introduces errors in the message being transmitted, i.e., the information transfer is unreliable (Shannon & Weaver, 1949).

In information theory, messages are evaluated by the amount of uncertainty they resolve. The uncertainty of a message can be precisely defined. As an example, consider an aerial photograph. Such a photograph can be scan-digitized via a television camera using 512 (2^9) scanlines and 512 (2^9) pixels per scanline. Each resulting pixel can typically take 256 (2^8) gray scale values. In an arbitrary aerial

photograph it is reasonable to assume that any pixel is equally likely to take on any of the 256 gray values. Given this, it can be shown that such a digitized photograph represents an uncertainty of 26 bits. Thus, 26 binary digits can differentiate each of the possible digitized photographs since $2^9 \times 2^9 \times 2^8 = 2^{26}$. Given this large number of possible messages, the receipt of a particular message resolves a great deal of uncertainty by eliminating such a large number of the possible messages.

Information theory also shows that if any of the possible messages is more probable than the others, then the uncertainty decreases. In a typical map, large areas have constant tone. The gray scale of a digitized map would vary in steps from one tone to another. Each tone would be maintained for some distance along any scanline. This implies that certain messages represented by a typical map are much more probable than others and this, in turn, implies that the uncertainty is much less. This could be taken to mean that the information imparted by a map is much less than an aerial photograph. The work of Sukhov (1970) relies on this view. However, Knöpfli (1982) has pointed out that this argument is based on the assumption of a noiseless communication channel. He suggests that the model of a noisy channel is more appropriate.

In a noiseless channel, the message sent is precisely the message received. The process of transmission adds no uncertainty to the message. Thus, all of the uncertainty that existed prior to the arrival of the message is eliminated. This is illustrated in Figure 2-1a, where the original message, X , is received without error as message Y . Each component of original message, X_n , is transferred unambiguously to the corresponding component of the received message, Y_n . In this case, once Y is received, the receiver knows the precise message sent. Few communication channels truly fit this description. In the case of cartographic communication between the

Figure 2-1. (a) A noiseless channel. (b) A noisy channel. (c) An example of equivocation. After Knöpfli (1982, p. 85).

cartographer and a map reader, the assumption of a noiseless channel is particularly inappropriate.

In a noisy channel, a given input message may result in any of several received messages. As a result, the receipt of a particular message does not inform the receiver precisely which message was sent. This is shown in Figure 2-1b. The fact that a given component of X could be received as any of several components of Y implies that the reception of a given message does not inform the receiver of the original message. That is, the receipt of a component, Y_n , only indicates to the receiver the original message contained one of several possible states, X_i, X_j, X_k , etc., as shown in Figure 2-1c. This confusion, or *equivocation*, on the part of the receiver diminishes the information content of the message. Considering again the use of a scanned aerial photograph as a means of communicating information about the environment, there is substantial ambiguity associated with the interpretation of a pixel having a given gray scale value. A particular gray value does not have a well-defined meaning. Meaning must be extracted from the image by the receiver on the basis of both the local and global changes in gray scale values. Another way of viewing this problem is to consider number of themes, or different classes of information, contained in an aerial photograph. Such themes include land use, land cover, landforms, drainage, population density, transportation networks, cultural

artifacts, and many others jammed together on one piece of paper. In some cases, the information associated with a few neighboring pixels will relate to all of these themes and more. These difficulties are compounded by various kinds of noise that occur in the processing of a photograph. This increases the equivocation of the transmission and, as a result, decreases the amount of information which is transferred.

Figure 2-2 provides a comparison between a satellite image, derived from Multispectral Scanner data from Landsat-3 adjusted to 1:250,000 scale, and a topographic map for the same scale and area. The image contains a vast amount of information on a variety of themes; it would be very appropriate for a study of land-use patterns or geological structures near Las Vegas. However, if the information needed from the image or map is to be used to navigate the main arterials of Las Vegas and the surrounding backroads by automobile, then the merits of the map become apparent. While the image contains some visible evidence of possible transportation routes, many of the relevant details are obscured by a wealth of competing information which, for the purposes of automobile route-finding, is noise. In contrast, the map shows a connected network of roads which is clearly distinguished from the other information on the map.

Communication via a map also can be represented by a noisy channel. However, information theory shows that it is possible to reliably transmit messages over a noisy channel if a suitable coding scheme is used. By limiting the total number of possible messages in the encoding process, the remaining capacity of the communication channel can be used to ensure that the smaller number of possible messages are received correctly. In digital electronic communication, this usually means that a certain number of bits are used to enhance the message bits in such a way that

the reception of an incorrect bit can be detected and corrected. In the case of cartography, the process of limiting the number of possible messages and coding them in such a way that they can be transmitted reliably is called generalization.

The cartographer often has message of limited scope to be transmitted. This is particularly true in thematic mapping. Much of the process of designing a map concerns the clear presentation of the cartographer's message to a map reader. In other words, the goal of map design is to create a communication pathway that functions well in the presence of various kinds of noise. There are a number of ways in which this goal can be approached. Information that is irrelevant to the intended message can be removed. Thus, details of population distribution are rarely presented on a meteorologist's forecast map. Once the important data has been selected, it is often helpful to simplify this data so that excessive detail does not confuse the larger pattern that contains the intended meaning. Information is often presented more clearly if it is grouped into related categories rather than as a continuous range of values. Such a classification allows each category to be displayed with an unambiguous symbol. For example, distinct symbols differentiating interstate highways from city streets allow certain information about the transportation network to be communicated more reliably than would a continuous mapping of road width to line width. The process of choosing effective symbols for use on a map is called *symbolization*. The cartographer can also extend the basic data by assuming a model for the phenomenon which the data represents. For example, by assuming a model of the behavior of an actual wind field, a map of wind vectors can be extrapolated from a few observations in space. Such additional inference on the part of the cartographer is called *induction*. These processes of selection,

simplification, classification, symbolization and induction are the components of the wider concept of generalization (Robinson, Sale, Morrison, & Muehrcke, 1984).

It follows from the considerations above that the tools of generalization are the means by which a cartographer transfers information via inherently noisy communication channels. These tools can allow the mapmaker to transmit a message about location or spatial variation to the map reader unambiguously. This implies that generalization is, in many ways, the core of cartography. The generalization process is the factor that creates a need for cartography in an age of high-quality aerial photography and satellite imagery. Remote sensing creates vast amounts of data that the experienced interpreter can use to great advantage. However, cartography is the medium of choice for presenting clear messages about phenomena on the earth's surface. These two basic methods for presenting spatial information differ in who performs the initial interpretation of the source data. With an aerial photograph or a satellite image, the viewer must perform virtually all of the interpretation. With a map, the mapmaker performs much of the interpretation of the basic data before and during the creation of the map in order to pass his or her ideas on to the map reader. While maps and images are different in fundamental ways, they are complementary in their description of the earth's surface. In some cases, the best of both worlds can be achieved by combining the two approaches in an imagemap (Peruzzi, 1975).

Limitations of Generalized Communication

While the use of generalization techniques can minimize the noise in the communication path from cartographer to map reader, the process of generalization also places limits on the communication that takes place. A perfect communication channel from the cartographer to the viewer of the map would allow the viewer to

reconstruct that portion of the cartographer's reality that relates to the map. In other words, the map reader could, by the process of reading the map, recreate all the information used by the cartographer in creating the map, along with all the cartographer's thoughts and assumptions about the material being portrayed. If such a map could be created, then the geographical concept which the cartographer was presenting would be transferred without error or omission to the map reader. Each of the subprocesses of generalization limit the accuracy and completeness with which information can be transferred via a map (Morrison, 1975, 1974).

The difficulty with the entire process of generalization is that the information removed in order to clarify the message cannot be recreated by the map reader. As a result, the reader can only receive a portion of the original concept. For example, in selecting the information for the map, the mapmaker omits data which, in some cases, may be related to the underlying message but clutters up the more important data. This data is irretrievably lost to the reader. Consequently, the reader may miss certain subtle aspects of the concept. The classification process provides a more extreme example. In the presentation of statistical information, there are many methods of grouping a distribution of interval data into classes, e.g., equal intervals, quantiles, and logical groupings. The results of a given classification can be used to create a choropleth map. The reputable cartographer will study the data distribution and select a classification method that highlights the important aspects of the data. However, other methods could be selected that lead the map reader to form erroneous conclusions about the distribution. Because the original information has been generalized, the map reader cannot know if the information has been presented honestly (Haggett, 1967). The intentions of the cartographer, whether

benign or malign, are not available to the receiver at the end of the communication channel.

Traditional Cartographic Generalization

As described above, the theoretical goal of generalization is to create a relatively error-free communication channel between the cartographer and the map reader. The traditional practice of generalization can be examined on the basis of the same sub-processes utilized previously: selection, simplification, classification, symbolization and induction. All of these components must be considered in light of the purpose of the map. Also, when considering generalization in any practical context, the scale of the planned map becomes central to the entire process. This is because the degree of generalization increases as map scale decreases. Thus, the generalization required to produce a general purpose world map is related to, but quite different from, the generalization needed for a large-scale topographic map.

In discussing the topic of generalization, it is helpful to differentiate between the common types of map features. These types are points, lines, and areas. The distinctions between these are somewhat fuzzy because, at a large enough scale, most physical features on the earth can best be represented as areas. However, at a given scale, it is usually clear which representation is appropriate for each kind of feature being mapped (Dent, 1985). Each of the component processes of generalization will be discussed in a separate section; however, in traditional cartography, they are often highly integrated in the mind of the cartographer.

Selection

The overall goal of selection is to identify the specific information that is to be placed on the map in order to communicate the message of the cartographer. Robinson, et al. (1984) maintain that selection and generalization are distinctly

different processes. While such a differentiation can be made, the selection process has many of the characteristics of other generalization processes, including the controls of map purpose, scale, graphic limitations, and data quality. Therefore, in this study, selection will be considered a subprocess of generalization. There is a great variety of measured and derived information that can be associated with a given location or region and only a small portion of that information can be presented on one map. Therefore, the first task of the mapmaker is to determine the types of information, or themes, that are to be portrayed. This is true for both base maps and thematic maps. For example, a decision must be made whether to include major highways on a base map. The best answer will reflect the map purpose, scale, any previous themes that have been chosen to appear on the map, and whether the quality of graphic reproduction to be used in creating the map can represent the additional information without confusion or clutter.

After the basic themes to be mapped have been identified, additional selection takes place as particular items within each theme are included or omitted. For example, the decision to include the locations of populated places on a map does not mean that all populated places must be represented on the map. A reasonable density of features must be maintained as a function of scale (Salichtchev, 1976; Srnka, 1970; Töpfer & Pillewizer, 1966). It may be possible to portray every hamlet on a 1:24,000 topographic map; however, if all such places were shown on a world map, an unreadable clutter would result. Thus, this part of the selection process is closely related to the scale of the map being produced. In the case of determining which populated places to represent, a number of approaches can be used to make the selection. One possibility is to represent all such places with a population larger than a certain value. For example, on a 1:250,000 map, all villages with more than

100 people might be portrayed, while on a 1:2,000,000 map, the population threshold might be 5,000. While such an approach is relatively simple to implement, it is too simple to work well in many cases. One difficulty occurs because the distribution of populated places on the earth is extremely non-uniform. Thus, a population cutoff value will result in areas with low population densities having no place names, while heavily populated areas will have too many. So the threshold may have to be modified to reflect the relative economic or social importance of the place. In densely populated areas, a medium-sized town might have little relative importance compared to major cities nearby and, consequently, it could be eliminated from the map. A small village, well removed from any large population centers, could be justifiably included on smaller scale maps because it might serve as an essential node for the nearby populace (Kadmon, 1972).

In addition to the map purpose, scale, and minimization of clutter, the selection process for information from one theme may be affected by information from another theme. Different themes may be related by some underlying mechanism. This means that the level of detail for each of the two related themes should be comparable. For example, most topographic maps include both hydrography and hypsography. Rivers and streams are usually represented by lines along their paths and the elevation surface is shown by contours. These two themes are highly related in that the drainage network is responsible for much of the variation in the elevation surface. Thus, valleys and other undulations that are represented on the elevation surface should correspond to the rivers and streams at the same locations (Panekoeck, 1962). This correspondence is very strong in areas with substantial rainfall where most streams are perennial. In terms of selection, this implies that the choice of which streams to include cannot be made independently from the generalization

process for the contours. The level of detail for the two themes must be properly matched. In general, the traditional process of selection relies on the cartographer's understanding of the geographical message to be transmitted, the geographical variation of the themes being presented, and the underlying geographical relationships between the themes.

Simplification

Once all the items to be portrayed on the map have been selected, many of them may have to be simplified. Simplification is the removal of unwanted detail that detracts from the message of the map. A cartographic line, either the representation of a linear feature or a portion of the boundary of an area feature, can contain so much detail that it becomes a jumble that obscures the more important trends and patterns the line is intended to portray. For example, an important goal in the representation of a coastline is to portray the geomorphic character of that coastline accurately on the map. A long smooth sandy beach looks much different from a rocky coastline. A map of a region along a coast should convey such differences over a range of scales. If every undulation of a complex submerged coastline is maintained as scale is decreased, then at some scale the coast may acquire an unappealing, jagged appearance that does not convey the correct impression of the character of the land-water boundary. Figure 2-3a shows a line which has some of the complexity of structure of an actual cartographic line as the scale decreases from right to left. In the smallest line, features of the line are coalescing and the structure of the line is not represented well.

On a somewhat more practical level, the process of representing very fine detail with a pen of specific width is a problem. If a cartographer attempts to show detail which, when represented at the scale of the map, is less than the line

Figure 2-3. (a) A narrow line drawn at several scales. (b) A wide line drawn at several scales.

width being used, then the line will tend to broaden and blur to varying degrees depending on the complexity of the line at different points (Brophy 1973). This will create an unattractive presentation of the feature and may make it difficult to differentiate features distinguished by different line widths. Figure 2-3b shows the same line as in Figure 2-3a, but symbolized with a wider line. The problems with line representation are accentuated by the use of the wider line.

Simplification for a relatively small scale map is often done on a reproduction of larger scale source material at the scale of the source material. The new linework is done with a felt pen with the width selected so that the lines will be the correct width when the drafting is reduced to the target scale (Zoraster et al. 1984). The width of the lines, as they are drawn, gives the cartographer a sense of the level of detail that can be maintained on the new map. In many cases, simplification involves a relatively straightforward smoothing of the line so that details that cannot be represented at the target scale are replaced by an averaged line that maintains the general trend and structure of the original feature. However, in some cases, particularly when the change in scale is large, attempts at simple smoothing will run into difficulties. A classic example is provided by a submerged coastline

indented by fjords. At large and, perhaps, at intermediate scales, all or most of the structure of the coastline can be represented realistically. At small scales, an average representation of such a coast would be smooth since virtually all the fjords would be much narrower than the line width; consequently, the coastal undulations could not be represented. While the mapmaker can provide a general impression of the coastline's position, the smoothed line would give an erroneous impression of the geomorphic character of the coast. To effectively communicate both the location and the character of the coastline, the cartographer will often exaggerate the largest fjords while eliminating the smaller features. Thus, a cartographer, with the necessary understanding of geomorphic structures, can present a more appropriate message about the coastline to the map reader.

A related difficulty can arise in regions which were covered by continental ice sheets during the most recent glacial period. These areas are often dotted with innumerable small- and medium-sized lakes. At large scales, all the lakes can be shown. However, as scale decreases, more and more of the lakes will become too small to represent at the new scale. If the scale is made too small, then very few lakes will appear on the map, because most of the lakes will be too small to represent if drawn at the correct size. Elimination of lakes that cannot be represented will convey an inaccurate picture of the landscape to the map reader. As with the problem of fjords, a solution depends on cartographic license, i.e., minor alterations in certain details are made to avoid larger errors in conveying the message of the map. In this case, certain lakes may be enlarged beyond their accurate representation at the new scale in order to represent numerous smaller lakes nearby that have been eliminated. Thus, a number of lakes within a small area are coalesced into one lake (Nickerson & Freeman, 1986). If this is done carefully, the relative areas

of land and water can be maintained as the scale decreases. Thus, the cartographer can convey the impression that the region contains numerous small lakes over a wide range of scales.

In some cases, the merging of small islands may be required as scale decreases. An example is shown in Figure 2-4. A portion of the coast of Chile is portrayed in Figure 2-4a at appropriate level of generalization and at a substantially smaller scale. The effect of reducing scale while maintaining all the geomorphic detail is

disastrous; the small islands and narrow fjords merge into an unreadable tangle. Some processing of the coastline is clearly necessary for use at the smaller scale. The elimination of all the islands is clearly inappropriate since much of the area off the coast is actually land and some of the islands are large enough to be represented at the smaller scale. Figure 2-4b shows one possible solution. Several of the islands are merged into larger islands, smaller fjords are eliminated, and the larger channels and fjords are maintained or enlarged to ensure an acceptable representation at the smaller scale. On reduction, the correct character of a submerged coastline with numerous offshore islands separated by narrow channels is maintained. This aspect of simplification is analogous to what would occur in a satellite image as the pixel size increases. At a certain resolution, many of the channels and fjords would be much narrower than a pixel, while most of the islands would continue to be resolved. As a result, most of the reflectance of the pixels covering the islands and narrow channels would be representative of land. A merging of features similar to those shown in Figure 2-4b would occur.

Classification

In many cases, the message that is to be transmitted via a map can be clarified by grouping portions of a set of data into related subsets or by dividing a continuum

Figure 2-4. (a) A detailed representation of the Chilean coastline and a reduced version of the same line. (b) A generalized version of the same coastline which is suitable for reduction.

of values into separate ranges. This process is called classification. For example, different roads can be grouped in various ways to assist in the presentation of the message. Roads can be divided into their administrative categories (interstate, U.S., state, ...). They could also be grouped by the number of lanes. Such classifications are examples of simple groupings based on clearly defined characteristics of the roads. However, if roads were to be grouped on the basis of the volume of traffic (number of vehicles) they carry per day, then more complex decisions must be made regarding classification. Each individual road will usually have a specific value for the traffic volume that is derived from observations. Few, if any, roads will have exactly the same value. If each numeric value was printed on a city street map, it is likely that the map would be very cluttered and that the general variation in the traffic flow, which should be the principal message of the map, would be obscured by all the detail. However, if all the separate values were grouped in some reasonable way, then a small number of classes could be symbolized clearly and distinctly.

If this classification is done well, the more important trends and patterns will be highlighted.

There are a number of ways to group a distribution of data (Robinson, et al. 1984). The choice of an appropriate method will depend on how the data is distributed in the range between the minimum and maximum values. In some cases, a logical grouping will be apparent. In the example of traffic volume on city streets, it might be that the data will naturally fall into three subgroups reflecting lightly, moderately, and heavily used roads. A boundary value for traffic volume would be chosen to differentiate between lightly and moderately used roads, and another value would be used to distinguish moderately and heavily used roads. Another possibility for classifying this type of distribution is to identify percentile groupings. For example, the 25th percentile is the value for traffic volume with 25% of the roads having less than that volume and 75% of the roads have more than that value. A common choice for classifying a distribution is to divide the distribution so that there are four categories, or quartiles, each with an equal number of observations. It is also common to divide the range of values up into equal intervals, typically so that the boundary values are even numbers. For example, an elevation surface ranging from 150 to 670 meters above sea level might be represented using boundary values of 100 meters, 200 meters, 300 meters, and so on. There are numerous other methods to classify data, some of which can be very complex. However, the underlying concept is the same. A large number of individual occurrences of a type of feature or a type of observation are grouped with similar occurrences to simplify the resultant map, thereby allowing the important characteristics of the data to be easily understood.

Classification is, in many situations, tied to the scale of a map. This is especially true when mapping nominal data. The number of types that are distinguished must be appropriate to the scale of the map. Considering land use/land cover as an illustration, if the classification is too coarse, then the entire region being mapped may fall into one category (e.g., agriculture) and the resulting map will be of minimal interest. On the other hand, if the classification is too detailed, then the resulting map is likely to be unreadable due to the density of information. In a number of land use/land cover classification systems, this problem is addressed by providing a hierarchical classification scheme (Anderson, Hardy, Roach, & Witmer, 1976). The highest level of such a system is very coarse and is appropriate for small-scale mapping. Lower levels in the hierarchy are more detailed and are intended for use at larger scales (Campbell, 1983; Taketa, 1974).

Classification is heavily used as part of the representation of statistical data in thematic mapping. However, it also plays an important role in base maps. These maps often contain roads and other features that are portrayed more clearly after having been classified into a few groups. Classifying cities represented as point symbols on the basis of population or administrative function allows important and well-known cities to be highlighted. This allows the map reader to get oriented with greater ease. Classification is also related to the selection process used when a map is being created using source material from a larger scale. Complete categories of data may be omitted because they are too detailed for the scale.

Symbolization

Given the information that has been selected and classified, problems arise concerning how to show that information on the map. This is the process of symbolization. In the case of point, line, and area features, this will involve the selection

of graphic attributes for each type of feature (Robinson, et al. 1984). For example, the graphic attributes for a line include its color, width, and style (e.g., solid, dotted, dashed). The choices in symbolization should be made judiciously to ensure that the different types of features are readily distinguished and that the important features stand out from the background and any ancillary information. In the case of base map information, the association of geographic features with point, line, or area representations is usually obvious, given the scale of the map. Thus, rivers are represented as lines and cities as points on a small-scale map. On a large scale map, both rivers and cities would be shown as areas.

In thematic mapping, more thought is often required to select the proper means of symbolization because the information is often abstract and, therefore, easily misrepresented if an inappropriate choice of symbols is made. For example, it is possible to symbolize the population within an administrative unit (e.g., states or counties) by showing the unit filled with a level of gray that represents a range of population. This means of showing population can be highly misleading. The difficulty arises from the tendency to interpret a map of quantitative data over areal units as a distribution. Symbolizing raw population values as areas does not accurately represent a distribution, unless each administrative unit is exactly the same area as all the others. This can be seen by considering a hypothetical region over which people are evenly distributed. If population values are associated with areal units of unequal size within the region and the values are classified into a set of ranges that can be portrayed using a set of gray-scale values, then large areal units will be shown as having large values and small units will be portrayed as having small values. This means that large areas will typically be covered with a dark gray and small areas will have a light gray. This will give the erroneous

infinitesimal width and the points had infinitesimal width and length), then there would typically be no conflict between features at the reduced scale because there was no conflict at the scale of the source material (Monmonier, 1986; Nickerson, 1988; Nickerson & Freeman, 1986). Since the lines and points have finite size, features that did not touch on the source material may overlap at the scale of the new map. These conflicts, which result from the process of symbolization, require moving one or both of the overlapping features so that they may be clearly distinguished. The choice of which feature to move will normally be based on a prioritization of feature types. The positions of the features in the more important classifications will be maintained while the less important features will be moved. Once the decision about which features are to be moved has been made, the cartographer must redraw the feature, attempting to minimize both the error in position and the distortion of the shape while maintaining a detectable space between other features.

Induction

Induction can be viewed as the final step in the process of generalization. This step involves the inference by the cartographer of new information to add to the map in addition to the actual observations being presented (Robinson, et al. 1984). For example, if a number of observations are plotted on a map and the cartographer interpolates a surface that represents the data portrayed on the map, the cartographer has induced information that is an extension of the source information. Such a step is based on a model which the cartographer assumes is valid for the situation being presented. Various interpolation techniques reflect such models. The addition of such information can clarify a pattern or trend that is not easily apparent from the observational data alone. Weather maps showing temperature values at reporting stations along with regions that are color coded

to represent temperature ranges are a concrete example. Induction can also occur when a reader studies a map and is able to add personal knowledge to the mapped information to develop a further understanding of some phenomena.

Difficulties in Traditional Generalization

The major problem in traditional generalization is the extreme reliance on the geographical expertise of the cartographer (Wright, 1942). That is, there is no correct generalization for a particular situation; generalization is only a reflection of the cartographer's understanding of the problem. Distinctions can be made between relatively good and relatively poor generalizations. However, the generalization process is highly subjective and traditional cartography has made little progress in developing techniques for controlling the generalization process more strictly. As a result, the generalizations produced by two cartographers may be strikingly different. Also, the same cartographer drawing the same map using the same source material, but at different times, will produce a different result. These differences will reflect the level of geographical knowledge and the experience of the cartographer at the time the map is made. This difficulty arises during each of the component processes that make up generalization. Thus, selection, simplification, classification, symbolization, and induction are all subjective and cannot be reliably recreated by different cartographers at the same time or by the same cartographer at different times. This can result in difficulties in interpreting maps. In a sense, the subjective nature of traditional generalization increases the noise in the communication channel, since the map reader is unlikely to hold the same views as the cartographer concerning the specific data or the larger problem being considered.

Chapter III

Computer Use in Cartography

Cartography is both an art and a science. Much of the art of cartography is manifest in the various aspects of traditional generalization. A substantial portion of the science of cartography includes tasks that are mathematical in nature and highly repetitive. As such, these tasks are good candidates for implementation on computers. This match between computers and cartography was recognized as early as 1959 (Tobler). The computer has already made important contributions to some aspects of cartography, and has a more important role to play in the future; however, the development of truly automated cartography is limited by the difficulty of applying computers to the subjective aspects of the field.

At present, there are a number of specific situations where the computer is successfully utilized. These include statistical analysis, environmental modeling, processing satellite spectral data, and applications in emergency response. Many of these applications are impossible without the aid of large, fast computers and, in these situations, machines are clearly adding new capabilities to previously existing ones. However, when considering the automation of other tasks, the applicability of computers is less clear. While providing advantages of speed, along with ease of revision and replication, computer processing often results in products of lower quality as a result of current technological limitations. In some cases, changing situations could force the reprogramming of computer software to meet mapping requirements. This activity is usually quite expensive and time-consuming. These

problems are solvable, at least in principle, with advances in technology and further software development. Automated cartography, as opposed to computer-assisted cartography, poses additional difficulties that may not be solved in the near future. The adaption of cartography to this new tool is an ongoing process which will not be completed for some time. While some of the effects of the computer can be unsettling to traditional cartographers, the process of adapting to new technology is really nothing new. The history of cartography is in large part the story of such adaptations throughout the centuries (Wilford, 1981).

The Promise of Computerization

Traditional cartography has involved several steps including data gathering, compilation (which includes the generalization process), production, and reproduction. Each of these processes can involve steps that are repetitive and time-consuming, and which can be expressed in mathematical or procedural terms. For example, the plotting of a reference graticule, the manipulation of statistical data, and the drafting of line work on a map can be very tedious. These procedures are prone to error in the hands of an inexperienced cartographer. Computers execute such repetitive tasks tirelessly and, when programmed correctly, perform the work without error. Computer systems are also designed to store and process huge quantities of data. This allows the creation of very large data sets organized to provide quick, easy, and repeated access to the information. These data sets are referred to as cartographic databases (Campbell, 1984). When these capabilities are combined with the rapidly developing techniques in computer graphics and human-computer interfaces, the potential for aiding the cartographer's task seems unlimited.

When working with location data in digital form, which makes up much of a cartographic database, the problem of the generation of a map projection becomes

trivial. Once the required mathematical procedures are encoded into the computer, the choice of projection involves only the selection of the appropriate procedure to be executed (Snyder, 1982, 1987). Because of the great speed of computers

at mathematical operations, the procedure can be used for each location in the data set. As a result, if the map projection coordinates are used for plotting, each location displayed will be as precise as all other locations. This contrasts with such traditional techniques as tracing from an optical blowup of a pre-existing map where the accuracy near the center of the image will be better than at the edges. Also, when changing projections by matching a graticule on the source map with one drawn on the new map, the points where linear features intersect lines on the graticule will tend to be more accurate. With access to computers, the choice of projection can be made on the basis of cartographic reasoning alone; thus, the availability of base maps using the same or similar projections is not a factor in the decision (Monmonier, 1982). Also, changing scales over limited ranges becomes trivial in the computer environment; in some computer graphics terminals, a change of scale is incorporated into the hardware of the device, thereby allowing scale changes with the push of a single button.

The ability of computers to quickly repeat operations implies that, once a computer map has been made, copies can be produced quickly and easily. Generally, the task of revision becomes proportional to the degree of alteration required in the map. The utilization of new data to produce an updated copy of a map can, once again, be as simple as pressing a single button. By comparison, any important change in a hand-drawn map may require complete redrafting of the map. More extensive changes to a computer map may also be very simple to make. For example, once a routine is set up to make a choropleth map based on a set of areal units

using data values associated with each unit, the same routine can be used with different data values for the same units. The only change would be to alter the list of data values. Many such routines are available that allow the user to select the classification intervals, title and legend information. By entering a few sets of numbers and some additional labeling information into a computer, a large series of maps can be produced with nearly the same effort as that required to produce a single map.

The ease with which information can be replicated and revised on a computer provides a basis for more effective display of information. In manual cartography, decisions must be made about the design of a map before the map is created. Experience usually allows adequate design choices to be made; however, once the process of drafting the map begins, the commitment to the basic design is almost irrevocable. New ideas that occur both during the creation of the map and with inspection of the finished product must be filed away in memory as experience for use in future maps. In contrast, the speed of computers and graphics devices makes it possible to try many different presentations of the same data using flexible mapping software (Campbell, 1984). The possible variations may include different data classifications and methods of symbolization. Because it is quick and easy to make changes of this nature, ideas on improving the presentation of the information can be applied to the current map as well as future maps. This iterative, computerized approach to map design allows the cartographer to rapidly converge to a satisfactory result.

Once a traditional map is completed, the information stored on the map can only be used in limited ways for further mapping or analysis. For example, geographical outlines can be traced onto another map. In computer mapping systems, the storage function of the map is separated from the display function. The digital

data used to create a map remains stored on the computer for any future use including the creation of additional maps, the performance of statistical summaries, and to support spatial modeling. This means that data gathering can be more effective because, once the data is captured into computer form, it can always be available for further processing and analysis in new and innovative ways. For example, a road network, digitized for the creation of a computer map, can also be used as input to a computer program that models traffic flow in a city. Similarly, digital terrain data, acquired for use as input to atmospheric models, can also be used to generate contour maps of an area. This is perhaps the most fundamental effect of adapting cartography to computers. To quote Dobson (1983, p. 139):

... the manually prepared map is simply a generalized 'picture' of reality. The new (digital) product is a flexible data base which can be incorporated readily into a variety of analytical as well as descriptive studies which can contribute to learning and gain greater value from the initial effort.

The separation of the storage and display functions that occurs with computer mapping also allows for "non-map" cartography. Certain cartographic analysis functions that are traditionally done using paper maps can be performed entirely on a computer without a map being made (Yoeli, 1975). For example, the areas of regions displayed on maps can be measured manually, but digital data on a computer describing such regions can be utilized as input into computer programs that will determine the area of each region much more quickly and probably more accurately than can be done manually. Most map mensuration procedures can be done effectively using the appropriate map data on the computer.

Because the computer can store and process huge amounts of data, it is possible to take many types of spatial information and integrate them into a single system with location being used to tie the data together. Such a system permits

spatial patterns in many different types of data to be compared, contrasted and combined, thus facilitating geographical study. This is the concept behind geographic information systems. These systems have already made important contributions in many fields that rely on spatial data. Geographic information systems are a rapidly expanding area both in research and in commercial applications (Coppock & Anderson, 1987; Dangermond, 1983).

The computer is remarkably applicable to many kinds of problems in cartography. Its abilities to store, process, and display vast quantities of data seem ideally suited for the making of maps.

Current Applications

Important work on the application of computers to mapping is proceeding in a number of areas. Large mapping organizations such as the United States Geological Survey (USGS) and the Defense Mapping Agency (DMA) are expending a great deal of effort to take advantage of computers in producing reference maps for the United States and other areas of the world. The amount of manual work required to produce and maintain the various topographic map series is tremendous and, in some cases, nearly prohibitive. For example, it will require about 55,000 7.5-minute quadrangle maps to cover the U.S. at 1:24,000 scale as part of one major USGS series. Since it takes about two man-years to hand scribe a single 1:24,000 map, it has not been possible to produce a complete series of these maps even though work has continued on this project for decades. Complete coverage is finally expected by the end of 1990. However, the work on such a map series does not end with complete coverage, it must be kept up to date to be useful. Revision of the planimetry should take place on a regular basis, especially in populated areas where cultural features can change very rapidly. Automation is the most promising solution to the problem of creating

and maintaining the tremendous number of topographic maps needed to cover a country as large as the United States. Consequently, USGS is researching means of automating much of the mapping process. The results of this work are being applied to the creation of new map series designed to be amenable to computer processing, as well as to the adaption of the traditional map series to modern techniques. The long-term goal is to transfer all map series to computer form by the year 2000 (Starr, 1986). Similar work on digital mapping is being done by national mapping organizations in Canada and Great Britain (Darragh, 1979; McMaster, Haywood, & Sowton, 1986), as well as other countries.

In addition to the creation of cartographic data bases via data capture from pre-existing paper maps, work is being done on automated techniques for acquiring the source information for maps (Monmonier, 1985, pp. 76-109). In particular, the use of remote sensing systems deployed on satellites is an active area of research. While remote sensing is not formally a part of cartography, it is an important source of the information necessary to produce cartographic databases. In many cases, current remote sensing techniques are highly adapted to computer capabilities. For example, Landsat data is created in digital form by scanners on the satellite. While a great deal of information can be extracted from an image generated from the unprocessed reflectance data, much more information can be provided by subsequent computer processing. The radiance values can be calibrated, pixel locations can be determined and the data resampled to a convenient coordinate grid. This allows the Landsat data to be matched with other spatial data from different sources to aid analysis. In addition, techniques such as histogram stretches, smoothing, band ratioing, and various classification techniques allow patterns to be identified in the data (Mausel, 1985; Simonett, 1983; Howarth & Boasson, 1983). The volume of

data involved in a Landsat scene would preclude the use of manual techniques to accomplish this type of processing. Landsat data provides a new source of raw data from which geographic information can be extracted in a highly automated manner. This type of spatial data could not be produced without computers and the utility of the data would be greatly limited without extensive digital processing. A long-term goal for the use of this type of remotely sensed data is the direct extraction of finished maps from the raw data.

An emergency response to the release of a radioactive or toxic substance into the atmosphere is one example of an application which must rely on computers and computer mapping. An effective response to this type of incident involves sophisticated computer programs that model the transport and diffusion of the

Other areas in which computers are applied to spatial or geographic problems include urban planning (Dangermond & Harrison, 1987), resource management (Reisinger & Davis, 1987), vehicle navigation (French 1987), the modeling of environmental phenomena (Dangermond et al. 1985), terrain representation (Belles, 1985), and military systems (Hardaway & Porter, 1985). In most of these examples, computers are utilized to provide very specific capabilities. When more generality is incorporated into a mapping or geographic information systems, the system is usually applied to a specific and limited area of interest, such as a single urban area. While these examples indicate that the computer plays an active and central role in cartography, the tendency to limit systems to specific situations or regions reflects some of the difficulties that arise in attempting to use the computer to solve more general problems.

Limitations of Automation

There are a number of difficulties that limit the extent to which computers can be used to solve general cartographic problems. A number of these problems are purely technical in nature and will, in time, be solved by advances in computer capabilities. At present, problems of this type occur in all phases of the handling of digital cartographic data.

One of the advantages of computers is the ability to process huge quantities of data. Thus, it is possible to create cartographic databases with high resolution that cover large areas, such as countries and continents. However, the generation and maintenance of such data causes many problems. Much of the current effort related to cartographic data acquisition involves transferring map data that already exists in map form into digital form for computer use. This process of data capture adds error to the digital representation of the data over and above any error in the

source map and tends to be very expensive. Some of the errors in the data capture, or *digitization*, process can be detected and corrected by human operators; however, there is always some loss of information and accuracy (Jenks, 1981). Consequently, while maps created from the resulting databases can precisely represent the data in the database, the database itself will contain errors that may be significant.

The size of many cartographic databases also creates problems, including the difficulty of updating, distributing, and using the data. The problems of maintaining the accuracy of high resolution spatial information are compounded by the fact that the speed with which such information changes increases as the level of detail increases. For example, a database that holds only the major interstate highways will change relatively slowly, while a database that contains all the roads in the U.S. would be constantly evolving. The distribution of very large data sets is a complex problem independent of the type of information that is being transferred. Difficulties include the selection of transfer media, standardization of the data format, and limited channel bandwidth. A useful database with high resolution will require a substantial degree of internal complexity in order to represent the detail adequately. The complexity of the software for handling the data will be commensurate with the internal complexity of the database. Such complexity is likely to be reflected in the cost of such software. In general, the reality of a complete, accurate, generally available cartographic database with sufficient detail to serve most mapping needs remains a distant goal.

Severe limitations exist in the output capabilities of most current computer systems. While graphics terminals have evolved greatly over the past decade, they are still inadequate for high-quality cartographic output. The top level graphics terminals have resolutions of about 100 dots/inch. It is not possible to clearly

present the detail and symbolism that appears on a typical topographic map at this resolution. The color choices available on most current devices are also insufficient for many mapping applications. Devices that produce paper copy of computer maps, i.e., hardcopy devices, also have such limitations. While hardcopy devices in low and moderate price ranges have higher resolution (300 dots/inch) than graphics terminals, this resolution is still too coarse for finished work. Color selection and quality on these devices is often more limited than on graphics terminals. Thus, map symbolism is limited, in some cases severely, by current computer graphics technology. In some situations, software can be written to mitigate these limitations. However, such software is expensive to produce and, in many cases, is slow to execute on all but the most powerful computers. Very expensive output devices can overcome these difficulties. However, such devices are out of reach for most people working in cartography. These problems have affected the decisions made by cartographers using computers. For example, the new USGS 1:100,000 map series will eventually be revised from digital cartographic data. Consequently, the symbols on the maps, produced by traditional techniques, have been limited to symbols easily processed by computer, such as solid rather than cased lines for major roads (McLaurin, 1975). The current technology also tends to limit the types of maps that are produced. It is relatively easy to use a computer to plot linework; however, it is more difficult to program a computer to produce some of the other useful cartographic representations, such as dot maps, flow maps, and contiguous cartograms.

The automation of cartography has effects which extend beyond the technical limitations of this new set of tools. The careless production of maps is one such example. If a novice can simply press a few buttons to make a map, there may be a

proliferation of poor quality and meaningless maps (Lai, 1985). There are a number of ways in which automation accentuates this problem. As mentioned above, it is straightforward to implement a large suite of map projections that become available to any user with a simple menu selection. For the knowledgeable mapmaker, this is a very helpful tool that allows the creation of improved maps. However, this capability can become something of a toy for a novice. Such a user may tend to select projections at random and plot any type of information. Problems, such as displaying area-aggregate data on a projection that is not equal area, can lead to misleading maps. The ability to change scale easily on the computer can also be a mixed blessing. The user can zoom in on an area of interest to examine it in greater detail. As with map projections, this can be of great assistance to those who understand the limitations of the data being studied. However, if the enlargement is too great, then the information displayed may essentially be noise and, hence, be meaningless.

Even more dubious maps can be created when mapping statistical data or the results of a modeling procedure. A programmed statistical package can generate numbers very easily. However, the generation of valid statistics from such a package requires a clear understanding of the meaning and utility of the various statistical procedures. The combination of computer statistics and computer mapping provides the opportunity to create endless maps that have no meaning. Such difficulties are, in some ways, an artifact of the limited educational opportunities available in computer cartography. Since few users of computer mapping packages are likely to have any specialized training in their use or even a minimal geographic education, this problem may reach epidemic proportions. The computer can give the cartographer a nearly unlimited suite of techniques to use in making a map.

However, it takes experience and training to use such techniques to extract useful information. The expense of the hardware and software needed to build a basic mapping system often inhibits the educational process. While prices are constantly falling, they are still high enough to be a problem for many universities.

While education in current techniques can resolve some of the difficulties mentioned above, some related but fundamental problems exist that require closer examination even by experienced cartographers. One such problem arises when digital cartographic data produced at given scale is to be used to create a map at a much smaller scale (Brassel, 1985). The computer can be programmed to simply plot all the detailed data at the new scale. However, without an appropriate generalization procedure, the resulting map will be unreadable, as discussed in Chapter II. Related difficulties can arise in geographic information systems due to the need to process numerous data sets describing different themes for the same region. The different data layers will, in general, all have different sources and, as a result, they will have different levels of accuracy and degrees of generalization. Typically, the output from such a system will be a map that displays the data after it has been combined in some manner. While the resolution of some of the data sets may be sufficient for the particular map being produced, other data layers may be too coarse for reliable use at the map scale. Because a computer will process any of the data layers indiscriminantly, maps can be produced that are of indeterminate accuracy. It may be very difficult to assess the validity of such a product. Because of the increase in both computer mapping systems and geographic information systems, it is important to develop a solid understanding of map generalization in the computer environment. Such an understanding should include techniques for changing the degree of generalization of a data set along with tools for measuring the degree of generalization.

Clarification of the generalization process in this way would be a major step in making more effective use of computer cartography as well as geographic information systems.

Chapter IV

Cartographic Databases

Cartographic databases play a central role in computer cartography. Major flows of information in computer cartography are from observations and traditional maps to the cartographic database and from the cartographic database to new maps. Because of its crucial position in the overall process, the structure and quality of a cartographic database places fundamental limitations on the maps that can be created from it and on the operations that can be applied to such a database. Therefore, it is very important to be aware of the different types of data structures that can be used to store map data and the implications of the selection of a particular type of structure on the utility of the database. Numerous factors affect the quality of digital cartographic data and some of these factors can be artifacts of the type of data structure. These factors must be understood if proper use is to be made of the database. The following sections will consider some important aspects of these kinds of databases along with some of the difficulties associated with them.

Digital Map Representation

Traditionally, the storage medium for a map has been the paper on which it is printed. As a result, all the information contained in the map is visible to the map reader, i.e. "maps are holistic representations of spatial reality" (Visser, 1990).

other roads. It is also clear which locations on the map are inside or outside the boundary of an area. The reader can integrate different map elements, such as labels, contained in areas with both the area itself as well as the boundary of the area. As a result, it is easy to determine if areas, such as states or counties, are adjacent. These are examples of problems which, although easy for a person to solve in viewing a printed map, can be relatively difficult and time-consuming to deal with on a computer. The brain has the ability to process a portion of a map in parallel while a computer must deal with each component of a map sequentially. To solve such problems easily using a cartographic database requires both the careful selection of the information to be saved in the database and the use of an appropriate data structure. For purposes of representing the holistic nature of a map in a computer, the kinds of information presented on a map can be divided into three categories: geometry, attributes, and topology.

Geometry

The geometrical aspects of cartographic data are in many ways the most obvious information that must be entered into the computer. Geometry includes the location of points, the path of lines, and the shape of areas. While the geometry of features on a map is obvious to the map reader, the representation of the same information on the computer requires a means of encoding the graphical positions shown on the map. This requires some kind of coordinate system (see Appendix A for a brief discussion of coordinate systems).

Representing the geometry of a map in a computer involves defining some kind of planar coordinate system on the surface of the map and entering the coordinates of locations of interest into the computer. In the case of a point feature, it is possible to indicate its location with one coordinate pair. However, a line or an

area is composed of an infinite number of points. Because computers have finite storage, all the points that make up a line or area on a map cannot be represented on the computer. Numerous ways exist of approximating the path of a line or the shape of an area utilizing a finite number of points stored on the computer. Some of these mechanisms will be discussed in the next section. In general, the greater the number of points used to represent a line or an area feature, the more accurate the representation will be. Thus, the quality with which the geometry of a map will be represented on a computer is largely a matter of economics, i.e., requirements for better geometric representation can usually be met by larger, faster (and usually more expensive) computers.

Attributes

Simply storing the geometry of the features that appear on a map is insufficient to reproduce the map from the digital database. On most maps, different kinds of features are symbolized in different ways, e.g., lines can have different widths and colors. For such distinctions to be made, it is necessary to associate other information, called *attribute* data, with each geometrical entity. Such attribute data allows lines representing roads to be distinguished from lines representing streams. As a result, the attributes can be used to control a plotting program so that the roads are drawn in one color and the streams in another color. This implies that there must be some type of encoding of all the various geographic features to be represented on the map (Selden & Domaratz, 1982).

These attributes, along with the geometry, are necessary to create a usable base map from a cartographic database. The idea of associating attributes with features can also be applied to thematic mapping. This can involve associating additional attributes with a given feature over and above those required to classify

the feature. Thus, a section of highway could be tied not only to a category, such as an Interstate, and a number (e.g., 80), but to such additional information such as number of lanes, date of construction, volume of traffic flow, accident rate, frequency of traffic jams. This kind of information can be used as part of an interactive query system whereby all the information about a feature selected from a map displayed on a computer terminal is shown to the user.

In addition, such data can be used to create thematic maps. In the highway example, numerous maps describing aspects of automobile transportation could be created using a database containing the necessary attributes and appropriate symbolization. This is commonly done with administrative areas for which statistical data is aggregated, such as census districts, counties, and states. In addition to a code identifying each area, other information, such as population, ethnic breakdown, property values, and number and types of businesses can be incorporated as attributes. Any number of thematic maps can be made by classifying one of the attributes and then symbolizing the areas in an appropriate way. Additional thematic maps presenting more than one of the attributes can be created to determine if the two attributes are spatially related. A minimum amount of attribute information is required for the creation of base maps, but additional data of this type increases the flexibility and utility of the database as a geographic tool.

Topology

Given the geometry and the attributes that represent a map, it is possible to reproduce the original map. In principle, it is also possible to perform most kinds of geographic processing. However, in practice, additional data describing the map is necessary to perform certain types of geographic manipulations with reasonable efficiency. General geographic operations, such as the study of networks (e.g., a

stream or road network) and the study of coverages (a set of areas where every location in the map area is assigned to one and only one area), are greatly optimized if information is provided to describe which features are connected or adjacent to other features. In the case of a stream network, it might be necessary to find all the tributaries of a river as part of a watershed analysis. Given only geometry and attribute data describing the rivers and streams in a cartographic database, it is possible to examine the locations of all the rivers and streams in the database and to determine which streams join the river of interest. However, because the total number of such features in a database may be huge, this approach will usually require substantial computation and, in fact, may be prohibitively expensive. If the information describing the feature connections is already contained in the database, then an exhaustive search will not be necessary (Peucker & Chrisman, 1975).

It has been argued that most geographically interesting problems can be expressed in terms of coverages, which commonly occur in thematic mapping (Beard, 1987). A study region, such as a state, is often divided into administrative areas, such as counties, for which statistical data are routinely gathered. Determination of which counties are adjacent to a particular county is often necessary in geographic processing, e.g., aggregating adjacent counties with equal values for an attribute to form a larger area for presentation on a map. Coverages also arise when a geographic distribution of values for a thematic variable is classified; familiar coverages of this type include land use/land cover and geologic maps. Areas with common classifications will combine to form a coverage of the area of interest. Knowledge of adjacency and the capability to determine which areas lie on either side of a boundary is crucial to identifying patterns in thematic data of this type. It is also important for more mundane operations associated with the creation and maintenance of this kind

of data, as will be discussed later. As with connectivity information in networks, adjacency information can be derived from attribute and geometry data alone, but the cost is likely to be high. Storage of the appropriate information in the database can make these procedures much faster and, therefore, more useful to cartographers and geographers. If the complete geometry, attributes, and topology of a map are input to a computer, then a large array of mapping and analysis procedures are possible. To realize these possibilities, an appropriate cartographic data structure is required.

Cartographic Data Structures

The two most common types of cartographic data structures are called vector structures and raster structures. These two structures are very different in terms of the kinds of operations they support most effectively (Goldberg, 1984; Walton, 1985). Discussion of these storage formats requires some computer terminology which is introduced in Appendix B.

Vector Data Structures

In vector data structures, all features are defined in terms of lists of one or more coordinate pairs. Thus, a point feature is defined by a single location. A linear feature is defined by a list of locations that define the path of the line. An areal feature is indicated by a list of locations that define its boundary. To form a representation of a feature, each location in the list is connected with some type of curve. Most commonly, the curve is a straight line (a *vector* in computer graphics parlance, which is distinct from the notion of a vector as used in computational geometry), so drawing a feature is basically a matter of connecting the dots that represent the locations stored in the list. It also is possible to connect the points with more complex curves such as parabolas, splines, or higher order polynomials;

however, a great many current graphics devices automatically transform coordinate lists into straight line vectors while only the most expensive devices support more complex curves. In this study, locations will be connected with straight line segments.

In a vector data structure, a location is typically stored with each component of the coordinate pair assigned to a word (typically two or four bytes) of computer storage. Thus, two words are used for each location. These two words, which represent a location, could be stored in a record. A list of locations then might be stored by creating a sequence of location records in the computer memory, i.e., a singly-dimensioned array of records. This is illustrated in Figure 4-1 for two linear features. A given record in the array is accessed with an index, or pointer, that indicates where in the list the particular record resides; thus, each point in the list could be plotted by setting an index to point at the first element of the list and passing the first coordinate pair to a plotting device along with an appropriate instruction. The index would then be increased by one so that it pointed at the second coordinate pair that would also be passed to the plotting device. This would be continued until all the coordinate pairs had been plotted.

Generally, attribute information for a map is also stored in record structures. Each cartographic feature in the database would typically be associated with a record. Such a record would contain information describing the feature. Many types of data could be placed in this record and the choice of data depends strongly of the kinds of uses envisioned for the database. A very simple record might consist of a single attribute field that would contain a code indicating the classification of the feature. This would allow an appropriate symbol to be associated with the feature when a map was plotted. Other information that could be included in such

Figure 4-1. Two linear features and their vector representations.

a record might be the name of the feature, the minimum and maximum coordinates of the feature, and the feature's length. The name could be used to label a map, the coordinate ranges can assist fast plotting from a large database containing many features that do not appear on a given map, and the length can be used as part of a selection procedure in some cases. Many cartographic databases, which are available from various sources, are in this form or in a variation of this form. A record containing the attributes of the feature is followed by a list of location records defining the geometry associated with the feature. This sequence is repeated for each feature in the database. Such a structure has the advantage of simplicity and is adequate for the most basic mapping processes. Figure 4-2 shows a simple example of attribute data added to the geometry.

In addition to information that assists in the efficient production of maps, the feature record may also contain thematic attributes of interest. The feature

Figure 4-2. Two linear features with a code and a name added as attributes. In this example, a river has a code of 100 and a road has a code of 200. The addition of the code allows the features to be symbolized differently. The name permits features to be labeled.

record for a city might, for example, include fields for the population or the number of high-rise buildings. The association of attribute information, which is not

~~inherently spatial in nature with geographic locations or features in this way is~~

data that is used only rarely. Also, extending the capabilities of the database with new attributes could result in frequent alteration of the record structure.

One solution to such problems relies on a unique index, or *key*, that is associated with every feature in the database. Utilizing this key, it is possible to decouple the individual attribute fields from the file(s) containing the basic feature records and the feature geometry. For example, each thematic attribute could be maintained in a separate file or *table*. Such a file would contain one record for each feature that was associated with the particular attribute, and each record would consist of the key for a feature as one field along with another field with the attribute value for that particular feature. Figure 4-3 provides an example. In producing a thematic map, the geometry for a feature would be associated with its attribute value by reading the file containing the basic feature records, and determining the feature code, geometry, and the key of the feature. Then the attribute file of interest would be searched using the key. Given the geocode, geometry and the newly found attribute for the feature, the necessary decisions regarding symbolization could be made. This mechanism of handling attributes can be extended to handle very complex attribute schemes by allowing table entries to be keys into other tables. These additional tables may contain data on a variety of subjects that do not have to be tied directly to the original cartographic database and, as a result, the amount of redundant data in the overall structure is minimized. Such a complex data structure is called a *relational* database which is also illustrated in Figure 4-3 (van Roessel, 1987; Waugh & Healey, 1987).

Given a vector representation of a map incorporating geometry and attributes, many kinds of map processing can be completed. However, the vector encoding still contains no explicit information about connectivity or adjacency. This could only be

Figure 4-3. A unique key (the ID field) has been added to the record for each feature. This key points into a table for additional attribute data which is, in this case, the classification of each road. Road classification, in turn, is used as a key into another table which contains the speed limit as a function of road classification. The use of a field, such as road classification, which exists in two different tables to connect, or join, those tables is the definitive characteristic of a relational database.

derived from an exhaustive examination of all the features and many of the locations that compose the features. As a result, problems which require the determination of topological relationships can be so computationally expensive that they cannot be handled. The solution is to add explicit topological information to the database. In the context of mapping, the use of topology involves the use of zero-, one-, and two-dimensional cartographic objects. Zero-dimensional objects are called *nodes* and indicate a junction or end point. A one-dimensional topological object is called a *chain*. It consists of a directed sequence of non-intersecting line segments with nodes at each end of the chain. A sequence of chains can be used to define a ring if none of the chains intersect and the sequence of chains closes, i.e., if the sequence of chains begins and ends at the same node. A two-dimensional object is referred to as a *polygon*, which includes an area along with a bounding ring along the outside of the area and zero or more non-intersecting, non-nested rings that define holes in the area. Other variations of zero-, one-, and two-dimensional entities can be

defined, but the objects mentioned above describe the basic topological concepts (Morrison, 1988).

Several points need to be made about these topological objects and their application to mapping. Nodes bound chains and chains bound polygons. Thus, zero-dimensional objects bound one-dimensional objects and one-dimensional objects bound two-dimensional objects. In a network, chains are connected by nodes. In a coverage, adjacent polygons are separated by chains. These relationships define the topology of a map. They are independent of the particular shapes or paths of the polygons and chains. The relationships between the polygons, chains, and nodes are maintained when the map is stretched, squeezed or distorted in any way that does not involve cutting or folding the surface. This invariance under distortion is the defining characteristic of topology.

A further important point is that geographic features do not form a one-to-one pairing with topological objects. Nodes are distinct from cartographic points that indicate a geographic feature. Although a node may specify a geographic location, this is not a requirement since a node exists mainly to form an essential part of the topology and the topology is independent of the choice of coordinate system. A chain can be described by a sequence of locations that specify the shape of the chain and by two nodes which are associated with the start and the end of the chain. Thus, a chain connects two nodes with a direction defined from the start node to the end node. However, the actual locations that described the chain are not important to the topology; only the connection between nodes and the direction are necessary. A linear geographic feature can be described by one or more chains. For example, a road will be divided up into separate chains at each intersection because chains are, by definition, non-intersecting.

There are numerous ways to store topological information in a cartographic database. All these methods rely on attaching an identifier, either explicitly or implicitly, to every topological object. Frequently, each node, chain, and polygon is numbered and that number is used to reference the object. On the other hand, in structures that store the objects in arrays, the array index can be used as the identifier without being maintained in the object record. The selection of a method is largely a function of the intended use of the data. For example, one common format, the USGS Digital Line Graph (DLG), associates all the geometric and most of the attribute information with chains (Luman, 1987). The identifiers are incorporated as explicit fields in the object records. Each chain includes the identifiers of the start and end nodes as well as the identifiers of the polygons to the left and to the right of the chain. The nodes and areas, which mainly serve a topological function, are relatively minor components of the overall database. Such a structure minimizes redundant information and, as a result, is relatively compact. However, efficiency in processing can be gained if nodes and areas contain more information, such as the chains that connect or bound them. This will simplify certain processes (e.g., tracing a network or finding chains that bound a specified polygon) at the expense of increased storage. Decisions during database design should reflect the relative importance of speed and space to the application. An example of a topological representation of three states and a surrounding background is illustrated in Figure 4-4.

Raster Data Structures

The other major class of cartographic data structures is called *raster* structures. In these data organizations, the geographic space of interest is divided into a rectangular mesh of grid cells. In the computer, this mesh is normally represented

Figure 4-4. The topological structure of three states showing: (a) The nodes, chains, and areas (area A4 is a background area which encompasses all features in the database), (b) the chain data (each chain has pointers to its bounding nodes, the areas being bounded, and a list of coordinates that define the chain), (c) the area data (each area includes a list of the chains that bound it as well as the name of the area), and (d) the node data (each node has a list of the chains to which it is connected).

by a two-dimensional array. One array is used to store the values associated with each grid cell for a specific attribute. The value for a given grid cell is accessed by the use of two indices that indicate where in the array the associated value resides. One index represents the selection along the x , or easting, axis of the mesh of grid cells and the other index represents the position along the y , or northing, axis of the mesh.

The term “raster” is derived from the way images are shown on a television screen. Here, the image is divided into a sequence of scanlines that range from the top of the screen to the bottom. Along each scanline, or raster, the intensity of the beam is modulated to create small regions of varying brightness or color. The smallest region for which the brightness or color can be controlled is called a picture element or *pixel*. The pixel of a television raster corresponds in a conceptual way to an individual grid cell in a rectangular mesh where the grid cell is the smallest portion of geographic space for which specific data is kept.

A map is normally encoded in raster structure by a number of arrays that match the geographic mesh. Each array stores the values for one thematic overlay, or feature type, on the map. Thus, an element in a given array contains the thematic value for the grid cell that is associated with the element. For example, one array might contain an elevation value for each grid cell. Another array could store land use/land cover categories for same grid cells. Important linear data, such as a road network, can also be stored in such a raster structure. An array could store ordinal values where a value of zero would indicate a grid cell that does not contain a road; larger integers would indicate where various categories of roads exist in the grid. Linear features can be poorly represented in a raster structure, particularly when the size of a grid cell is large. If a continuous road network is required, then

each road must be represented as a string of adjacent grid cells. In a coarse grid, such grid cells could be one or more kilometers on a side and, thus, be much wider than any road. On the other hand, with a high-resolution grid having a cell size of only few meters, a road may be represented quite accurately. Thus, the accuracy of the geometric representation of a map in a raster structure is a function of the size, or resolution, of the grid cell. As the resolution becomes finer, i.e., the grid cells become smaller, the geometry is portrayed more precisely. However, if a given area must be covered by the raster database, the amount of storage required in the computer increases rapidly as the grid cells get smaller.

Attribute data is handled naturally in a raster structure. A value for each attribute is determined for each grid cell. These values are stored in an array. Normally, each attribute is maintained in a separate array in the computer. Because the grid has spatial relationships built into the array representation, topology is not usually isolated as separate information. Determination of adjacency and connectivity can usually be determined using basic grid operations. The many kinds of operations that can be performed on raster data have been summarized by Berry (1987).

Comparison of Raster and Vector Structures

Most kinds of geographic information are represented naturally in either the raster structure or the vector structure, but rarely both. For example, linear data, such as a stream or road network, can be easily and compactly stored in a vector data structure. On the other hand, terrain elevation values are simply expressed in a grid structure. Each structure also supports certain geographic processes more effectively. Plotting linear base map data is done easily and quickly from a vector

data structure. In large part, this is due to the fact that each line object is represented only by the locations that define its path, whereas in a raster structure, every element in the array must be processed to draw the map and processing the additional information will slow the procedure down. In a vector structure, the overlay of two polygon coverages representing different thematic variables is a difficult process because, in general, the two sets of polygons will not match and the two coverages must be intersected to create a new coverage with unique values for each polygon. In a raster structure, the same polygon overlay process is trivial; it involves simple logical or arithmetic operations on matching array elements. Thus, the choice between raster and vector data structures is based on the kinds of data to be represented and on the kinds of geographic processing that will be required. In many cases the best choice is unclear and, as a result, research is proceeding on various alternative and hybrid data structures that will combine the strengths and avoid the weaknesses of the basic vector and raster structures (Peuquet, 1983). For this study, vector structures are the most applicable and raster data will only be mentioned in passing in subsequent chapters.

Overlays and Tiles

Because of the quantity of data which may be maintained in a cartographic database, it is often impractical to store all the information in a single file. Various mechanisms are used to divide up cartographic data into manageable portions. One common technique for structuring such databases is to group all information regarding a major class of features into a separate file or *overlay*. For example, a database might group features into overlays for transportation, hydrography, and political boundaries. Each overlay would be stored in a separate file. This allows

one file. This can simplify and accelerate the processing. The larger cartographic databases often have so much information that the subdivision into overlay files is insufficient to produce files of manageable size. As a result, databases are often split up into files that cover smaller geographic areas. These component files are sometimes called *tiles*. For example, separate files might cover one state of the U.S. or a one degree quadrangle. This can be done in a manner analogous to existing topographic map series which are commonly divided into quadrangles of a size appropriate to the scale of the maps. This division can cause problems if the database files for adjacent maps do not match. This can seriously limit the utility and flexibility of the database. A better approach is to allow features in adjacent map files to point to each other. This requires that the geometry and the topology match at the boundary. While substantial effort is often necessary to produce such a “seamless” electronic map, the additional quality and utility of such a database can often make the work worthwhile.

Data Capture

The information that is placed into cartographic databases comes from many sources. Increasingly, the data is processed directly from surveys and other observations. For example, with relatively minimal processing, LANDSAT data can be input into a raster database as a separate thematic overlay. Straightforward mathematical transformations allow the reflectance values to be determined for grid cells on the ground from the pixel values in the LANDSAT scene. However, most data, even that from remote sensing satellites, currently requires substantial human interaction to derive the necessary map information. Thus, many applications would require a LANDSAT scene to be processed to create a land use/land cover overlay and this usually requires the information processing capabilities of people. While

the direct input, or capture, of observational data for cartographic databases is an area of focused research and development, the bulk of data capture processing involves the transformation of existing maps and photographs into computer form, i.e., digitizing.

Tablet Digitizing

Manual digitizing systems have been the workhorse of most data capture systems (Marble & Lauzon, 1984). While a number of approaches have been used in the past, current systems have generally settled on free cursor systems that use electromagnetic technology to locate the cursor with respect to the surface of a tablet. In a typical operation, a map is placed on the digitizing tablet and the cursor is used to locate items on the map. A computer graphics system is often integrated into the digitizing system to provide feedback to the operator. The first step in the manual process is the preparation of the maps to be digitized. This step may include acquiring the needed maps, determination of what information will be extracted and identification of control points to allow the map to be registered on the tablet. The map may be annotated to assist in these processes. The map is then placed on the digitizing tablet and the control points are identified with the cursor. Given the geographic locations of the control points and their corresponding tablet coordinates it is possible to define a transformation from the cursor position at any point on the tablet to geographic coordinates (or some map projection grid). Once the preprocessing and registration steps are complete, the digitization process can proceed. Points on the map are located by placing the crosshairs that exist on most cursors over the point of interest and pressing a button on the cursor. The digitizing system will then store the location of the point in an appropriate data structure.

There are two common modes of entering linear data. One method is to have the operator identify a series of points along the arc that define its path. In point mode digitizing, points are digitized only where the arc changes direction; a relatively modest number of points can be used to define the feature. The other mode of line digitizing is stream mode. In stream mode the operator traces the line by moving the cursor crosshairs continuously along the line. The system samples the location of the cursor and records locations at either equal time intervals or whenever the cursor moves a given distance increment. Stream mode digitizing tends to produce large quantities of location data. Much of the data is redundant, e.g., hundreds of points may be used to represent a straight line that can be described with equal accuracy by two points. On the other hand, point mode digitizing relies on the person digitizing to make an appropriate selection of points to define a feature which, in turn, depends on the person's understanding of the cartographic requirements and experience.

Most manual digitizing systems allow the digitized locations to be displayed on a computer screen as they are being collected. This allows the operator to detect obvious blunders and to correct them before they become incorporated into the main database. This may be the first in several steps of checking and editing the data. An additional quality-control step is to plot the data at the same scale as the original map on clear plastic and to place the plot over the map. If done carefully, this can identify most significant locational errors. The correction of many of these errors is usually accomplished using interactive graphics software that allows manipulation of the locational data already in the computer. Redigitization is usually avoided unless absolutely necessary.

In some cases, the raster data must be converted into attributed, topologically structured vector data to be most useful. Software exists to identify the pixels that make up the line work of the map. However, current software has difficulty handling certain problems such as line interruptions for labelling and road intersections. A person must examine each problem area and interact with the software to complete the various vector elements. Once the vectors have been extracted from the raster data, software similar to that which is often used in manual digitizing systems is utilized to define the topological structure of the vectors and to tag the various features with the required attributes.

Manual and scan digitizing are applicable in different environments. Raster scanning is important for large organizations that need to capture data from large numbers of maps and also require all the information contained on each map. It can also be the more natural mechanism for input into raster data structures. The time required to completely and accurately capture a map in digital form is minimized by this technique. For smaller organizations with limited needs and limited budgets, manual digitizing is often more cost effective. When only a small portion of the information on a map must be captured and only a modest number of maps must be processed, the expense of raster scanning systems may make them impractical. The large amount of post-processing may also be a disadvantage for simple applications. Scan digitizing also works best when map separates are used which may not be practical for small operations. These systems satisfy needs at opposite ends of the spectrum of data capture environments. Other types of current and future techniques may fill some of the middle ground (Selden, 1987).

Design Considerations for Cartographic Databases

Numerous factors must be considered in designing and dealing with cartographic databases. Some of these factors are common to most any process of building and using a complex technology-based system. Others are more specifically related to aspects of cartography. Some of these are considered below.

Matching Structure to Function

In database design, the basic problem is to identify the operations that need to be performed and the data that can support these operations. If this process is incomplete, then the structure developed for the data may be either too simple or too complex for the required functionality. An overly complex design may result in unnecessary storage and processing requirements along with undue difficulty in understanding and maintaining the structure. Perhaps the simplest set of requirements for a vector cartographic database would be to produce a base map including such features as coastlines and political boundaries on a monochrome graphics device for familiar areas at a standard scale. In such a case it may not be necessary for any information, other than the geometry, to be stored. Simply drawing the lines in the database with the same symbolization, regardless of feature type, may be sufficient to produce a recognizable image in some areas. The only information in such a simple database would be strings of coordinate pairs without any attributes. Such a structure is often referred to as a *spaghetti* format and it can only be used for the most trivial mapping applications. While such a structure may be selected if the operational requirements are very simple, this type of structure is not easily extended to more complex operations. One very challenging aspect of database design is to consider not only the current set of requirements but to envision the effects that future development may have on mapping requirements. If significant

extension of capabilities is at all likely, then more robust designs may be appropriate because they would support both the current and many of the future requirements.

The addition of attribute data to the geometric information permits additional mapping operations. For example, if codes are associated with the various coordinate strings, then the various types of features can be symbolized differently, thereby producing a higher quality map. If attributes are associated with a set of polygons covering an area, then the polygons can be filled with different colors or patterns on the basis of the attribute values producing a choropleth map. Given a

coverage, such as the states in the U.S., it is possible to create a precisely matched set of polygons that form the coverage and to use this set to produce the required maps. Each polygon is a line which begins and ends at the same point; one such exists for each state. Such a structure maintains each boundary twice, e.g., the border between Texas and Oklahoma exists both in the Oklahoma polygon and in the Texas polygon. Such a structure is shown in Figure 4-5. This simple organization permits easy plotting of choropleth maps using data aggregated by state. However, if the polygons of such a coverage ever change, then the effort required to update the polygons would be substantial because all the boundaries between adjacent polygons are maintained in two unrelated places in the database. Thus, any editing must occur twice. If such editing is not done carefully, then splinter polygons and other related errors can occur. Such problems can be avoided if the appropriate topological structure is used. If each string between two adjacent polygons is maintained as a separate entity with information indicating which polygons the string bounds, then the geometry of any boundary exists at a unique location in the database. As a result, a change in the boundary can easily be made by replacing

Figure 4-5. Three areas maintained as separate polygons. Note that the common boundaries are stored twice.

the old coordinate string with a new one. Thus, the structure of the database determines how easily it can be used and maintained as well as the level of functionality supported.

Another example of the close ties between structure and function in a vector cartographic database is related to line simplification as part of the generalization process. If lines on a map could be considered as independent entities, then each line could be expressed with the minimum number of points required for the scale

and purpose of the map. The simplification would only be a function of the shape of the line. However, most lines on a map are *not* independent of other lines and other features. Roads intersect other roads, cross streams at bridges, and often terminate at other roads. Streams merge to form larger streams, and rivers drain into lakes and oceans. Consequently, the independent simplification of a line may cause such connections to be violated if the geometry of a line is considered in isolation. Since such connections often form a crucial component of the recognizable pattern of a

portrayal are maintained. This implies that lines must be simplified between nodes that will exist on a target map and that the nodes themselves cannot be moved or omitted. Thus, if generalization associated with scale change is a requirement for a cartographic database, then topological information should be included in the data structure.

Resolution and size

If a cartographic database is to be handled easily, the resolution of the information in the database must be consistent. That is to say, use of a database will be facilitated if the level of generalization is comparable for all the information in the database. In a vector database, the resolution largely corresponds to the spacing of the points used to represent a line or the boundary of an area feature. If the points are closely spaced, then most of the details of a feature's shape should be preserved and the representation will be of high resolution. If the points are far apart then details of the feature's form must be omitted and the shape will be stored at low resolution. If the feature is to maintain a recognizable shape as fewer points are used to represent it, then some strategy must be developed for identifying the points that contribute most to the perception of form. It is very important that the resolution be consistent for each feature. If it is not, then the portrayal of the features at almost any scale will be confusing and unappealing. Figure 4-6 shows an example of two reservoirs south of Livermore, California which have been digitized at significantly different resolutions. This figure demonstrates that a digital cartographic database should be associated with a particular scale and all the features which make up the database should be represented at a resolution consistent with that scale.

Figure 4-6. Two lakes digitized at very different resolutions. Variable resolution such as this can create problems in the display and processing of cartographic data.

These statements imply that the level of generalization in the representation of a feature must be consistent with the scale at which it will be reproduced and that such a degree of generalization will only be appropriate for a limited range of scale changes. In general, map data should not be displayed at a larger scale than the

Figure 4-7. Data digitized at 1:2,000,000 scale at high-resolution are plotted on a 300 dot per inch printer at the same scale. Notice the clutter of information along the southern shore of Lake Michigan. On the original maps in the National Atlas of the United States of America (1970), the same information is presented with distinguishable symbolization along with other data such as labels. The higher resolution used to create the original maps supports a density of information that cannot be easily maintained with lower resolution plotting devices.

may actually be somewhat smaller than would be expected due to the device used for plotting the maps. For example, if all the data from a typical color map is captured accurately with a high-resolution scanning device, all of the digitized data cannot be clearly displayed on a low-resolution black-and-white printer. This does imply some limitation on the usefulness of much of the information in cartographic databases for many computer mapping operations which can only support low-cost printers. Figure 4-7 shows four overlays from the USGS 1:2,000,000 DLG data at the intended scale plotted on a 300 dot per inch laser printer.

In thematic mapping, data, which has been over-generalized to the point of caricature, has often been used to provide a base map. For example, a caricature of state boundaries may be appropriate as a base map for thematic data aggregated by state since the thematic material will not be overshadowed by unnecessary detail (Dent, 1985). It may be possible to use digital cartographic data at larger scales than intended as a source of caricature for such base maps. However, in such an application it may be safer to use greatly over-generalized rather than slightly enlarged data which might give the impression of being an accurate representation of physical features on the earth. Jenks (1981) has suggested that simplifications of digital map data, which are obviously different from the portrayal at full-resolution and yet clearly recognizable as simplifications of the original features, are good candidates for use as base maps for thematic material. Perhaps this can be extended to enlargements of digital map data beyond their intended scale.

If the displayed information is integral to the purpose of the map, enlargement or caricature must be avoided. As scale is decreased from the production scale of the data, a map will tend to become more and more cluttered and the linework will become increasingly jittery and unreadable. The limited scale range over which vector cartographic data is useful is a truly fundamental problem of database design. One possible solution to this problem is to maintain numerous databases for an area with each database designed for use over a small range of scales. The various scale ranges would have to overlap to some degree to allow maps of any scale to be created. These databases would form a hierarchy with the lowest level being associated with the most detailed information which would ever be required, e.g, 1:25,000. Without processing, this data might serve for maps with scales as small as 1:40,000. The

next level might serve for maps from 1:40,000 to 1:65,000. Such an arrangement could, in principle, be extended to cover all the scales of interest.

However, many practical matters limit the viability of such an approach. One such matter is the size of all the databases that would be required. If any significant area is to be covered at the relatively small scale of 1:2,000,000, large quantities of information must be captured and stored. For example, USGS now has DLG coverage of the conterminous U.S. at two scales. 1:2,000,000 DLG data has been available for some time and requires about 40,000,000 bytes (40 Megabytes) of storage. For a newer database at 1:100,000 scale, the same coverage requires about 25,000,000,000 bytes (25 Gigabytes). Given these examples, it should be apparent that the cost of processing and storing many such databases to cover a range of scales is a very large task. Extension of such databases to cover the globe clearly would make the problem much worse. At present, only the largest systems could handle the volumes of data required to implement this method for handling widely varying scales. This approach might be manageable for a substantial range of scales for small and very small scale maps; however, other methods for handling digital map data must be considered for medium and large scales. Another difficulty with such a hierarchy is the problem of matching the different levels when each level is compiled separately. This can result in a map created at the small-scale end of one level's ranges looking very different from a map of only slightly different scale that is produced from the large-scale end of the adjacent range. The generalization of the original source material for each level may be substantially different. As a result, maps at a specified scale derived from two adjacent levels may have significant discrepancies.

Another approach to maintaining cartographic data is to capture data once at the largest scale that will ever be used. Thus, only one large database is created

and maintained. Maps for other scales are produced from this master database by selection and simplification of the information contained in the master database. For this method to succeed, the data structure used must include an appropriate attribute coding scheme as well as certain topological information. In addition to limiting data capture and database maintenance to one set of data, this approach can avoid the inconsistencies that arise from map data compiled at different scales. All the information is ultimately derived from the same source so consistency can be maintained if appropriate techniques are utilized in generalization. Resolution of the difficulties associated with using large-scale databases to produce small-scale maps is of great interest to mapping agencies such as DMA (Langran, 1984). Data aggregation and generalization has also been identified as a major problem facing the development of geographic information systems (Boyle, Dangermond, Marble, Simonett, & Tomlinson, 1983).

A detailed examination of all aspects of this approach is beyond the scope of this study. Consideration of selection, simplification, classification, symbolization, and induction as an integrated process is well beyond current capabilities. Thus, individual aspects of the problem need to be chosen which are amenable to careful examination. In this study, the simplification component of generalization will be the area of focus. More specifically, line simplification will be considered in the context of vector cartographic databases.

Chapter V

Line Simplification Algorithms

A great deal of literature exists on all aspects of generalization and its implementation on computers (Shea & McMaster, 1989; McMaster & Shea, 1988). This is to be expected since, as has been suggested above, generalization is one of the fundamental components of the mapmaking process in both manual and automated methodologies. Progress has been made in many areas; however, the subject that has been most amenable to computer implementation has been that of line simplification in vector databases. In addition to the relevance of line simplification to scale change as part of generalization, line simplification has a number of other very practical advantages in the computer environment. These include faster plotting, reduced storage requirements, and more rapid processing functions such as vector-to-raster conversion and mensuration techniques (McMaster, 1987a). Numerous approaches to this problem have been examined and dozens of algorithms have been developed. Some of the more important and accessible algorithms will be discussed below. While this section will attempt to provide a representative view of the available techniques, many techniques of interest will only be mentioned in passing. Those approaches that are implemented as part of this study will be described in the greatest detail.

Classification of algorithms

Because of the large number of line simplification algorithms that have been described, classifying these techniques is essential before a general discussion can

take place. One useful classification considers the mathematics and geometry on which a technique is based to categorize algorithms into nine basic types, along with an additional miscellaneous category. These nine categories are: point selection, low pass filtering, angle detection, digital elevation model (DEM) smoothing, tolerance bands, point relaxation, domain transformation, mathematical fitting, and epsilon filtering (Zoraster et al. 1984).

A number of these groupings can be eliminated from the present study for various reasons. DEM smoothing is a method for simplifying contour lines that are derived from a DEM (Chen, 1987; Monmonier, 1983). The basic technique is to smooth the DEM before the contours are generated. This results in simpler contour lines that are appropriate for a map at a reduced scale. While this is a valid technique, it is only useful for contour lines generated from a DEM. It cannot be easily adapted to other map data and, because this study is mainly concerned with other kinds of cartographic lines, this method will not be considered further.

A common example of a domain transformation procedure is the Fourier transform. In this technique, a function (i.e., a given line) is represented by a weighted sum of sine and cosine terms of increasing frequency. Once the original function is transformed, the smaller details of the lines path can be removed by eliminating the high frequency terms from the Fourier series. This technique has a number of problems, the most telling of which is that this procedure only works on closed curves. Since many of the lines available in cartographic databases are not closed, and especially because geocoding and topological considerations suggest that many closed curves should be broken up into lines that are not closed, this technique also will not be considered further.

Two other classes of algorithms will also be eliminated from further study. Mathematical fitting can include at least two approaches. The first attempts to describe a smooth function which fits a set of points exactly. Splines provide an example of this type of technique. While such a function provides a smooth representation of the set of points which may be useful for certain display purposes, it does not eliminate points or in any other way simplify the line. Approximate fits can represent a sequence of locations with less data than the original sequence. However, techniques such as least-squares are not easily applied because an appropriate function must be selected as a model. There is, at present, no solid basis for making such a selection with cartographic data. Other approximate fitting techniques have similar difficulties (Vanicek & Woolnough, 1975; Pavlidis, 1983; Sampson, 1982).

The epsilon filtering approach encounters some practical difficulties. In this technique, a circle with a selected radius is rolled along the line being generalized. The generalized curve results from tracking the circumference of the circle along the line (Perkal, 1965; Rechlin, 1972; Chrisman, 1983). Features that are narrower than the radius are eliminated because the circle cannot roll into an indentation which is narrower than it is. One major difficulty with this technique is that a drastically different result is obtained from filtering the two sides of a given line. This is because indentations on one side of a line are bumps on the other side and a circle can roll over bumps which are narrower than it is. There is no obvious basis for choosing which generalization is more appropriate. In addition, the implementation of this procedure on a computer is not obvious. This algorithm has been programmed as part of a commercial software package and is not described in detail in the open literature.

Figure 5-1. Retention of every third point of a short chain.

The remaining categories (point selection, low pass filtering, angle detection, tolerance bands, point relaxation) will be examined more closely. A miscellaneous technique related to the fractal dimension of cartographic lines will also be included.

In the discussion of the techniques, a cartographic line will be a set:

$$P = \{p_i = (x_i, y_i) | i = 1, \dots, n\}, \quad (5-1)$$

where (x_i, y_i) are the two coordinates of the point p_i , and where n is the number of points which delineate the cartographic line.

Description of algorithms

Point selection

The simplest approaches to line simplification are typically point selection methods (Tobler, 1964; Rhind, 1973). Such methods employ a simple mechanism for selecting the points to be retained from the original line and make no attempt to examine the shape or any other characteristic of the line. The classic example is the retention of every k^{th} point from the original sequence. To ensure that topological connections are maintained, it is common to start at the first point in the chain and always include the last point, even if it would not normally be selected. Thus,

$$P_s = \{p'_j = (x_i, y_i) | i = (j-1)k + 1, j = 1, \dots, m-1; p'_m = (x_n, y_n)\}, \quad (5-2)$$

where $(m-1)k + 1 < n$, P_s is the simplified chain, and m is the number of points in the simplified line. A simple example is shown in Figure 5-1.

The main advantage of this procedure is that it is extremely simple, easy to implement, and fast. However, because no attempt is made to identify cartographically important points, gross geographic misrepresentations are possible, such as not including the corner point of a right angle boundary as part of a political border. Another approach that falls in this category is to rely on a random number to select one of the points from each sequence of k points along the line. This can avoid some of the artifacts that result from equally spaced points along the line. These algorithms are mainly useful in removing points from lines that have been

application of these methods to generalize cartographic lines is risky as points are selected without any consideration of the geographic features being represented. An additional difficulty is that the choice of k is not obviously related to a requirement for generalization and must usually be selected by a knowledgeable user.

Low pass filtering

Low pass filtering techniques include any weighted or unweighted averaging of the coordinates describing the line (Tobler, 1966). Moving averages include the

Figure 5-2. The average of three equally weighted points in a window is shown at (x'_i, y'_i) . Greater weighting of the central point, as in Equations 5-4, moves the average to (x''_i, y''_i) which is closer to the central point, (x_i, y_i) .

can affect the degree and nature of the smoothing. For example, if the new points are determined according to

$$\begin{aligned} x''_i &= \frac{(x_{i-1} + 3x_i + x_{i+1})}{5}, \\ y''_i &= \frac{(y_{i-1} + 3y_i + y_{i+1})}{5}, \end{aligned} \quad (5-4)$$

rather than Equation 5-3, then the smoothed point will be somewhat closer to the central point of the window. Notice that the denominator in the equations is equal to the sum of the weights. This must always be true if useful results are to be obtained from a smoothing function. The change in the position of the averaged point which results from the use of these two different weighting schemes is illustrated in Figure 5-2.

If all the $x_i, i = 2, \dots, n-1$ are processed using these averaging equations, the line is smoothed but there is no decrease in the number of points. In contrast, if non-overlapping, but adjacent, windows are used, then the number of points is reduced. One averaged point represents all three of the points in the window. So with the equations shown above, then the resulting line will have one-third the number of points of the original, i.e., a substantial simplification will have occurred. The number of points in the window can be increased to allow for greater simplification.

In general, the following formulae can be used, assuming equal weights,

$$\begin{aligned} x'_i &= \frac{\sum_{j=i-l}^{i+l} x_j}{k}, \\ y'_i &= \frac{\sum_{j=i-l}^{i+l} y_j}{k}, \end{aligned} \quad (5-5)$$

where $l = (k - 1)/2$ when k is odd.

$$\begin{aligned} P_s &= \{p'_1 = (x_1, y_1), \\ p'_j &= (x'_i, y'_i) | i = (j - 2)k + \frac{(k + 1)}{2} + 1, j = 2, m - 1 \\ p'_m &= (x_n, y_n)\}, \end{aligned} \quad (5-6)$$

where $(m - 3)k + \frac{(k + 1)}{2} + 1 < n$, and x'_m and y'_m are defined as in Equations 5-5. This formula applies when the width of the window, k , is odd. Minor modifications are required if k is even. By varying the window width and the weights, line features of various scale lengths can be either highlighted or removed. It is most common to eliminate small variations; however, large variations can also be removed. In this study, the width of the window is varied to control the degree of simplification and the weights are always equal to 1.

These methods are fast and simple to implement. However, as with point selection algorithms, no attempt is made to identify the presence of important characteristics along the line and, as a result, important features may be smoothed away along with insignificant ones. A selection of the window size and weights must be made, and such a choice is not obviously related to cartographic requirements for generalization.

Angle detection

Angle detection algorithms attempt to isolate important points along a digital curve by assuming that points at which the direction of the line changes substantially are perceptually the most important. The association of high information content

Figure 5-3. The Attneave cat represented with 38 points of high angularity connected by straight lines.

with large changes in direction was initially noted in psychological research. Attneave (1954) demonstrated that a sleeping cat could be represented efficiently by connected 38 points of high angular change with straight lines. Figure 5-3 shows the well-known Attneave cat and illustrates the effectiveness of such points in generating caricatures of complex shapes. This notion has been discussed in the cartographic literature regarding the identification of the critical points which are most characteristic of cartographic lines (Marino, 1979).

A simple approach (Rosenfeld & Johnston, 1973) is to describe the angle at each vertex, i.e., at each point along the chain except the endpoints, with a value called a direction cosine that is chosen so that it varies from -1 for a vertex which forms a straight line segment with no bend to $+1$ for a vertex where the chain doubles back on itself. Thus, large values for the direction cosine indicate highly angular vertices. To define the direction cosine for a vertex located at some point, p_i , along the chain, two vectors must be defined. One of these vectors goes from p_i to the previous point in the chain, p_{i-1} . The other goes from p_i to the subsequent point in the chain, p_{i+1} . Thus, from Equation A-2 (see Appendix A),

$$\begin{aligned} \mathbf{u}_i &= (x_i - x_{i-1}, y_i - y_{i-1}), \\ \mathbf{v}_i &= (x_i - x_{i+1}, y_i - y_{i+1}), \end{aligned} \tag{5-7}$$

Figure 5-4. Two vectors defined by three adjacent points along a chain.

Figure 5-5. A simple threshold for α_i of 120° is used to simplify a short chain. The original chain is on top, and the simplified chain is the heavy line at the bottom.

where u_i and v_i are two vectors mentioned above. These vectors are shown in Figure 5-4. Using these vectors, a direction cosine can be defined from Equation A-7:

$$c_i = \frac{\mathbf{u}_i \cdot \mathbf{v}_i}{\|\mathbf{u}_i\| \|\mathbf{v}_i\|}. \quad (5-8)$$

When expressed in terms of the coordinates which make up the chain using Equations 5-7, A-8, and A-3, the following formula is obtained:

$$c_i = \frac{(x_i - x_{i+1})(x_i - x_{i-1}) + (y_i - y_{i+1})(y_i - y_{i-1})}{\sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}}. \quad (5-9)$$

When computed in this way, small values for c_i , close to -1 , indicate that the line is continuing in approximately a straight line. A line can be simplified by leaving out points with c_i less than some threshold value. For example, if the threshold is chosen to be -0.5 , then all points, x_i , with vertex angles, α_i , that are greater than 120 degrees will be eliminated as in Figure 5-5. Here, α_i is measured from u_i to v_i .

Such a basic approach will simplify the line; however, it will tend to select small, sharp variations in line direction as significant points and will miss less extreme variations of more global importance. This is because the angularity is only examined in the narrowest region around the point. A more sophisticated method looks over longer distances, i.e., uses a wider window, before and after the current point in an attempt to attain such a synoptic view. Here, a sequence of direction cosines is defined for each point, p_i , in the chain as:

$$\begin{aligned} \mathbf{u}_{il} &= (x_i - x_{i-l}, y_i - y_{i-l}), \\ \mathbf{v}_{il} &= (x_i - x_{i+l}, y_i - y_{i+l}), \\ c_{il} &= \frac{\mathbf{u}_{il} \cdot \mathbf{v}_{il}}{\|\mathbf{u}_{il}\| \|\mathbf{v}_{il}\|}. \end{aligned} \tag{5-10}$$

c_{il} is called the l -cosine at p_i . All the c_{il} are found for some limit, k , around p_i . This sequence is searched, starting with the widest window ($l = k$) and narrowing it, search for the first maximum of c_{il} . More precisely, an h is found such that:

$$c_{ik} < c_{i,k-1} < \cdots < c_{ih} \not\leq c_{i,h-1}. \tag{5-11}$$

We now define c_{i,h_i} to be c_{ih} and call it the cosine at p_i . c_{ih} measures the angularity at a point p_i using the widest possible window within the selected limit, m . A simple example is shown in Figure 5-6. A curvature maximum is said to exist at p_i if $c_{i,h_i} \geq c_{j,h_j}$ for all j such that $|i - j| \leq h_i/2$. This last test ensures that no greater c_{j,h_j} exists between p_{i-h_i} , p_i , and p_{i+h_i} , i.e., in the window for which the angularity at p_i was determined.

This is a relatively straightforward method for finding the significant points along a curve. It was developed for general purpose examination of line shape on the computer without any specific concern for its applicability to cartography. It works well in many situations; however, some polygonal shapes can cause problems

Figure 5-6. Determination of the cosine at p_i with $k = 3$. Notice that the angle from line 1 to line 1' is greater than the angle from line 2 to line 2', as is the angle from line 3 to line 3'. This implies that $c_{i3} < c_{i2} \not< c_{i1}$, which shows that $h = 2$ in this example.

for this technique. In addition, there is no theoretical basis for the choice of the parameter k . Other methods use more complex tests in an attempt to find the most important points of maximum deflection (Rosenfeld & Weszka, 1975; Davis, 1977). Different approaches based on similar concepts have also been developed (Rutkowski, 1981; Rosenberg, 1972).

Tolerance bands

This class of algorithms shows substantial promise for cartographic applications. Several of these techniques were designed specifically for cartographic applications. As with the angle detection techniques, an attempt is made to find the critical, or most important, points along a chain. A range or tolerance is selected, typically on the basis of the width of the pen used to draw the line or the size of a pixel on the graphic device being used. The procedures attempt to omit or ignore variations in the line that are less than the tolerance and to retain points where the deviation in the shape of the line becomes larger than the tolerance.

A number of these algorithms rely on the determination of the perpendicular distance from a point, p_i , in the chain to a line between two other points in the chain, p_j , and p_k , where $i \neq j, j \neq k$, and $i \neq k$ as shown in Figure 5-7a. This

Figure 5-7. (a) The perpendicular distance from a point to a line between two other points. (b) The vectors used in computing the perpendicular distance.

requires the definition of two vectors as follows:

$$\begin{aligned} \mathbf{u} &= (x_j - x_i, y_j - y_i), \\ \mathbf{v} &= (x_k - x_j, y_k - y_j). \end{aligned} \quad (5-12)$$

Thus, $\mathbf{u}$ is the vector from p_j to p_i , and $\mathbf{v}$ is the vector from p_j to p_k . The perpendicular distance can be seen from Figure 5-7b to be

$$d_{ijk}^{\perp} = \|\mathbf{u}\| \sin \theta. \quad (5-12)$$

From Equations A-3, A-9, and A-10, it follows that

$$\begin{aligned} d_{ijk}^{\perp} &= \frac{\|\mathbf{u} \times \mathbf{v}\|}{\|\mathbf{v}\|} \\ &= \frac{(x_i - x_j)(y_k - y_j) - (x_k - x_j)(y_i - y_j)}{\sqrt{(x_k - x_j)^2 + (y_k - y_j)^2}}. \end{aligned} \quad (5-13)$$

One of the earliest examples of this class of techniques was developed by Lang (1969). The procedure starts with $j = 1$, and k set to some pre-selected value, k_0 . p_1 is already saved as part of the simplified chain. The perpendicular distance from p_i to the line between p_1 and p_k , i.e., $d_{i1k}^{\perp}$, is computed for each $i = 2, \dots, k-1$, and compared against the selected tolerance, ϵ . If all of the distances are less than the tolerance, then p_k is retained as part of the simplified chain. Otherwise, k is assigned to the previous point, i.e., $k \leftarrow k-1$, and the $d_{i1k}^{\perp}$ is computed again with the new value for k . This continues until a point to be retained is found. If k is

Figure 5-8. The early steps of the Lang algorithm with $k_0 = 5$. In (a), the distances, $d_{i1k}^\perp$, from p_i to the line between p_1 and p_5 , are less than ϵ for $i = 2, 3, 4$, so p_5 is retained in the simplified chain. In (b), where a narrower tolerance is used, $d_{315}^\perp$ is greater than ϵ . Therefore, k is decreased from 5 to 4, and a new series of distance tests begins in (c).

decremented until it is set to 2, then p_2 will be saved because no intermediate points exist to cause the test to fail. Thus, a point to be saved will always be found. Once a point is retained, the algorithm restarts with the new point acting as p_1 in the description above. The procedure moves step by step along the chain until the last point is reached. This last point is always saved. This is illustrated in Figure 5-8.

A variation on this approach is more commonly used because it is usually faster than the method above (Douglas & Peucker, 1973). Here the point, j , is called the *anchor*, and the point, k , is referred to as the *floater*. The initial anchor point is saved in the simplified chain. The procedure starts with $j = 1$, and $k = n$, where n is the total number of points in the sequence. That is, the initial line is from the first point to the last point of the sequence to be simplified. The $d_{ij}^\perp$ are computed for all $i = j + 1, \dots, k - 1$. The largest $d_{ijk}^\perp$ is found. If the largest $d_{ijk}^\perp > \epsilon$, then the point, p_i , with the maximum $d_{ijk}^\perp$ becomes the floater, i.e., $k \leftarrow i$, and the process is repeated. As each floater is found, the index, k , is saved in a

list so that the positions of all the floaters can be recovered in the order opposite to that in which they were discovered. That is, the most recently generated floater is on the top of the list. Such a list is called a *stack* in computer terminology. If the largest $d_{ijk}^\perp < \epsilon$, then all the points between j and k are eliminated. The point p_k now becomes the anchor, i.e., $j \leftarrow k$, and the new anchor point is saved in the simplified chain. The floater is reset to index on top of the stack, i.e., the most recently generated floater. The perpendicular distance tests are performed again with the new anchor and floater. This process is repeated until all the points in the chain are within ϵ of the simplified chain. This technique is illustrated in Figure 5-9. The algorithm is applied to a closed curve by setting the first anchor to be p_1 , and first floater to be p_n , even though these are the same point. In this case, the initial tolerance test just searches for the point along the chain which is furthest from the beginning and end of the chain. After this, the technique proceeds as above.

A very similar approach was developed at about the same time by Ramer (1972). Related but distinct methods are also available (Opheim, 1982; Dettori & Falcidieno, 1982).

Point relaxation

Point relaxation methods simplify a line by allowing the points along the line to relax their positions a tolerance, d , in any direction. The simplest method maintains the position of the retained points (Williams, 1978). More sophisticated approaches will result in fewer lines (Williams, 1981). However, in the cartographic applications being considered here, retaining the locations of the simplified points is a desirable quality. As a result, the earlier algorithm will be discussed here.

Figure 5-9. The Douglas-Peucker algorithm. (a) In the first step the maximum perpendicular distance is for p_3 . This maximum is greater than ϵ . Consequently, it becomes the floater and its index, 3, is placed on the stack. (b) The distance computations are now performed for the original anchor and the new floater. In this case, no points are greater than ϵ from the test line, so the floater becomes the anchor and is marked as part of the simplified chain. It is also removed from the stack. This returns 7 to the top position. (c) The new anchor, 3, and the new floater, 7, define the test line. p_4 is found as a distance maximum greater than ϵ and will be retained. (d) The last test line has no points greater than ϵ from it, so the procedure is complete, with p_1, p_3, p_4 , and p_7 making up the simplified chain.

It is appropriate to express the cartographic line P in polar coordinates in this method, i.e.,

$$P = \{p_1 = (0, 0), p_i = (r_i, \theta_i) | i = 2, n\},$$

where

$$\begin{aligned} r_i &= \sqrt{(x_i - x_1)^2 + (y_i - y_1)^2}, \\ \theta_i &= \arctan\left(\frac{y_i - y_1}{x_i - x_1}\right), \end{aligned} \tag{5-13}$$

which follows from Equations A-1. The goal of this technique is to find the largest k for which a line from p_1 to p_k is within the tolerance, ϵ , of all the $p_i, i = 2, \dots, k$. The

Figure 5-10. All the lines from p_1 which pass within ϵ of p_i lie within the wedge whose boundaries are tangent to the circle of radius ϵ around p_i . These lines define a range of angles referred to as S_i in Equation 5-14.

set of all lines from p_1 which pass within the tolerance of p_i is, using the notation of Williams (1978):

$$S_i = \{\theta = \phi | \theta_i - \delta_i \leq \phi \leq \theta_i + \delta_i\}, \quad (5 - 14)$$

where $\delta_i = \arcsin\left(\frac{\epsilon}{r_i}\right); i \neq 1$. These relationships are illustrated in Figure 5-10.

The lines which pass within ϵ of the first i points are:

$$T_i = \bigcap_{j=2}^i S_j, \quad (5 - 15)$$

i.e., T_i is the intersection of the S_j for the points from p_2 to p_i . This is shown in Figure 5-11. For computing purposes, T_i can also be expressed as:

$$T_i = \{\theta = \phi | \theta_{min} \leq \phi \leq \theta_{max}\}, \quad (5 - 16)$$

where

$$\begin{aligned} \theta_{min} &= \max(\theta_j - \delta_j), \\ \theta_{max} &= \min(\theta_j + \delta_j); \quad 2 \leq j \leq i. \end{aligned} \quad (5 - 17)$$

The algorithm begins by assigned p_1 to the simplified line. Then the T_i are calculated until this set becomes empty, i.e., $\theta_i < \theta_{min}$ or $\theta_i > \theta_{max}$. When this happens,

Figure 5-11. The Williams technique is illustrated for the first three points of a chain. The two ranges of angles, S_2 and S_3 , overlap in a narrower range of angles, T_3 . As additional points are considered, the intersections, T_i , will tend to narrow until, for some point p_m , no line can lie within ϵ of all the points, $i = 1, \dots, m$. When this occurs, p_{m-1} is retained in the simplified chain.

p_{i-1} is included in the simplified line and the procedure is repeated with p_i acting as p_1 in the description above. Williams (1978) also discusses some optimizations of the formulae above. However, these appear to make some assumptions about the relationships between the p_i which may not be warranted for all cartographic lines. Since this algorithm will be applied to large numbers of such lines, only the basic equations, which are the most robust, have been implemented.

Other approaches to point relaxation exist. However, most are limited in application to closed curves such as lakes (Montanari, 1970; Oommen & Kashyap, 1983). They will not be considered here because of such limitations.

Fractal Techniques

Fractal geometry is a relatively recent advance in mathematics and graphics which grew, in part, out of consideration of certain aspects of cartographic lines. In standard geometry, points are considered to have a dimension of 0, lines have dimension 1, and areas have dimension 2. Thus, dimension is an integer quantity. In many geometric and graphics applications, this is entirely adequate. However, in other situations the concept of integer dimensionality can become blurred. Coastlines, as represented on a map, are one classic example.

Simple geometric shapes such as straight lines, polygons, and circles have unique and unambiguous lengths or perimeters. For a straight line, one measurement will suffice. Measurement of the length of a circle or some other simple curve requires a somewhat more complex approach. One procedure is to step along the length of the curve with a specific unit of distance and to count the steps taken. Multiplying the number of steps times the unit distance approximates the length of the total curve. This is illustrated in Figure 5-12a. If a large step size is chosen, then the approximation of the length will be relatively crude. However, an improved estimate will be found if a smaller unit of distance is selected as shown in Figure 5-12b. As the step size gets smaller, the length approximations will approach a limiting value. Such a well-defined limit is the true length of the line. A line with an unambiguous length has dimensionality of 1.

In the case of certain kinds of cartographic lines, such a well-defined length cannot be found. The measured length is inherently a function of the step size used to measure it, as well as the scale at which the line is represented. The length will always increase as the step size decreases and does not approach a limit over a very wide range in the step size. This effect was first noted by Richardson (1961) and is

Figure 5-12. Two approximations of the length of the curve using (a) coarse and (b) more refined step sizes.

illustrated in Figure 5-13. The measured length of the coastline will also increase as the scale of representation increases (Maling, 1968; Baugh, 1976). The length of a coast on a small-scale map will be less than the length of the same coast measured on a large-scale map. This reflects the underlying characteristics of the boundaries that occur in nature as well as cartographic lines used to represent these boundaries. This arises because, in the case of a complex coastline, there is no scale at which land-water boundary becomes simple and smooth. In a certain statistical sense, the shape of a coastline remains the same regardless of the scale at which it is examined over a substantial range of scale.

Lines which have an unambiguous length, when measured as above, clearly have a dimension of 1. However, the dimensionality of lines with an ambiguous length is less obvious. Because the length of a coastline, and its representation on a map, is ambiguous, it seems different from lines with dimension 1. However, it clearly does not cover an area in the normal sense, so it does not have dimension 2. The concept of *fractal dimension* allows for lines with non-integer dimensionality, i.e., the dimension of a line can be any value D , $1 \leq D < 2$. The fractal concept was developed to address complex lines such as coastlines (Mandelbrot, 1967; Goodchild, 1980). While the length of a coastline does not approach a limiting value as a

Log_{10} (Length of Side in Kilometres)

Figure 5-13. The length of various curves as a function of the step size used to measure them. With a simple curve, such as a circle, the measured length rapidly approaches a constant value which is the true length of the curve. In contrast, for complex coastlines such as that of Britain, the length continues to increase as the step size decreases. This increase in length is such that, when the relationship of step size to total length is plotted on a log-log graph, the resulting line is essentially straight. From Richardson (1961, p. 169).

function of step size, the measured line length often does exhibit regularity. As

when using step size ϵ , it follows that:

$$\begin{aligned}
 N\epsilon &= \epsilon^{1-D}, \\
 \ln N + \ln \epsilon &= (1 - D) \ln \epsilon, \\
 \frac{\ln N}{\ln \epsilon} &= -D, \\
 D &= \frac{\ln N}{\ln (1/\epsilon)}. \tag{5-19}
 \end{aligned}$$

Here D is the fractal dimension. For a simple line, $D = 1$, as would be expected. The land frontier between Spain and Portugal gives $D = 1.14$ and the very complex west coast of Britain gives $D = 1.25$. Thus, the perceived complexity of the cartographic line is correlated with the fractal dimension in a fairly direct way. A simplification algorithm based on the stepping approach discussed above has been proposed because it will maintain the fractal dimension of the line over a certain range of scale change (Muller, 1987b).

The goal of the procedure is to find points along the track of the line that are the step distance, ϵ , apart in straight line distance. This is analogous to walking a pair of dividers along the line (Shelberg et al. 1982). This requires the computation of various distances between points on both the original and the simplified chain.

These distances between points are computed in the usual way according to Equa-

Figure 5-14. The relationship between a point on the simplified chain and the points of the original chain.

Figure 5-14. If p'_s lies exactly on a point of the original line, that point is considered to be p_i , and p'_s is still considered to lie between p_i and p_{i+1} . An index, j , is set to $i + 1$, i.e., $j \leftarrow i + 1$, and the distance from p'_s to p_j is computed. If $d_{sj} < \epsilon$, then j is incremented, i.e., $j \leftarrow j + 1$, to allow the distance to the next point of the original chain to be tested. As j moves along the points of the original chain, one of two things will happen. Either j will move through all the points remaining in the original chain, in which case, the last point, p_n , will be added to the simplified chain, and the procedure will stop, or a j will be found where $d_{sj} \geq \epsilon$. When this occurs, several different possibilities must be handled.

The simplest case occurs when $d_{sj} = \epsilon$. When this happens, p_j is assigned to the simplified chain, $i \leftarrow j$, $j \leftarrow i + 1$, and $s \leftarrow s + 1$, i.e., s now points to the newly added point on the simplified chain. These assignments are illustrated in Figure 5-15. When these operations are complete, the procedure can now continue at the beginning of this description.

There are two situations that must be handled when $d_{sj} > \epsilon$. One possibility is that $j = i + 1$, i.e., that the entire step, ϵ , fits on the one line segment from p_i to p_{i+1} . Thus, the next point of the simplified chain can be found by moving a distance ϵ along the same line segment that p'_s already lies on. The following

Figure 5-15. The procedure when the next point lies exactly on a point from the original chain. (a) The indices when the matching point is found. (b) The indices after the new point has been added to the simplified chain.

Figure 5-16. The procedure when the next point lies on the current line segment. (a) The indices at the beginning of the operation. (b) The indices after the new point has been added to the simplified chain.

formulae perform this operation:

$$\begin{aligned} x'_s &= x_i + \frac{d_{is} + \epsilon}{d_{ij}}(x_j - x_i), \\ y'_s &= y_i + \frac{d_{is} + \epsilon}{d_{ij}}(y_j - y_i). \end{aligned} \quad (5-20)$$

This new point, p'_s is included in the simplified chain and $s \leftarrow s + 1$. Since the simplified point is still on the same line segment, the other indices, i and j , remain set to the same values. The basic procedure is ready to begin again. This operation is shown in Figure 5-16.

If $j > i + 1$, then a more complex procedure which relies on trigonometry is required. A typical situation is shown in Figure 15-17a. The line segments between

points used in these computations will be referred to using the same subscript notation as the distance variables already mentioned, i.e., the line segment from p'_j to p_j will be referred to as $l_{s,j}$. First, a triangle is formed from the points p'_j , p_j , and p_{j-1} . The angle of this triangle between $l_{s,j-1}$ and $l_{j-1,j}$ will be called α , as shown in Figure 5-17b. This angle can be computed from the law of cosines, because the lengths of the three sides of the triangle are known. Thus,

$$\alpha = \arccos \frac{d_{j-1,j}^2 + d_{s,j-1}^2 - d_{s,j}^2}{2d_{j-1,j}d_{s,j-1}}. \quad (5-21)$$

Now a new triangle is formed by shifting the endpoint of the side opposite α from (x_j, y_j) towards (x_{j-1}, y_{j-1}) until this side has length ϵ . This can be done because of the relation of lengths $d_{s,j-1}$ and $d_{s,j}$ to ϵ , i.e., $d_{s,j-1} < \epsilon < d_{s,j}$. This is illustrated in Figure 5-17c. This new side will intersect $l_{j-1,j}$ at some point $(x'_{s'}, y'_{s'})$. This new triangle also has α as one of its angles. The other two angles of the triangle will be called β , which is measured from $l_{s,s'}$ to $l'_{s',j-1}$, and γ , which measured from $l_{s,j-1}$ to $l_{s,s'}$. Noting that $d_{s,j-1} \sin \alpha = d_{s,j-1} \sin(\pi - \alpha) = l_{s,s'} \sin \beta$, and that $l_{s,s'} = \epsilon$, then β can be calculated by:

$$\beta = \arcsin \frac{d_{s,j-1} \sin \alpha}{\epsilon}. \quad (5-22)$$

Now γ is easily found from the fact that $\alpha + \beta + \gamma = 180$. Since all the angles and the lengths of two of the sides are known, the third side can be determined by:

$$d_{j-1,s'} = \frac{d_{s,j-1} \sin \gamma}{\sin \beta}. \quad (5-23)$$

$d_{j-1,s'}$ is the distance from p_{j-1} that must be traversed to reach $p'_{s'}$, and proceeding in a manner similar to Equations 5-20:

$$\begin{aligned} x'_{s'} &= x_{j-1} + \frac{d_{j-1,s'}}{d_{j-1,j}}(x_j - x_{j-1}), \\ y'_{s'} &= y_{j-1} + \frac{d_{j-1,s'}}{d_{j-1,j}}(y_j - y_{j-1}). \end{aligned} \quad (5-24)$$

Figure 5-17. (a) The geometry for $j > i + 1$. Note that because $d_{j,j} > \epsilon$ and $d_{j,j-1} < \epsilon$, the next step must lie on the line between p_{j-1} and p_j , i.e., on $l_{j-1,j}$. (b) The first triangle for which the lengths of all the sides are known, thus allowing the computation of α . (c) The second triangle used to compute the angles, β and γ . On the basis of these angles, $d_{j-1,s'}$ can be determined. When this is known, the position of the new point, p'_s , can be calculated. (d) The indices after the completion of this operation.

This procedure is completed by assigning $p'_{s'}$ to be p'_s and $i \leftarrow j - 1$. The index, j , is already set to the correct value. The positions of the various indices on completion of this operation are shown in Figure 5-17d. The procedure is now ready to begin again.

The procedure is initialized with $i \leftarrow 1$ and $j \leftarrow 2$ and the first p'_s is equal to p_1 . The step length, ϵ , can be chosen on the basis of a requirement for simplification. Thus, if n is the number of points, p_i , in the original line, and the cartographer wishes to express the line at a new scale with $n/10$ points, then the total length of the original digitized line, L , is computed and $\epsilon = \frac{L}{n/10} = \frac{10L}{n}$. The required step size is, therefore, 10 times the average distance between successive points in the original line.

Fractals have other roles to play in cartography. For example, they can be used to create a more realistic image of an area when relying on too little data (Dutton, 1981). By adding statistical variation to a line or surface that has appropriate characteristics, a more natural impression may be given than could be produced relying only on the measured data. While this is useful in certain applications such as flight simulation, it is not appropriate for many cartographic applications. More central to cartographic research, is the use of fractal dimension to study the structure of cartographic lines. Battenfield (1989) has proposed that the fractal dimension and the change in fractal dimension at specific scales be used as a means to classify such lines.

Algorithm summary

The algorithms described above will be used in the remainder of this study. They cover a wide range of the available techniques. The nine original classes of algorithms have been grouped into low, medium, and high potential algorithms by Zoraster et al. (1984). The techniques to be used here include one from the low potential group (point selection), two from the medium potential group (angle detection, low pass filtering), and two from the high potential group (tolerance bands, point relaxation). The fractal technique, which was proposed after the work by Zoraster et al. was written, has been included. Thus, six basic techniques for simplification will be used. For ease of reference, each algorithm will usually be referred to by a one word descriptor as shown in the following table:

Algorithm	Descriptor
point selection	skipping
low pass filtering	smoothing
fractal related	walking
point relaxation	relaxing
tolerance band	banding
angle detection	angling

Other proposals have been made for the classification of line simplification algorithms. McMaster (1987) has proposed that algorithms be classified on the basis of the size of the region of the chain around a given point which is considered in making the decision on whether to retain that point. The proposed categories include: (1) independent point routines, (2) local processing routines, (3) constrained extended local processing routines, (4) unconstrained extended local processing routines, and (5) global routines. The six techniques being examined fit into four of these five categories. The skipping technique examines no other points and is the classic example of a procedure in category 1. Category 2 techniques only examine the immediately adjacent points. No such techniques are used in this study. The smoothing and angling methods look at a window with a width that varies with the degree of simplification required. Thus, they examine more points than a local processing routine. They are constrained to a fixed number of points in the window. Consequently, they are in category 3. The walking technique is also a constrained extended processing routine because of the fixed step distance. Such a step will cover a variable number of points. However, the fixed distance limits the extent of the search. The relaxing routine is unconstrained. For example, a very long straight line represented by an arbitrarily large number of points in the original line will be simplified to two points by this technique. The range of search is unconstrained in that any number of points will be searched over any distance as long as they can be represented by one line which lies within the tolerance of all the points. Thus, this technique is in category 4. The banding algorithm is a global technique in that it begins by examining the complete chain and is an important example of a category 5 algorithm. Thus, most of the possible categories in this classification are represented.

Of the six algorithms used in this study, two important groups can be dis-

use the structure of the original chain to derive the simplified chain. In contrast, the other three techniques do make such an attempt. The angling and banding procedures look specifically for critical points in the original chain. The relaxing technique searches for sequences of points which lie close to a straight line. The first group of techniques will be referred to as form-independent procedures in this study, while the second group will be called form-sensitive procedures.

While these techniques are highly varied, they have some common traits. All the algorithms maintain the positions of the first and last points of the cartographic lines that are simplified. In addition, all have a relatively straightforward geometrical interpretation and, consequently, the mathematics associated with each approach is manageable. This relative simplicity also implies that the programming associated with each has minimal difficulties.

Chapter VI

Methodology

This work took place in several stages. First, two publicly available cartographic databases were placed in a common format to allow comparable processing and display to take place. Both of these databases are in a "spaghetti" format; therefore, the data was enhanced with the addition of topological coding which allowed more effective application of the simplification algorithms. Generalized databases were then produced using various combinations of scales and simplification techniques. Maps, which were created from these derived databases, were examined visually. This allowed subjective judgments to be made regarding the efficacy of the individual techniques as well as the overall approach to handling scale change. A mathematical evaluation of the derived databases was also performed to provide a more objective basis for comparison. Similar studies limited to small sets of cartographic lines have been done by White (1985), Muller (1987b), and McMaster (1987b). A detailed description of the databases and their preparation, along with the various steps in evaluating the line simplification algorithms is provided below.

Cartographic databases utilized

This work relied on two pre-existing databases. One database was produced by the National Center for Atmospheric Research (NCAR) and is available as part of their graphics package, NCARLIB. It will be referred to as the NCAR cartographic database (NCARCDB). The other database, World Data Bank II (WDB II), was produced by the Central Intelligence Agency and is publicly available from the

National Technical Information Service. These databases have very simple structures consisting of geometry and some attribute data that allow the differentiation between major classes of features.

NCARCDB

The NCARCDB contains three distinguishable classes of data: (1) continental coastlines and major lakes, (2) international political boundaries, and (3) United States coastlines, state boundaries and major lakes. Within these three classifications, no differentiation is possible. The data is made up of a series of string records. Each string record consists of a string header followed by list of up to 200 coordinates (100 coordinate pairs). The coordinate pairs are latitude, longitude values. The string header is composed of fields for the number of coordinates, the feature classification (this field specifies one of the three feature groupings mentioned above), the minimum and maximum values for the latitudes and longitudes in the list of coordinates, and a string id number. International coastlines are represented by 16838 coordinate pairs, international political boundaries by 11304 coordinate pairs, and the U.S. data includes 21744 coordinate pairs. Documentation in the graphics package claims that international coastlines were recorded at about one degree intervals at 0.01 degree resolution, although the data points appear to occur somewhat more frequently than this reported resolution. None of the higher resolution data for the United States was used in this study. The original maps used as a source for digitizing are not specified in the documentation.

WDB II

WDB II is a much larger database than NCARCDB (Gorny, 1977). Because of its size, it is divided into separate files covering continental areas as shown in Figure 6-1. In addition, each continental data set is divided into overlay files

WDB II Area Files

Figure 6-1. The division of WDB II into files covering continental areas. From Gorny (1977, p. 1)

for coastlines and lakes, major rivers, and international political boundaries. The North American area also includes an overlay containing U.S. states and Canadian provinces. The data is made up of a series of string records. Each string record consists of a string header followed by list of coordinate pairs. The coordinate pairs are latitude, longitude values. The string header is composed of fields for the number of coordinates, the feature classification, and a string id number. The feature classifications are shown in Figure 6-2. This database was digitized from various map sources which range in scale from 1:1,000,000 to 1:3,000,000. The complete database contains over six million locations.

Common data format

To allow for comparison between the two databases, the more detailed attribute information in WDB II was first used to select features appropriate for

Features and Subordinate Classifications/Ranks in WDBII

- I. International boundaries or limits of sovereignty
 - 01 Demarcated or delimited
 - 02 Indefinite or in dispute
 - 03 Other line of separation or sovereignty on land
- II. Coast, islands, and lakes
 - 01 Coast, islands, and lakes that appear on all maps
 - 02 Additional major islands and lakes
 - 03 Intermediate islands and lakes
 - 04 Minor islands and lakes
 - 06 Intermittent major lakes
 - 07 Intermittent minor lakes
 - 08 Reefs
 - 09 Salt pans—major
 - 10 Salt pans—minor
 - 13 Ice shelves—major
 - 14 Ice shelves—minor
 - 15 Glaciers
- III. Rivers
 - 01 Permanent major rivers
 - 02 Additional major rivers
 - 03 Additional rivers
 - 04 Minor rivers
 - 05 Double-lined rivers
 - 06 Intermittent rivers—major
 - 07 Intermittent rivers—additional
 - 08 Intermittent rivers—minor
 - 10 Major canals
 - 11 Canals of lesser importance
 - 12 Canals—irrigation type
- IV. Internal boundaries
 - 01 First order admin

Figure 6-2. The feature classifications used in WDB II. From Gorny (1977, p. 2)

small-scale maps. Then the remaining categories were collapsed to the same three categories available in the NCARCDB. That is, only coastlines, major lakes, and political boundaries were used. The shores of major lakes, such as the Caspian Sea and the Great Lakes, were considered to be coastlines for this work. Thus, only two overlays, coastlines and political boundaries, have been utilized. The elimination of the numerous smaller features in WDB II was important due to the quantity of data in that database. If all the data were plotted at small-scale, the result would

be incomprehensible in many cases. This also simplified the creation of maps with similar content from the two databases. The NCARCDB was processed in its entirety; however, due to the size of WDB II, this project limited its scope to North and South America along with Europe (excluding the Soviet Union).

While the databases contain large amounts of geometric information, they include only a modest amount of attribute data and no topological information. The limited attribute information permits some control over the symbolization of the features. The lack of topological structure means that the application of the simplification algorithms could cause problems. The coordinate strings begin and end at arbitrary locations that are unconnected to anything else in the map structure. Important points, such as intersections between political boundaries and coastlines are not identified in the location strings. The application of a simplification routine, without consideration of the other features, could result in the omission or movement of significant points causing gaps and overshoots to appear in the maps produced from the simplified data. Therefore, it was essential to process these databases into a common format that includes an adequate topological structure. A common format for the geometric, attribute, and topological data also permits the application of software developed for one database to the other database with little, if any, alteration.

A procedure for developing topological information from unstructured data can be complex. Since such a procedure was not the main focus of this study, the process was simplified by the decision to ignore the rivers overlay in WDB II. Because rivers often define political boundaries for portions of their lengths, complicated geometries can result which are difficult to recognize and process automatically. The line intersections that exist between political boundaries and coastlines

normally only occur at the endpoint of at least one segment, thus limiting the search required. Exceptions were handled via interactive graphics software (see Appendix C).

The topological information includes nodes that are defined at intersection points in the databases. The intersection points considered encompass those between political boundaries and coastlines, as well as the intersections between features within the same overlay. All coordinate strings extend from one node to another node, i.e., they are chains. Pointers connect each chain to its bounding nodes. The topological structure is redundant in that nodes will have pointers which connect them to the chains that begin and end at them. In addition, each chain has space for pointers to the areas on its left and right. The recognition and processing of area features is beyond the scope of this project. These spaces are included to allow extension of this basic structure to more complex applications.

Three types of files are maintained as part of this effort: (1) chain files, (2) point files, and (3) node files. One separate chain file exists for coastlines and major lakes, and another for political boundaries. A chain file contains one record for each chain in the overlay. The structure of this record is:

chain number	attribute code	latitude minimum	latitude maximum	longitude minimum	longitude maximum	point location	point number	start node	end node	left area	right area
-----------------	-------------------	---------------------	---------------------	----------------------	----------------------	-------------------	-----------------	---------------	-------------	--------------	---------------

The chain number is used to identify the chain in computer processing, and is the key, i.e., the field most commonly searched for, in the record. The attribute code classifies lines within an overlay. The next four fields define the bounding box of the chain. This is a very simple mechanism to optimize processing the chain file. For example, if a map is to be plotted for a coordinate window which is smaller than the area covered by the map data, all chains that have a bounding box outside

the window can be skipped without any processing of the coordinates that define the chain. The next field is a pointer into the point file associated with the overlay, this is followed by the number of points that are used to define the chain. The last fields that are used are the node numbers for the beginning and ending nodes which bound the chain, i.e., they are pointers into the node file.

A point file exists for each overlay. It is simply a list of coordinate records. Each coordinate record contains a latitude and a longitude. For general mapping applications where many different projections are used, geographic coordinates provide the greatest flexibility. Transformations to various map projections are executed by the software that displays the information. The point file is essentially unformatted. Its utility is based entirely on the pointers and lengths that exist in

Evaluation of simplification algorithms

The existence of the structured cartographic files permits experimentation with the various simplification algorithms. The format of the databases allowed simplification techniques to be executed on individual chains so their behavior could be examined in detail. In addition, it was possible to produce complete maps at a variety of scales, on various projections, and with differing degrees of simplifications, thereby acquiring a more synoptic view of the performance of the algorithms. Two methods of evaluation were performed. The various maps produced were examined subjectively to determine the relative success or failure of the various techniques and to detect general characteristics in performance of each method. A more quantitative approach was also carried out, where a number of mathematical measures of performance were calculated.

Subjective evaluation

The judgement of the cartographic adequacy of the simplified databases was made largely on subjective criteria. Various series of maps were generated that both highlight the differences between the various techniques at given level of simplification and track the effectiveness of the techniques as scale decreases. Some of the series are focused on determining the relative merits of the algorithms. These are compared with the results of the mathematical evaluation. Other series are used to define the extremes at which simplification techniques are effective at producing smaller-scale maps from a cartographic database.

As an introduction to the behavior of the techniques, databases were created using the NCARCDB with a modest degree of simplification. This provided some experience with the algorithms using a small, manageable data set. Databases were also created from the WDB II data at a range of scales. These databases

represented much greater changes of scale and level of simplification. As a result, they provide more extreme tests of the algorithms. Maps created from the simplified databases were examined both at relatively large scale, to show the details of the algorithms' behavior, and at the scale appropriate to the degree of simplification. These maps provide a qualitative assessment of the advantages and disadvantages of the various techniques. The level of simplification for these maps was based on a common tolerance parameter which was used for each technique (see Chapter VII for a discussion of how tolerances are defined and used in this study). A given tolerance gives only approximately the same degree of simplification with the various algorithms. Consequently, a more direct comparison of the techniques is provided by comparing simplifications of an individual chain based on the different algorithms, subject to the constraint that the number of points retained must be equal or nearly so. The requirement for representation with a specific numbers of points isolates the performance of the various techniques from vagaries associated with the choice of tolerance.

Common map projections were used to plot the data. The normal aspect Mercator projection was used for some relatively large-scale maps of specific details because the simplification computations were performed on this projection (see Chapter VII). The Miller cylindrical projection was used for some maps of the NCARCDB because it provides a compromise between the extreme distortions of the Mercator and the cylindrical equal area projections, while maintaining a rectangular view of the whole world which is convenient for some computer graphics operations. For small-scales views of large portions of the world, the orthographic projection was utilized as it provides a good approximation of the true shape of features as seen on a globe. Because an orthographic map tends to be seen as a view of a globe, the

distortions are understood and compensated for easily by most map readers (Dent, 1987).

Mathematical evaluation

To quantify the comparison of the simplification procedures, mathematical measures of the success of the techniques in representing cartographic lines with fewer points were computed and summarized for each simplified database. To allow easy comparison, the measures for a given technique were plotted as a function of the number of points used to represent the original database. A wide variety of mathematical approaches to the evaluation of line simplification algorithms were carefully examined by McMaster (1986). From a total of thirty techniques, six were identified on the basis of statistical studies as being the most independent and informative. These six measures were:

- (1) Percent change in the number of coordinates.
- (2) Percent change in the standard deviation of the number of coordinates per inch.
- (3) Percent change in angularity.
- (4) Total vector displacement per inch.
- (5) Total areal displacement per inch.
- (6) Percent change in the number of curvilinear segments (McMaster, 1986, p. 115).

In this study, the last four measures are utilized as measures for algorithm comparison. The first measure is used here to provide the basis for comparison between algorithms, i.e., the four measures are plotted as a function of the number of coordinates used to represent the database. Thus, the change in the number of coordinates is used as the independent variable rather than as a measure. The standard deviation of the point density was omitted as a measure. This measure is based the assumption that a large standard deviation is an indicator of good

performance because critical points are unlikely to occur at a constant frequency along a chain. While this feature of critical points is usually true, it seems that this measure relies on the assumption that good critical point identification goal is equivalent to good simplification. This may not be true in all cases. In addition, natural shapes sometimes do have a somewhat regular structure which could produce critical points at relatively constant intervals. The choice of the number of coordinates as the independent variable and the omission of the standard deviation of the coordinate density is consistent with more recent work by McMaster (1987a, 1987b, 1989). The four measures mentioned have been adapted to examine the large number of lines in a complete database. The first two of these measures are based on quantities intrinsic to a digital line. Thus, they can be applied to the original chains as well as to the simplified chains. The last two measures quantify the distance or space between two chains, the original chain and the simplified chain. The details of these measures are as follows:

1. *Relative change in angularity.* This measure attempts to quantify the “wiggleness” of the chain. It is based on the assumption that this angularity partially characterizes the chain and that retaining this characteristic on simplification is desirable. The angularity is computed from the angle, β_i , which is the supplement of the the vertex angle, α_i , discussed in Chapter V. This is shown in Figure 6-3. The direction cosine associated with α_i and defined in Equation 5-9 can be used to determine β_i . Thus, if c_i is the direction cosine for point p_i , then the angularity $\beta_i = \arccos(c_i) + \pi$. Thus, β_i has been adjusted so that it is zero for two adjacent segments oriented in the same direction and π (or $-\pi$) for a segment that doubles back on

Figure 6-3. The angle, β_i , used to compute the angularity for a chain, and the angle, α_i , associated with the direction cosines discussed in Chapter V.

the previous segment. The directions, β_i , are summed for the entire chain.

That is,

$$B = \sum_{i=1}^n |\beta_i|, \quad (6-1)$$

where B is the angularity of the chain and $|\beta_i|$ is the absolute value of the angle β_i . If the angularity for the original chain is B and the angularity for the simplified line is B' , then the first measure is $m_1 = B/B'$, with the relative change being used for comparison. Notice that the angularity of the original chain is divided by the angularity of the simplified chain. With large

degrees of simplification, it seems clearer to say that the angularity of the chain was reduced by a factor of 25 on simplification rather than saying that the simplified chain had a proportion 0.04 of the angularity of the original chain. To provide a summary measure for the complete database, the m_1 for all the chains were averaged. This average value was plotted as a function of the number of points used to represent the complete database.

2. *Relative change in the number of curvilinear segments.* The individual angularity values, β_i , defined above may be positive or negative. A sequence of β_i with the same sign means that the chain is curving consistently to the left or consistently to the right. A portion of the chain with consistent

Figure 6-4. The curvilinear segments for a short chain.

curvature is called a curvilinear segment. The curvilinear segments for a short chain are shown in Figure 6-4. The number of curvilinear segments is counted for both the original chain, giving A , and for the simplified chain, giving A' . The second measure $m_2 = A/A'$, using the relative change. As for the first measure, the m_2 for all the chains were averaged. This average value was plotted as a function of the number of points used to represent the complete database.

3. *Average vector displacement.* This is a measure of the degree that the simplified chain differs from the original line. It measures the perpendicular distance from points on the original chain to line segments on the simplified chain. It is most obviously applied to simplified chains where the points, p'_j , of the simplified chain are all points in the original chain, p_i , as shown in Figure 6-5a. For each point, p_i , on the original chain which lies between p'_j and p'_{j+1} on the simplified chain, $d_{i,j,j+1}^\perp$ is computed as in Equation 5-13 and as shown in Figure 6-5b. These $d_{i,j,j+1}^\perp$ are computed over the length of chains and summed. The result is divided by the total number of points in the simplified chain, n' , to give the standardized value used for m_3 . McMaster uses the total vector displacement per unit length; however, this makes comparison of different levels of simplification for a given chain difficult because the number of points per unit length may vary widely. When this

Figure 6-5. (a) The perpendicular distances between points on the original chain and the corresponding line segment of the simplified chain. (b) The vectors on which the computation of $d_{i,j,j+1}^+$ is based.

measure is applied to a simplified chain where the points are not necessarily part of the original line, some additional bookkeeping is required so that the distance from a point on the original chain is computed for the correct line segment of the simplified chain. A summary measure was again computed by averaging the values for m_3 for all chains and graphing the results as for the first two measures.

4. *Total areal displacement per unit length.* This is also a measure of the degree that the simplified chain differs from the original line. The p_i which lie between p'_j and p'_{j+1} , along with the line segment from p'_j to p'_{j+1} , form a polygon as shown in Figure 6-6a. The area of the trapezoid defined by each pair of adjacent points in the polygon, p_k and p_{k+1} , and the x -axis is computed according to

$$a_k = \frac{1}{2}(y_k + y_{k+1})(x_k - x_{k+1}). \quad (6-2)$$

The first and last points are considered adjacent as part of this process. When these areas are summed over the entire polygon, the area of the polygon is computed as shown in Figure 6-6b. The value may be positive or negative; however, since area is always positive, taking the absolute value

Figure 6-6. (a) A polygon made up of points from the original and simplified chains. (b) The area of the polygon equals the area of the dotted polygon minus the area of the dotted and hatched polygon. This can be computed by finding the areas of trapezoids formed by adjacent points of the polygon and the x -axis. The areas of the trapezoids between j and $i - 1$, $i - 1$ and i , plus i and $j + 1$ are added to give the dotted area. The trapezoid from $j + 1$ to j gives the dotted and hatched area. See Monmonier (1982, p. 114-118) for more details.

gives the correct result. When the line segment from the simplified chain crosses the original chain, the polygon is divided into two polygons at the intersection point and the areas of these polygons are computed in the same way. This subdivision of polygons is also important for handling the smoothing technique where the simplified points do not lie on the original line. In the same situation, bookkeeping is required to match points on the original chain with the correct line segments of the simplified chain. An actual example of the polygons used in the area determination for the smoothing algorithm is shown in Figure 6-7. The result for the complete chain is divided by the chain length, in kilometers, to get the area per unit distance used as m_4 . The same averaging was performed as with the other measures to get a summary value for display purposes.

Figure 6-7. The area computation used for the smoothing algorithm where the points of the simplified chain do not lie on the original chain. The vertices of the dotted line are at the points of the simplified chain. The different hatch styles indicate where the points of the original chain correspond to the segments of the simplified chain.

The summary measures have been plotted to identify important trends and patterns. To provide a common reference, the measures are plotted as a function of the number of points used to represent the test dataset. This provides a means of observing not only the relative performance of each algorithm, but also allows the behavior of an individual algorithm to be studied as the number of allowed points decreases, which is related to decrease in applicable scale.

The combination of subjective and quantitative evaluation is, perhaps, appropriate as it reflects the combination of art and science which plays such an important

role in cartography.

Chapter VII

Results

A description of the results covering each aspect of the methodology is presented in the following sections.

Database Formats

significant difficulty. The original database has the desirable characteristic that, with very few exceptions, each endpoint of a line is exactly co-located with a point in any line to which it is connected. Thus, although no topology is provided, the data are geometrically integrated. This greatly simplifies the extraction of the

One characteristic of the database, which should be mentioned, is the existence of some relatively small, narrow islands which are symbolized as lines rather than area features. These lines were left as isolated lines bounded by two nodes without any other connections.

Structuring the WDB II data was more difficult for two reasons. First, and perhaps most obvious, is the larger size of the database. This caused some implementation difficulties and prompted some additional thinning of the data over what could be accomplished with the attribute coding alone. Many relatively small islands are coded as major coastlines in the original database. Interactive graphics software was used to flag any islands which did not have corresponding representations in the NCARCDB. In spite of the limitation of the study to the Americas and non-Soviet Europe and the elimination of all lines except major coastlines and international boundaries, the dataset used in the study still contained about 1,200,000 coordinate pairs. This made some aspects of the processing somewhat time-consuming, particularly the extraction of topological structure.

The structuring was also complicated by the fact that, unlike the NCARCDB, the endpoints of lines did not, in general, match points in the lines to which they were connected. A partial solution to this problems is to perform a "fuzzy" test for proximity of points within a specified tolerance rather than requiring the exact matching of coordinates. The difficulty in this solution is that a tolerance which is large enough to find all the correct matches is large enough to include many nearby points which should not be topologically connected. This problem was particularly bad where long, relatively straight lines were split in the original data. The beginning of one line often overlaps the end of the adjacent line by a substantial distance that is much greater than the resolution of the database. This is illustrated

Figure 7-1. Two lines with overlapping ends. This type of error occurs many times in WDB II.

in Figure 7-1. This may reflect confusion on the part of the digitizer as to precisely where to restart a line or may, in some cases, be an artifact of inaccurate edge-matching between adjacent map sheets. Such a situation is also complicated by the fact that the closest point between the two lines does not, in general, identify the points which should be matched on either line. Software was written which resolved most, but not all, of these difficulties. In situations which required the movement of points, the necessary correction was made to the appropriate point to force the match and then the shift was linearly feathered over the line to the next node. The feathering avoids unrealistic jumps in the line at the match point by spreading the change out over the nearby portions of the line. The linear feathering concept has been used to edge-match adjacent digital map sheets (Walker, 1989) and to shift

Figure 7-2. The technique for matching a linear features in WDB II is demonstrated in (A) and (B). The two boundary nodes in (A) mark the points which should be matched. The internal nodes, which represent other match points in the database, are to be left unchanged by the process. $len1$ and $len2$ are the lengths along the lines from the internal nodes to the corresponding boundary nodes. The matched lines will meet at the connection point shown in the inset. The connection point is chosen so that $s2/s1 = len2/len1$, where $s1$ is the distance from the end of line 1 to the connection point and $s2$ is the distance from the end of line 2 to the connection point. The shift required to match the boundary points decreases linearly to zero for the other points along the line as the internal node is approached. The results of this process are shown in (B).

groups have also identified the importance of solving this problem. The CIA has announced that a more structured version of this data will be available early in 1990. Improvements in database structure and in the geometric consistency between different features will greatly enhance the ease with which users can incorporate such data into existing research programs and application systems.

Normal and Tangential Tolerances

In order to provide roughly comparable levels of simplification, a parameter was required which would control each of the techniques. Two of the techniques, the banding and the relaxing algorithms, are defined in terms of such a tolerance.

Figure 7-3. Tolerances are used to control the line simplification algorithms. The normal tolerance, ϵ_N , bounds the side-to-side variation of the chain, while the tangential tolerance, ϵ_T , controls variation along the length of the chain. Experimentation has shown that ϵ_T must be about four times ϵ_N to produce comparable levels of simplification with the various techniques.

This tolerance is used to bound the variation in the shape of a chain within a distance measured perpendicular to the general direction of the chain; thus, such a measure could be considered a *normal* tolerance, ϵ_N . The walking algorithm uses a distance measure which is generally along the length of the chain; consequently, this measure could be considered a *tangential* tolerance, ϵ_T . The distinction between the two tolerances is illustrated in Figure 7-3. The remaining procedures (skipping, smoothing, and angling) process a chain in groups of k points. A value for k was derived from a specified tangential tolerance by dividing this tolerance by the length of the chain and multiplying the result by the number of points in the chain, i.e.,

$$k = \frac{n\epsilon_T}{L}, \quad (7-1)$$

where n is the total number of points in the chain and L is the total length of the chain. This gives the average number of points along the line per tolerance length. Some experimentation produced a very rough relationship between the two tolerances. It was found that a normal tolerance of one-quarter the tangential tolerance, i.e., $\epsilon_T \simeq 4\epsilon_N$, would produce roughly comparable levels of compression in the number of points used to represent a chain. Using a single parameter, ϵ_T , comparable levels of simplification could be selected for each technique.

Qualitative Characteristics with Limited Scale Change

Individual Chain Evaluation

As a first test, individual chains from the NCARCDB were simplified with each of the techniques using a tangential tolerance of 200 km to provide comparable levels of simplification. These lines were plotted at large enough scales to allow specific characteristics to be highlighted. All the figures in this subsection (Figures 7-4 to 7-13) are drawn using a heavy solid line for the simplified chain and a thinner dashed line for the original chain. The border between Angola and Zaire is used to demonstrate the performance of each of the algorithms on a chain with a few obvious significant points. The skipping, smoothing, and walking techniques retain none of the significant points (see Figures 7-4a to 7-4c). This is expected since none of these procedures make any attempt to discover important points along the chain, i.e., they are form-independent. The smoothing algorithm is also characterized by incorporating points which are a substantial distance from the original chain. The other methods, i.e., the form-sensitive techniques, do much better at retaining the important points. The banding and angling methods differing slightly on the precise position of one critical point and on the respective omission and retention of an additional point of minor significance (see Figures 7-4e and 7-4f). The relaxing technique does find two of the critical points precisely; however, it also shows a problem with the other critical points by moving past the critical point before the angular tolerance used in the technique is exceeded (see Figure 7-4d).

The east coast of the United States illustrates two difficulties. First, the shape of Florida varies greatly between the various techniques. With the form-independent techniques (see Figures 7-5a to 7-5c), the problem is exacerbated by the fact that the shape is a function of the particular starting point of the chain. If a different starting

Figure 7.4 The Angola-Zaire border simplified using $\epsilon_s = 200$ km with the (a) skinning,

(b) smoothing, (c) walking, (d) relaxing, (e) banding, and (f) angling algorithms.

point is used, e.g., the mouth of the Mississippi River rather than the United States–Mexico border, then the shape of Florida is substantially different. This is less of a problem with the form-sensitive methods. The second difficulty is illustrated in the banding method (see Figure 7-5e) where a spike has degenerated further to become a loop south of Chesapeake Bay. The Chilean coastline provides another example. The form-independent techniques (see Figures 7-6a to 7-6c) capture the general shape of the chain but tend to eliminate the features that suggest the submerged nature of the southern coastline. The latter is especially true of the walking technique where the details of the coastline are, in general, smaller than the walking step and are removed to produce a bland coast with little distinctive variation. The form-sensitive methods (see Figures 7-6d to 7-6f) do retain some of the complexity of the submerged coast; however, in the case of the banding and angling techniques the retained points produce spikes that appear somewhat unnatural.

Additional experiments were run on individual chains from the NCAR CDB where the number of points used to represent the simplified chain were decreased over a substantial range. The east coast of the United States is used again to show a weakness of the form-independent techniques. With each of these techniques, the shape of Florida changes drastically as the number of points is decreased (see Figures 7-7 to 7-9). The form-sensitive techniques show a more gradual and consistent change from a very accurate representation to a caricature (see Figures 7-10 to 7-12). The coastline of Alaska highlights a difficulty with the smoothing algorithm (see Figure 7-13). As the number of points retained decreases, the simplified chain moves inward, away from the concave coastline. Movement away from a concave shape is a characteristic of such a low-pass filter. With a closed curve, a loss of

area will tend to occur with this technique and other similar filters as simplification becomes extreme.

(a)

(d)

Figure 7-6. The Chilean coastline simplified using $\epsilon_T = 200$ km with the (a) skipping, (b) smoothing, (c) walking, (d) relaxing, (e) banding, and (f) angling algorithms.

Figure 7-7. The eastern and Gulf coasts of the United States represented with a variable number of points, L , chosen by the skipping algorithm.

Figure 7-8. The eastern and Gulf coasts of the United States represented with a variable number of points, L , chosen by the smoothing algorithm.

Figure 7-9. The eastern and Gulf coasts of the United States represented with a variable number of points, L , chosen by the walking algorithm.

Figure 7-10. The eastern and Gulf coasts of the United States represented with a variable number of points, L , chosen by the relaxing algorithm.

Figure 7-11. The eastern and Gulf coasts of the United States represented with a variable number of points, L , chosen by the banding algorithm.

Figure 7-12. The eastern and Gulf coasts of the United States represented with a variable number of points, L , chosen by the angling algorithm.

Global Evaluation

Each of the six main techniques was examined over relatively modest changes in scale using the entire NCARCDB. The purpose of this was to acquire a general overview of the behavior of the algorithms when applied to a large set of chains with varying shapes before applying more quantitative comparisons and also before pressing the algorithms to extremes in scale change and compression. This preliminary examination was performed on the NCARCDB in order to deal with a small data set. The NCARCDB was plotted on the Miller cylindrical projection, which provides a rectangular view of the world with somewhat less areal distortion than the Mercator projection (see Figure 7-14). The original database included 14277 points. The NCARCDB is intended for use at very small scales and the presentation of the data at this scale ($\sim 1:175,000,000$) reflects an appropriate level of generalization although along the submerged coast of Chile and many of the Melanesian islands the lines coalesce to produce a somewhat smudged appearance. Using a tangential tolerance of 200 km (equivalent to a normal tolerance of 50 km) the data was simplified with each of the basic techniques. The simplified data sets each had about 4000 points (ranging from 3863 to 4830 points). When the simplified data sets were plotted in the same format, the differences are quite subtle (see Figures 7-15 to 7-20). All these maps are somewhat oversimplified for the scale. The smoothing and walking techniques seem to soften the shapes in most places. This is to be expected for the smoothing technique. The relaxing, banding, and angling techniques tend to sharpen most corners and highlight the important angular changes as they are designed to do. The skipping technique shows a mixture of both smoothing and sharpening.

Figure 7.14 The original NCAR CDR on a Miller cylindrical projection

Figure 7-15. The NCARCDB simplified with the skipping algorithm.

Figure 7-16. The NCARCDB simplified with the smoothing algorithm.

Figure 7-18. The NCAR CDB simplified with the relaxing algorithm.

Figure 7-19. The NCAR CDB simplified with the banding algorithm.

A second set of very small-scale maps was plotted using the orthographic projection to give a different and, in some ways, a more correct view of the chains as they are simplified. Once this was done, two difficulties were highlighted which were not apparent when the simplified data was plotted on the Mercator or Miller cylindrical projections. First, the border between the United States and Canada from the Pacific Ocean to Lake of the Woods was drawn as a straight line with the banding algorithm. Secondly, a dramatic difference exists between the density of retained points near the equator and those near the poles (see Figure 7-21). Both of these problems are related to the use of map projections in the simplification process. Discussion of these problems and other considerations associated with the use of map projections in the simplification of cartographic databases will be deferred until Chapter VIII.

Figure 7-21. An orthographic view of NCAR CDB after simplification with the banding algorithm. Notice the straight line representation of the United States-Canadian border and the variation in the density of points between the equator and the poles.

Qualitative Characteristics with Extreme Scale Change

Individual Chain Evaluation

Specific chains from the WDB II data were selected to examine, in detail, the effects of extreme compression of the number of points used to represent the chain. For the individual chain examples discussed previously, 30–40% of the points from the original NCARCDB chain were retained in the simplified representation. In the examples that follow, only 0.5–5% of the original points are retained. Such extreme compression reflects the degree of simplification necessary for large changes in scale. To provide a fair comparison between the various techniques, the selected chains were simplified subject to the constraint that the number of retained points must be approximately constant. Thus, the tolerance used for each technique was varied until the number of retained points was relatively close to a specified value. Each chain is plotted on the Mercator projection at a scale where the entire chain can be plotted. This allows the differences between the simplified chain and the original data to be seen clearly. For these maps (Figures 7-22 to 7-24), the simplified chain is drawn with a heavy line and the original chain is drawn with a thin line (see Appendix C).

The first example, the border between Paraguay and Brazil, is a relatively straightforward cartographic line with no problematic features (see Figures 7-22a to 7-22f). As might be expected, none of the techniques have major difficulties and they all capture, to some degree, the basic shape of the line. The banding algorithm produces a clearly superior result in this case. The west coast of the United States is examined next. Here, all the form-independent techniques give a poor representation of the Olympic peninsula, while all of the critical point techniques successfully

find the northwest corner of the peninsula (see Figures 7-23a to 7-23f). The banding algorithm seems to do better than the others in representing San Francisco Bay; however, this technique produces a spike at the mouth of the Columbia River. The third example, the border between Guatemala and Mexico, highlights the relative weakness of the form-independent techniques (see Figures 7-24a to 7-24f). Here, the sharp corners reflecting arbitrary political agreements between countries, which ignore any natural features, are a crucial component of the shape of the chain. The form-independent techniques do not capture any of the important corners and, as a result, miss much of the character of the chain. The form-sensitive methods do much better, either finding the corners exactly or coming very close. The attempt to find the most basic structural elements of the chain pays dividends in this case. The angling methods does the best at finding the corners with the banding slightly missing some corners and the relaxing technique missing all the corners by a somewhat larger amount.

Figure 7-22. The Paraguay-Brazil border represented using 15 points selected by the (a) skipping, (b) smoothing, (c) walking, (d) relaxing, (e) banding, and (f) angling algorithms.

Figure 7-23. The west coast of the United States represented using 49–52 points selected by the (a) skipping, (b) smoothing, (c) walking, (d) relaxing, (e) banding, and (f) angling algorithms.

Figure 7-24. The Guatemala-Mexico border represented using 19-21 points selected by the (a) skipping, (b) smoothing, (c) walking, (d) relaxing, (e) banding, and (f) angling algorithms.

Global Evaluation

Maps of North America and northern Europe were plotted on the orthographic projection at scales of 1:45,000,000 and 1:12,000,000, respectively, using data simplified by each algorithm. The tolerance was chosen to be approximately correct for the scale of the map to give the overall impression of the performance of the techniques. The general sense for all the methods is one of modest success. The

broad shapes of the features appear easily recognizable and portions of all the maps are rendered with appropriate levels of detail. However, significant problem areas do exist. Most of the techniques have difficulty with complex coastlines. In particular, the submerged coasts of British Columbia, Baffin Island, and Greenland cause problems (see Figures 7-25 to 7-30). The relaxing algorithm has approximately the right level of detail but appears too angular. The walking algorithm looks the best in these areas. All the other techniques retain much too much detail. The banding and angling methods, especially, produce chains which are very angular and jittery. In general, the form-independent methods look better on complex coastlines than the form-sensitive methods. However, the form-independent methods have noticeable problems with political boundaries such as the U.S.-Mexican border and the borders between Mexico, Guatemala, and Belize. The corners which occur along these borders are somewhat altered or smoothed with the form-independent techniques as is expected from results described above. These same general comments can be made about the maps of northern Europe (see Figures 7-31 to 7-36). The form-independent techniques look better on the complex coastlines, although the walking technique is a bit angular, but seem to over-generalize the simpler lines.

However, it is becoming clear that simply retaining critical points is insufficient to produce natural looking cartographic lines over large changes in scale.

Figure 7-25. The WDB II simplified to 1:45,000,000 with the skipping algorithm.

Figure 7-26. The WDB II simplified to 1:45,000,000 with the smoothing algorithm.

Figure 7-28. The WDB II simplified to 1:45,000,000 with the relaxing algorithm.

Figure 7-29. The WDB II simplified to 1:45,000,000 with the banding algorithm.

Figure 7-30. The WDB II simplified to 1:45,000,000 with the angling algorithm.

Figure 7-31. The WDB II simplified to 1:12,000,000 with the skipping algorithm.

Figure 7-33. The WDB II simplified to 1:12,000,000 with the walking algorithm.

Figure 7-34. The WDB II simplified to 1:12,000,000 with the relaxing algorithm.

Figure 7-36. The WDB II simplified to 1:12,000,000 with the angling algorithm.

Quantitative Evaluation

In order to refine the examination of the different techniques, the quantitative measures described in Chapter VI were applied. The measures were computed for a series of tangential tolerances ranging from 3 to 80 km. They were averaged for all chains in the database. Each technique produces a somewhat different number of points for a given tolerance, so a direct comparison of values is inappropriate. Instead, each measure is plotted as a function of the logarithm of the number of points used to represent the simplified database for each technique. The wide range of tolerance values provides a broad region of overlap between the various techniques which allow effective comparisons to be made.

The plot of the number of curvilinear segments shows little to distinguish between the techniques (See Figure 7-37a) with the exception of the angling method which causes the least reduction. A modest distinction can be made on the basis of the angularity measure, with the form-independent techniques producing greater changes in angularity than the form-sensitive methods (see Figure 7-37b). This mirrors the results of the subjective evaluation, where the walking and smoothing algorithms produce especially gentle representations of the chains. The skipping technique produces a somewhat intermediate result in terms of perceived angularity. All of the form-sensitive techniques enhance perceived angularity with the angling method producing the most angular result. The average distance between points of the original chain and the simplified chain along with the area between the two curves seem to be the most easily interpreted measures of the accuracy of representation. Jenks has used the distance measure as being the best match to perceived closeness in line representation (1989). In both these measures, the banding technique appears distinctly superior to the others (see Figure 7-37c and 7-37b). In

the context of these two measures, the relaxing technique could be considered the next best algorithm on the basis of a relatively good result in the distance measure.

Figure 7-37. Quantitative measures of simplification performance are plotted as a function of the logarithm of the number of points used to represent the original database. The four measures used are (a) the average change in the number of curvilinear segments, (b) the average change in the total angularity, (c) the average distance to points, and (d) the average area per km.

Chapter VIII

Technique Enhancement and Map Projection Effects

The examination of the various techniques has demonstrated strengths and weaknesses in all of them. In many cases, these weaknesses are systematic, i.e., certain situations, such as submerged coastlines or angles in otherwise straight political borders, cause problems for a given technique. As a result, consideration of these situations in the context of a particular algorithm may allow the development of refinements or extensions to the basic technique which can improve performance in the difficult situation. The results of Chapter VII highlight the banding technique as being the most promising. The detailed examination of its performance on individual chains shows that it does a good job of identifying critical points and of providing an efficient representation of the basic geometrical structure of the chain. In addition, the quantitative measures show that it does well in maintaining the number of curvilinear segments and angularity, and it is clearly superior to the other techniques on basis of the average distance and area measures. Thus, this method seems to be a good candidate for more intense study.

While the banding technique seems very promising, in some situations it can produce undesirable results. This is especially true along complex coastlines such as the submerged coasts of Iceland, Greenland, Norway, and British Columbia. In these areas, the procedure tends to retain only the mouths and landward extremes of each fjord, resulting in a coastline consisting only of jagged spikes. This is because the length of the fjord is greater than the tolerance while the breadth is

much less than the tolerance. The resulting chain is fuzzy when plotted at the scale associated with the tolerance and very jagged when examined at larger scale. Another difficulty, which is in some ways related to the problem with submerged coastlines, is the representation of any straight section of the chain with very few (usually two) points regardless of the length of the section. Extensions were added to the basic algorithm in an attempt to deal with these situations.

Enhancement of the Banding Technique

The banding algorithm tends to select points of high angularity. While this is often ideal for creating recognizable caricatures of objects using very few points, the procedure can produce results which appear very stilted and unrealistic. This problem becomes particularly severe in areas of submerged coastline with many fjords where substantial simplification can produce a jumble of spikes. This difficulty with spikes has been discussed by Miller and Kok (1988). Spikes are found by looking for points of very high angularity which are immediately preceded and followed by two points which are very close together as in Figure 8-1. Each point in a simplified chain is examined to determine if it is the apex of a spike. If so, it is removed from the simplified chain. The search for spikes is repeated until no additional spikes can be found. The iteration of the search can eliminate some deep, narrow bays in addition to simple spikes. This is what a cartographer might well do to simplify such complex coastlines although a cartographer might retain or remove bays based on other information, such as the presence or absence of important ports.

The second problem, the elimination of all points along a straight or nearly straight line causes difficulties with political boundaries that follow meridians, parallels, or other surveyed lines that may be considered straight on some projections

Figure 8-1. The variables used to detect a spike at point, p_i , include the vertex angle, α_i , and the distance between p_{i-1} and p_{i+1} .

but curved on others. If the basic banding algorithm is used without alteration, only the points defining the beginning and the end of the straight segment will be retained. If the data is always plotted on the same projection, this is not a problem. However, if the data is transformed to another projection where the actual line is curved, the simplified version will still be straight.

To address this difficulty, the tangential tolerance is utilized, along with the normal tolerance, ϵ_N , on which the algorithm is based. This tangential tolerance, ϵ_T , is used to force points to be retained along the line even if the maximum perpendicular distance is less than ϵ_N . This ensures that some points will be maintained along a straight line. Consequently, the simplified data can be used more reliably with other map projections. If some significant variability does occur along the chain, the maximum distance is retained even if it is less than the tolerance as shown in Figure 8-2. If the line is essentially straight, then the midpoint of the section is retained. This enhancement has two positive effects on the shape of simplified chains. First, enough points are retained along straight sections to allow subsequent transformation of the chain to other map projections where the chain might not be straight. In addition, the retention of a few non-critical points adds to the natural look of the chain. When only critical points are retained, the simplified

Figure 8-2. Retaining a point which is less than the normal tolerance, ϵ_N , from the current test line because no critical points occurred along the chain for a distance greater than the tangential tolerance, ϵ_T .

chain is often obviously a caricature. The mechanism described is a simple way of enhancing the appearance of the simplified chain in some situations.

Several examples are presented to show the improvement that can result from this refined technique. The coasts of Norway and Greenland are shown at comparable levels of simplification for the banding and refined methods (see Figures 8-3 and 8-4). The refined procedure produces a less jagged chain and a number of bays have been eliminated. In addition, some of the less complex chains (e.g., the eastern border of Norway) have a slightly smoother, more natural appearance. These two techniques do retain different numbers of points for the same tolerance because the refined technique retains additional points along relatively straight sections. This is somewhat offset by the elimination of spikes and small bays along the complex coastlines. A better comparison is provided when the simplification is constrained to retain the same number of points. The coast of British Columbia is shown simplified with these two methods using such a constraint (see Figure 8-5). Again, the refined technique produces a less jagged coastline and some of the bays have been eliminated or truncated. Enough complexity is maintained to suggest the nature

of the coastline when plotted at the scale associated with the selected tolerance. Simplifications of the coast of Iceland show similar improvements (see Figure 8-6).

Figure 8-3. The coast of Norway simplified using the (a) banding and (b) refined techniques.

Figure 8-4. The coast of Greenland simplified using the (a) banding and (b) refined techniques.

Figure 8-5. The coast of British Columbia represented with about 150 points selected by the (a) banding and (b) refined techniques.

(a)

Examination of the plots of the quantitative measures of algorithm performance shows comparable behavior by the two methods on the curvilinear segment and area measures (see Figure 8-7). The refined technique reduces angularity somewhat more than the banding technique, although not by a great deal. On the average distance to points, the banding algorithm does better. However, the refined technique is comparable to or better than any of the other techniques. As a final comparison, plots showing northern Europe were drawn at a scale appropriate to the level of simplification (see Figure 8-8). The scale associated with a tangential tolerance was determined by assuming that it would be twice the pixel size. The pixel size was determined from parameters associated with the graphics device and the graphics package which were used (see Appendix C). This choice of tolerance produces reasonable maps at approximately the right level of detail. Differences between the basic banding technique (Figures 8-8a and 8-8c) and the refined technique (Figures 8-8b and 8-8d) can be discerned on the most complex chains, such as the submerged coastlines, and on simpler chains, such as many of the political borders and the east coast of Britain. In most cases, the refined algorithm gives a superior result.

Thus, the refined technique has produced some improvement in cartographic presentation at the cost of a modest loss in quantitatively measured performance. This highlights a fundamental distinction between two aspects of numerical generalization. Statistical generalization is concerned with the accurate representation of the original data with the least possible redundancy to minimize storage and processing requirements. Quantitative measures permit the degree of distortion associated with a minimal representation to be detected and controlled. Thus, statistical generalization is concerned with the retention of mathematical measures in

Figure 8-7. The quantitative measures plotted as in Figure 7-37 with the addition of the refined technique in the heavier solid line.

Figure 8-8. Northern Europe plotted at two scales using the banding (a,c) and refined (b,d) techniques.

filtering cartographic data. In contrast, cartographic generalization is concerned with the effective display of information given the purpose and scale of the map. Consequently, the quantitative measures may be less important in determining the success or failure of a given technique in cartographic generalization (Brassel & Weibel, 1988). The refined technique described here performs less well than the banding technique in the context of statistical generalization, but would be considered an improvement in terms of cartographic generalization. The distinction between these two types of generalization is a function of scale. Over limited scale change, they are the same or very similar. Thus, mathematical measures of performance can serve to quantify the success of cartographic generalization nearly as well as for statistical generalization. It is with increasing scale change that cartographic generalization becomes clearly distinct from statistical generalization.

Other problems occur with the banding technique. The method occasionally misses corner points by a small amount. Figure 8-9 is the same as Figure 7-24e and illustrates this limitation of the technique. This difficulty has been reported in the literature (Liao, 1981) and a two-stage technique was proposed in which the banding algorithm is used to find approximate critical points and, in a second step, the regions near these potential critical points is searched for more angular points. This could improve the performance of this technique on chains such as the one shown here. Thus, the strategy of addressing the specific limitations of a technique with relatively simple refinements can be applied to numerous simplification problems.

While enhancements such as these can improve the performance of a technique, overuse of this approach could cause the remedies to different problems to interfere with each other. In addition, as more complete databases are utilized, the number and complexity of the refinements may become unmanageable. For example, the

Figure 8-9. The Mexico–Guatemala border simplified with the banding algorithm. Note that two of the corners of the border are missed with this technique.

Figure 8-10. A fjord with a river entering it is diagrammed in (a). A possible generalization of this situation is shown with the heavier line in (b).

addition of rivers to the test data set used in this study would cause the apexes of many bays or fjords to become nodes which would prevent a bay from being eliminated from the coastline unless the river entering the bay was also eliminated. This problem could be solved by collapsing the narrow bay to a single line extension of the river to the coastline with the bay removed as shown in Figure 8-10. However, this would require the identification of bays with specific geometrical characteristics. Such identification, utilizing basic computational geometry, could be time-consuming because the relative positions of points along an indeterminate length of coast, and even points from different chains along the coast, would have to be examined. The number of points which required intercomparison would increase combinatorially as the size of the features which required examination grew so that such an approach would eventually become prohibitively expensive.

A possible solution to the problem of identifying features along a chain requiring special processing may eventually grow out of work in neural networks. In a neural network, many relatively simple logic units, referred to as neurons, are interconnected in a complex way. The determination of whether a particular neuron

sends a signal down its output connections is controlled by the various stimuli it receives from its many input connections. In such networks, the interconnections are allowed to adapt to the kinds of signals the network receives from the outside world. This is an attempt to use a model of the physiology of the brain to develop machines that can perform some the complex tasks which human brains to quickly and naturally, such as pattern recognition. The application of neural networks to recognizing the structure of cartographic lines has been discussed in the literature (Mower, 1988) and may eventually have practical application in the area of map generalization by allowing a machine to identify characteristic features of lines the way a cartographer would. Specific responses to each class of characteristics could then be executed.

Map Projection Considerations

The problem of simplifying a line that is straight on one projection for use on another projection on which it will be curved, as raised in Chapter VII, hints that the use of map projections must be carefully considered in such processing, particularly at the small scales being considered in this study. Computations for the simplification techniques were performed on the normal aspect Mercator projection. This was selected initially for convenience and without careful thought. On transformation to the orthographic projection, the difficulty with straight lines became obvious. Another problem was also apparent. The level of detail maintained for the chains changed significantly between the equator and the pole. This difficulty is due to the large difference in scale between points near the equator and points near the pole on the Mercator projection used for the simplification calculations. When the original data is plotted on the orthographic projection without simplification, the level of detail appears relatively constant. However, when simplified data is plotted on this projection, the level of detail increases with latitude. This is an artifact of the constant tolerance used during the computations performed on the Mercator projection. Near the equator the specified tolerance is matched to the scale at each point with relative accuracy and the level of simplification is controlled properly. However, as the poles are approached, the scale increases while the tolerance remains the same. In effect, this means tolerance becomes smaller, and this results in the retention of greater detail as shown in Figure 8-11. Careful consideration of the effects of the map projection is needed in this situation.

An ideal map projection for performing the computational geometry calculations involved in line simplification would be a planar representation of the entire globe which had no scale, area, or angle distortion. With such a projection, all the

computations could be executed without error. However, no such projection exists (Maling, 1973). It is therefore necessary to select a projection which allows the inherent distortions to be minimized or corrected. The computations that are most commonly done involve the determination of angles (and the related trigonometric functions) and of distances. The Mercator projection is a good candidate for use in line simplification because the distortion due to angles and distances can be determined very easily on the sphere. There is no angular distortion around a point because this projection is conformal. Thus, over limited distances planar angles determined on the map projection have little distortion and no correction is required. The scale does vary significantly from point to point, but the variation is easily computed as being equal to $1/\cos \phi$, where ϕ is the latitude. If ellipsoidal formulae are being utilized, then the scale variation is $(1 - e^2 \sin^2 \phi)^{1/2}/\cos \phi$, where e is the eccentricity of the ellipsoid. Thus, the scale variation can be compensated for to a large degree. As a result the Mercator projection, when used properly, appears to

be a good choice for the type of calculations required for line simplification.

While the scale variation at a point can be compensated for exactly, the scale along a line, other than a parallel, varies and the precision of the compensation depends on the amount and sophistication of the calculation which is incorporated into each algorithm. For example, every distance calculation could involve integration of the scale along the measurement line to get an exact value. A simpler approach might be to compute the geographic coordinates of the two points with an inverse projection function and use formulae for great circle distance. Determination of great circle distance is straightforward on the sphere but complex in its own right on the ellipsoid. However, these approaches would be only slightly more exact than taking the average of the scales at the two points as representative of the scale

along the line, as long as the two points are not too far apart. While adjusting the scale for every two-point distance calculation is certainly possible, such correction terms would permeate most simplification algorithms, adding significantly to the complexity of the techniques. In this study, a less exact correction was used. The distance from point to point along each chain was computed, corrected for scale error, and summed. Thus, the linear distance along the chain was computed accurately. When compared to the unadjusted length of the chain, a correction factor, which is representative of the entire chain, is determined. This correction factor was utilized to adjust the tolerance used to simplify the chain. This projection scaling of the tolerance allows a relatively consistent density of points to be retained. This approach is not ideal, particularly when processing chains of great latitudinal extent (the Chilean coastline being an extreme example). However, it is simple and its application seems to produce approximately the right the degree of adjustment as shown in Figure 8-12. Further examination of this problem did not seem required at this time. All calculations performed on WDB II utilized the scale correction as described above. In the case of the basic banding technique, it is possible to perform the simplification on the sphere efficiently (Burt, 1989). However, this approach is applied most conveniently to tolerance band methods on the sphere; its extension to more general computations would be somewhat complex.

Figure 8-12. An orthographic projection showing the effects of compensation for the variable scale implicit in the use of the Mercator projection for performing the simplification computations. Notice the relatively constant density of points as a function of latitude.

Chapter IX

Summary and Conclusions

Generalization is one of the fundamental themes in both geography and cartography. Much of geography is concerned with discerning general patterns in and relationships between phenomena distributed over the surface of the earth. Without generalization, any attempt to describe geographic behavior is liable to be overwhelmed with detail. Geographic generalization allows much of this detail to be collapsed into a simpler pattern from which some degree of understanding can be derived. Generalization plays a similar role in cartography. Cartographic generalization's component processes of selection, simplification, classification, symbolization, and induction are the mechanism by which a particular message regarding spatial information is reliably transferred from the cartographer to the map reader. A vast amount of information can be presented in ungeneralized form, e.g., in an aerial photograph, but the interpretation of this welter of data is left to the viewer. A generalized map allows a clear statement to be made which reflects the purpose of that map. Proper cartographic generalization also reflects the limitations imposed by data quality and graphic reproduction. The other crucial factor in this process is the effect of scale. Effective generalization cannot be performed without consideration of the scale of a map because the degree of generalization changes as the map scale changes.

Map generalization in traditional cartography is based on a mixture of science and art where certain well-defined rules are obeyed but the process is largely

subjective. While fine maps can be produced in this way, particularly by experienced mapmakers, the details of traditional generalization are difficult to describe precisely. This becomes important when attempting to process cartographic information with computers. Computers must be programmed with precisely defined steps and subjective procedures can be very difficult to describe in the necessary way.

The application of computers to cartography is a relatively recent development. Computers are ideally suited to perform many of the repetitive and tedious tasks which occur in mapmaking, but the influence of the computer extends beyond the simple automation of tiresome tasks. The ability to store map data in cartographic databases on computers is one example of a development which has had a fundamental impact. In contrast to the printed map, which has limited use beyond basic display and measurement functions, a cartographic database allows map data to be used and reused many times for different purposes. The variety of functions which a given cartographic database can support is controlled by various characteristics of the database. One such characteristic is the structure of the database. Simple structures can support relatively limited functions, while more sophisticated structures allow more complex and varied manipulations. The range of applications for a cartographic database is also limited by the scale associated with the database. The scale of a cartographic database is related to the data source from which it was derived and on the mechanism by which it was transformed into digital form for use on the computer. Such a database can be only safely used over a narrow range of scales close to this nominal scale. Mapping applications that require coverage at a wide range of scales face the problem of supporting myriad cartographic databases, each designed for use over a limited range of scales. Supporting a large number of

databases can be expensive. An alternative is to attempt to extend the useful range of scales over which a single database can be applied. This requires generalization of the digital map data for use at scales significantly smaller than the nominal scale of the database.

Cartographic generalization on computers is a very complex problem that is currently being approached by examining the details of its component processes. This study has taken a close look at the simplification process as a means of extending the useful range of scales of cartographic databases. Two generally available digital cartographic databases were placed into a data structure that supported the manipulations associated with such an examination. One of these databases, the NCARCDB, was designed for very small-scale maps and was used to study simplification over modest changes in scale. The other database, a subset of WDB II, was designed for much larger scale maps and was used to examine simplification over extreme changes in scale.

Six line simplification algorithms (skipping, smoothing, walking, relaxing, banding, and angling) were implemented and used to simplify both individual chains selected from the databases as well as the complete databases. The results were examined subjectively and quantitatively. Individual chains and the corresponding simplified chains were plotted to provide a detailed view of the behavior of the various techniques as shown in Figure 9-1. Maps were made from the complete databases to provide a more synoptic view of the effectiveness of the techniques. Figure 9-2 is an example. Subjective examination of these plots and maps highlighted the strengths and weaknesses of the different algorithms. The subjective evaluation was supplemented by a quantitative evaluation based on several mathematical measures of algorithm performance. Two of these measures (total angularity and the

Figure 9-1. The border between Angola and Zaire from the NCARCDB simplified with the banding algorithm. The original chain is the dashed line and the simplified chain is the solid line.

number of curvilinear segments) were based on the relative change in a geometric characteristic of a chain as it was simplified. The other two measures (average distance and the area per unit length) quantified the deviation of the simplified chain from the original chain.

On the basis of this work, one algorithm (the banding technique) was identified as being most promising. This technique did well in many situations but did have significant difficulties in other situations, e.g., long straight chains and complex, submerged coastlines. Refinements were made to the procedure to improve the performance in these problematic areas. The use of different map projections for the maps used in the study highlighted the importance of considering the distortions produced by the projection. In small-scale mapping, the map projection has a significant effect on the distance and angular relationships between locations. As a result, line simplification techniques must be adapted to correct for this problem.

A number of conclusions can be drawn on the basis of this study:

1. Line simplification techniques can play a role in extending the useful range of simple cartographic databases. Even the poorest results produce maps that are recognizable given a general familiarity with the shapes of the countries and continents. Overly simplified maps provide acceptable caricatures of the

Figure 9-3. A coarse map simplified from WDB II that is suitable for use as a base for limited kinds of thematic maps.

degree of simplification is more closely tailored to the scale of the map, other applications are possible. Figure 9-4 shows a less simplified map of Europe that might be appropriate for the presentation of other thematic data requiring a more precise locational reference.

Figure 9-4. A more detailed map simplified from WDB II that is suitable for use as a base for more general kinds of thematic maps simplified from WDB II.

2. Of the six basic algorithms extracted from the literature, the banding (Douglas-Peucker) technique seems the most promising. The banding algorithm performs the best on the basis of quantitative measures. The results of the subjective

evaluation also support the choice of this technique for further study and enhancement.

3. Specific difficulties with a technique can often be remedied with simple enhancements to the basic procedure. For example, the jaggedness of complex coastlines produced by the banding algorithm was to some degree remedied in the refined algorithm.
4. Quantitative measures of performance provide useful means of comparing line simplification techniques. Such measures form an integral part of statistical generalization. Their application to cartographic generalization is equally valid over small changes in scale. However, with larger scale change, perceptually-oriented measures, which mimic subjective evaluation, would be more appropriate if they could be developed.
5. The effects of the choice of map projection must be considered when processing small-scale map data on the computer. The normal aspect Mercator projection has been identified as an appropriate choice on which to perform the computations associated with the simplification algorithms. Given that the Mercator projection is conformal, angle computation without correction causes minimal error. Distance computations require a correction which is easy to compute. An average scale correction for each chain was used in this study.
6. Computer processing of cartographic data places constraints on the structure of the data and, consequently, on the producers of cartographic data. The complex processing associated with geographic information systems obviously requires topologically integrated data structures with sophisticated attribute coding schemes to function effectively. Similar structuring and attributing is also necessary to perform basic cartographic functions, of which map generalization

is an important example. A process as basic as line simplification requires information in addition to the geometric description of the lines. In particular, a basic topological structure is necessary to implement this process effectively. While the databases used in this study required enhancement to meet such basic requirements, the producers of such data sets, typically government agencies, are now aware of these problems, and work is in progress to replace many of the older data sets with improved versions or to supersede them with more sophisticated

and attributed data of larger scale. The gradual increase in availability of more sophisticated databases will simplify many aspects of computer cartography by providing the underlying data in a form which supports the necessary processing.

The limitations in the performance of the techniques studied here indicates that a very wide range of further work on line simplification techniques and on the larger issue of map generalization is needed. Additional work is needed on evaluating and enhancing the currently available procedures, as well as on quantifying the effects of the tolerance and line symbolization on the effectiveness of simplification techniques. A better understanding of what constitutes a properly simplified line is necessary, particularly over large changes in scale. New algorithms may well have a role to play. More forward-looking work will eventually be required on the recognition of cartographically significant features in a digital database so that the most appropriate techniques for simplification can be applied to them. The other component processes of generalization also require a great deal of work.

to granitic rules to produce a system capable of acceptable map generalization.

The selection of appropriate algorithms for various simplification and generalization situations could easily produce a vast number of rules which would have to be applied to identify the most appropriate way to solve each problem. The development of rule-based systems will most likely play a role in this area. Early work has already been done (Nickerson & Freeman, 1986), although the basic set of techniques, rules for their application, and methods of identifying the relevant characteristics of the problem at hand must be developed more extensively before robust, operational systems of this type can be built. An alternative approach may follow along the lines suggested by McMaster and Shea (1988), in that the cartographer may never be eliminated from the generalization process but will take the role of creatively identifying and applying the most appropriate techniques from an continuously expanding suite which would include numerous line simplification algorithms.

References

- Anderson, J.R., Hardy, E.E., Roach, J.T., and Witmer, R.E. (1976). *A Land-use and Land-cover Classification System for Use with Remote-Sensor Data*, U.S. Geological Survey Professional Paper 964, U.S. Government Printing Office.
- Attneave, F. (1954). "Some Informational Aspects of Visual Perception", *Psychological Review*, 61, pp. 183-193.
- Baugh, I.D.H. and Boreham, J.R. (1976). "Measuring the Coastline from Maps: A Study of the Scottish Mainland", *The Cartographic Journal*, 13, pp. 167-171.
- Beard, M.K. (1987). "How to Survive on a Single Detailed Database, *Proceedings of the International Symposium on Computer-Assisted Cartography, (AUTO CARTO VIII)*, ACSM, ASP, pp. 211-220.
- Belles, R.D. (1985). *Generating Color Terrain Images in an Emergency Response System*, Lawrence Livermore National Laboratory Report UCRL-92361.
- Berry, J.K. (1987). "Fundamental Operations in Computer-Assisted Map Analysis", *International Journal of Geographical Information Systems*, 1, pp. 119-136.
- Boyle, A.R., Dangermond, J., Marble, D., Simonett, D.S., Tomlinson, R.F. (1983). *Review and Synthesis of Problems and Directions for Large Scale Geographic Information System Development*, NASA Contract Report 166563, Environmental Systems Research Institute.
- Brassel, K.E. (1985). "Strategies and Data Models for Computer-Aided Generalization", *International Yearbook of Cartography*, 25, pp. 11-28.
- Brassel, K.E., and Weibel, R. (1988). "A Review and Conceptual Framework of Automated Map Generalization", *International Journal of Geographical Information Systems*, 2, pp. 229-244.
- Brophy, D.M. (1973). "An Automated Methodology for Linear Generalization in Thematic Cartography", *Proceedings of American Congress on Surveying and Mapping, 33rd Annual Meeting*, pp. 300-314.
- Burt, J.E. (1989). "Line Generalization on the Sphere", *Geographical Analysis*, 21, pp. 68-74.
- Butenfield, B.P. (1989). "Scale-Dependence and Self-Similarity in Cartographic Lines", *Cartographica*, 26, pp. 79-100.

- Campbell, J. (1983). *Mapping the Land: Aerial Imagery for Land Use Information*, Association of American Geographers.
- Campbell, J. (1984). *Introductory Cartography*, Prentice-Hall.
- Chen, Z.T. (1987). "Contour Generalization by a 3-Dimensional Spatial Low-Pass Filtering", *Proceedings of the Second Annual International Conference, Exhibits and Workshops on Geographic Information Systems, (GIS '87)*, ACSM, ASP, 1, pp. 375-386.
- Chrisman, N.R. (1983). "Epsilon Filtering: a Concept for Automated Scale Changing", *Proceedings of American Congress on Surveying and Mapping, 43th Annual Meeting*, pp. 322-331.
- Coppock, J.T., and Anderson, E.K. (1987). "Editorial Review", *International Journal of Geographical Information Systems*, 1, pp. 3-11.
- Dangermond, J. (1983). "A Classification of Software Components Commonly Used in Geographic Information Systems", *Design and Implementation of Computer Based Geographic Information Systems*, IGU Commission on Geographical Data Sensing and Processing.
- Dangermond, J., Derrenbacher, W., and Harnden, E. (1985). "World Map of Desertification Hazards: Data Base Development and Application, United Nations Environment Programme, 1983-1984", *Technical Papers, 45th Annual Meeting*, ACSM, pp. 288-297.
- Dangermond, J., and Harrison, J. (1987). "Urban Geographic Information Systems: The San Diego Experience", *Proceedings of the Second Annual International Conference, Exhibits and Workshops on Geographic Information Systems, (GIS '87)*, ACSM, ASP, 2, pp. 387-395.
- Darragh, I. (1979). "How Computers are Improving Our Maps", *Canadian Geographic*, October, pp. 44-49.
- Davis, H.F. (1967). *Introduction to Vector Analysis*, Allyn & Bacon.
- Davis, L.S. (1977). "Understanding Shape: Angles and Sides", *IEEE Transactions on Computers*, C-26, pp. 236-242.
- Dent, B.D. (1985). *Principles of Thematic Cartography*, Addison-Wesley.
- Dent, B.D. (1987). "Continental Shapes on World Projections: The Design of a Poly-Centred Oblique Orthographic World Projection", *The Cartographic Journal*, 24, pp. 117-124.
- Dettori, G. and Falcidieno, B. (1982). "An Algorithm for Selecting Main Points on a Line", *Computers and Geosciences*, 8, pp. 3-10.

Dickerson, M.H., Gudiksen, P.H., and Sullivan, T.J. (1983). *The Atmospheric Release Advisory Capability*, Lawrence Livermore National Laboratory Report UCRL 52802-83.

Dobson, J.E. (1983). "Automated Geography", *Professional Geographer*, 35, pp. 135-143.

Douglas, D., and Peucker, T. (1973). "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature", *The Canadian Cartographer*, 10, pp. 112-122.

Dutton, G.H. (1981). "Fractal Enhancement of Cartographic Line Detail", *The American Cartographer*, 8, pp. 23-40.

French, R.L. (1987). "Automobile Navigation in the Past, Present, and Future", *Proceedings of the International Symposium on Computer Assisted Cartography*

phy, (AUTO CARTO VIII), ACSM, ASP, pp. 542-551.

Goldberg, M. (1984). *Geocoded Data Structures and Their Applications to Earth Science Investigations*, NASA Technical Memorandum 86113, Goddard Space Flight Center, Greenbelt, Maryland.

Goodchild, M.F. (1980). "Fractals and the Accuracy of Geographic Measures", *International Association for Mathematical Geology Journal*, 12, pp. 85-98.

Gorny, A.J. (1977). *World Data Bank II, Volume 1 — North America, and General User Guide*, National Technical Information Service, PB-271 869.

Haggett, P. (1967). "Maps as Models", *Models in Geography*, Methuen & Company, Ltd., pp. 698-713.

Hardaway, G.M., and Porter, E.D. (1985). "The Terrain Analyst Work Station".

- Lai, P. (1985). "Mis-application of Automated Mapping, An Assessment", *Proceedings of the International Symposium on Computer-Assisted Cartography*, (AUTO CARTO VII), ACSM, ASP, pp. 322-326.
- Lang, T. (1969), "Rules for the Robot Draughtsmen", *The Geographical Magazine*, 42, pp. 50-51.
- Langran, G. (1984). *Map Design for Computer Processing: Literature Review and DMA Product Critique*, Naval Ocean Research and Development Activity Report 72, NSTL Mississippi.
- Laskowski, P.H. (1989). "The Traditional and Modern Look at Tissot's Indicatrix", *The American Cartographer*, 16, pp. 123-133.
- Liao, Y. (1981). "A Two-Stage Method of Fitting Conic Arcs and Straight Line Segments to Digitized Contours", *Proceedings of IEEE Pattern and Image Processing Conference*, pp. 224-229.
- Luman, D.E., (1987). "Applying USGS Digital Line Graph Data in a Microcomputer Environment", *The American Cartographer*, 14, pp. 321-343.
- Maling, D.H. (1968). "How Long is a Piece of String", *The Cartographic Journal*, 5, pp. 146-156.
- Maling, D.H. (1973). *Coordinate Systems and Map Projections*, George Philip and Son.
- Mandelbrot, B.B. (1967). "How Long is the Coast of Britain? Statistical Self-Similarity and Fractal Dimension", *Science*, 156, pp. 636-638.
- Marble, D.F., and Lauzon, J.P. (1984). "Development of a Conceptual Model of the Manual Digitizing Process", *Proceedings of the International Symposium on Spatial Data Handling*, pp. 146-171.
- Marino, J. (1979). "Identification of Characteristic Points Along Naturally Occurring Lines: an empirical study", *The Canadian Cartographer*, 16, pp. 70-80.
- Mausel, P.W. (1985). "Characteristics and Techniques of Computer-Assisted Processing of Spectral Data", *The Surveillant Science: Remote Sensing of the Environment*, John Wiley and Sons, pp. 292-309.
- McLaurin, J.D. (1985). "The U.S. Geological Survey 1:100,000-scale Digital Cartographic Data Base". Paper presented at the International Symposium on Computer-Assisted Cartography (AUTO CARTO VII), Washington, DC.
- McMaster, P., Haywood, P.E., and Sowton, M. (1986). "Digital Mapping at Ordnance Survey", *AutoCarto London*, 1, pp. 13-23.
- McMaster, R.B. (1986). "A Statistical Analysis of Mathematical Measures for Linear Simplification", *The American Cartographer*, 13, 2, pp. 103-116.

- McMaster, R.B. (1987a). "A Automated Line Generalization", *Cartographica*, 24, pp. 74-111.
- McMaster, R.B. (1987b). "The Geometric Properties of Numerical Generalization", *Geographical Analysis*, 19, pp. 330-346.
- McMaster, R.B. (1989). "The Integration of Simplification and Algorithms in Line Generalization", *Cartographica*, 26, pp. 101-121.
- McMaster, R.B., and Shea, K.S. (1988). "Cartographic Generalization in a Digital Environment: A Framework for Implementation in a Geographic Information System", *Proceedings of the GIS/LIS Conference*, 1, pp. 240-249.
- Miller, S.W., and Kok, A.L. (1988). "Selective Generalization of Depth Contours for Production of Nautical Charts", *Proceedings of American Congress on Surveying and Mapping, 48th Annual Meeting*, pp. 66-75.
- Monmonier, M.S. (1982). *Computer-assisted Cartography, Principles and Prospects*, Prentice-Hall.
- Monmonier, M.S. (1983). "Raster-mode Area Generalization for Land Use and Land Cover Maps", *Cartographica*, pp. 65-91.
- Monmonier, M.S. (1986). "A Practicable Model of Cartographic Generalization", *AutoCarto London*, 2, pp. 257-266.
- Montanari, U. (1970). "A Note on Minimal Length Polygon Approximations to a Digitized Contour", *Communications of the Association for Computing Machinery*, 13, pp. 41-47.
- Morrison, J.L. (1974). "A Theoretical Framework for Cartographic Generalization With Emphasis on the Processing of Symbolization", *International Yearbook of Cartography*, 14, pp. 123.
- Morrison, J.L. (1975). "Map Generalization: Theory, Practice, and Economics", *Proceedings of the International Symposium on Computer-Assisted Cartography, (AUTO CARTO II)*, ACSM, ASP, pp. 99-112.
- Morrison, J.L. (Ed.). (1988). "The Proposed Standard for Digital Cartographic Data", *The American Cartographer*, 15.
- Mower, J.E. (1988). "A Neural Network Approach to Feature Recognition Along Cartographic Lines", *Proceedings of the GIS/LIS Conference*, 1, pp. 250-255.
- Muller, J.C. (1987a). "Optimum Point Density and Compaction Rates for the Representation of Geographic Lines", *Proceedings of the International Symposium on Computer-Assisted Cartography, (AUTO CARTO VIII)*, ACSM, ASP, pp. 221-229.
- Muller, J.C. (1987b). "Fractals and Automated Line Generalization", *The Cartographic Journal*, 24, pp. 27-34.

- Nickerson, B.G. (1988). "Automated Cartographic Generalization for Linear Features", *Cartographica*, 25, pp. 15-66.
- Nickerson, B.G., and Freeman, H. (1986). "Development of a Rule-Based System for Automatic Map Generalization", *Proceedings of the Second International Symposium on Spatial Data Handling*, pp. 537-556.
- Oommon, B.J., and Kashyap, R.L. (1983). "Scale Preserving Smoothing of Islands and Lakes", *Proceedings of the International Symposium on Computer-Assisted Cartography, (AUTO CARTO VI)*, ACSM, ASP, 1, pp. 243-251.
- Opheim, H. (1982). "Fast Reduction of a Digitized Curve", *Geo-Processing*, 2, pp. 33-40.
- Pannekoek, A.J. (1962). "Generalization of Coastlines and Contours", *International Yearbook of Cartography*, 2, pp. 55-74.
- Pavlidis, T. (1983). "Curve Fitting with Conic Splines", *ACM Transactions on Graphics*, 2, pp. 1-31.
- Perkal, J. (1958). "An Attempt at Objective Generalization", Translated from: "Proba Obiektywnej Generalizacji", *Geodezja i Kartografia*, VII, 2, pp. 130-142; in Discussion Paper #9, Michigan Inter-University Community of Mathematical Geographers, University of Michigan — Ann Arbor, 1965.
- Peruzzi, D. (1975). "A Cartographic Challenge in the Multiscale Mapping of Land Use", *International Yearbook of Cartography*, 15, pp. 133-142.
- Peucker, T.K., and Chrisman, N. (1975). "A Cartographic Data Structures", *The American Cartographer*, 2, pp. 55-69.
- Peuquet, D.J. (1983). "A Hybrid Structure for the Storage and Manipulation of Very Large Spatial Data Sets", *Computer Vision, Graphics, and Image Processing*, 24, pp. 14-27.
- Ramer, U. (1972). "An Iterative Procedure for the Polygonal Approximation of Plane Curves", *Computer Graphics and Image Processing*, 1, pp. 244-256.
- Rechlin, A.I.M. (1972). "The Epsilon Generalization", *Annals of the Association of American Geographers*, pp. 578-581.
- Reisinger, T.W., and Davis, C.J. (1987). "Integrating Geographic Information and Decision Support Systems: A Forestry Industry Application", *Proceedings of the Second Annual International Conference, Exhibits and Workshops on Geographic Information Systems, (GIS '87)*, ACSM, ASP, 2, pp. 578-584.
- Rhind, D.W. (1973). "Generalization and Realism Within Automated Cartography", *The Canadian Cartographer*, 10, pp. 51-62.
- Richardson, L.F. (1961). "The Problem of Contiguity: an Appendix to the Statistics of Deadly Quarrels", *General Systems Yearbook*, 6, pp. 139-187.

- Robinson, A.H., Sale, R.D., Morrison, J.L., and Muehrcke, P.C. (1984). *Elements of Cartography*, John Wiley and Sons.
- Rosenberg, B. (1972). "The Analysis of Convex Blobs", *Computer Graphics and Image Processing*, 1, pp. 183-192.
- Rosenfeld, A., and Johnston, E. (1973). "Angle Detection on Digital Curves", *IEEE Transactions of Computers*, C-22, pp. 875-878.

on Digital Curves", *IEEE Transactions on Computers*, C-24, pp. 940-941.

- Rutkowski, W.S. (1981). "Shape Segmentation Using Arc/Chord Properties", *Computer Graphics and Image Processing*, 17, pp. 114-129.
- Salichtchev, K.A. (1976). "History and Contemporary Development of Cartographic Generalization", *International Yearbook of Cartography*, 16, pp. 158-172.
- Sampson, P.D. (1982). "Fitting Conic Sections to Very Scattered Data: An Iterative Refinement to the Bookstein Algorithm", *Computer Graphics and Image Processing*, 18, pp. 97-108.

- Selden, D.D. (1985). "An Approach to Evaluation and Benchmark Testing of Cartographic Data Production Systems", *Proceedings of the International Symposium on Computer Assisted Cartography (AUTO CARTO VII)* ACSM, ASP.

- Snyder, J.P. (1987). *Map Projections: A Working Manual*, U.S. Geological Survey Professional Paper 1395, U.S. Government Printing Office.
- Srnka, E. (1970). "The Analytic Solution of Regular Generalization in Cartography", *International Yearbook of Cartography*, 10, pp. 48-60.
- Starr, L.E. (1986). "Mark II: The Next Step in Digital Systems Development at the U.S. Geological Survey", *AutoCarto London*, 2, pp. 200-205.
- Sukhov, V.I. (1970). "Application of Information Theory in Generalization of Map Contents", *International Yearbook of Cartography*, 10, pp. 41-47.
- Taketa, R.A. (1974). *Generalization in Land-Use Mapping*, M.A. Thesis, San Jose State University.
- Tobler, W.R. (1959). "Automation and Cartography", *The Geographical Review*, 49, pp. 526-534.
- Tobler, W.R. (1964). "An Experiment in the Computer Generalization of Maps", Technical Report #1, ONR Task No. 389-137, Contract Nonr 1224(48), Office of Naval Research Geography Branch.
- Tobler, W.R. (1966). "Numerical Map Generalization and Notes on the Analysis of Geometric Distributions", Discussion Paper #8, Michigan Inter-University Community of Mathematical Geographers, University of Michigan — Ann Arbor.
- Tobler, W.R. (1976). "A Transformational View of Cartography", *Proceedings of the International Symposium: The Changing World of Geodetic Science*, The Ohio State University, pp. 270-280.
- Töpfer F. and Pillewizer, W. (1966). "The Principles of Selection, A Means of Cartographic Generalisation", *The Cartographic Journal*, 3, pp. 10-16.
- Vanicek, P., and Woolnough, D.F. (1975). "Reduction of Linear Cartographic Data Based on Generation of Pseudo-Hyperbolae", *The Cartographic Journal*, 12, pp. 112-119.
- van Roessel, J.W. (1987). "Design of a Spatial Data Structure Using the Relational Normal Forms", *International Journal of Geographical Information Systems*, 1, pp. 33-50.
- Visvalingam, M. (1989). "Cartography, GIS and Maps in Perspective", *The Cartographic Journal*, 26, pp. 26-32.
- Walker, H. (1985). "Spatial Data Handling Capabilities in Emergency Response", *Geographic Information Systems in Government*, 1, pp. 295-307.
- Walker, H. (1989). "Use of the 1:2,000,000 Digital Line Graph Data in Emergency Response", *Proceedings of the International Symposium on Computer-Assisted Cartography, (AUTO CARTO IX)*, ACSM, ASP, pp. 472-482.

- Walton, R.G., (1985). *Strategies for Associating Data and Location in a Geographic Information System*, Master of Science Project Report, University of California, Los Angeles.
- Waugh, T.C., and Healey, R.G. (1987). "The GEOVIEW Design: A Relational Approach to Geographical Data Handling", *International Journal of Geographic Information Systems*, 1, pp. 101-118.
- White, E.R. (1985). "Assessment of Line Generalization Algorithms Using Characteristic Points", *The American Cartographer*, 12, pp. 17-27.
- Wilford, J.N. (1981). *The Mapmakers*, Vintage.
- Williams, C.M. (1978). "An Efficient Algorithm for the Piecewise Linear Approximation of Planar Curves", *Computer Graphics and Image Processing*, 8, pp. 286-293.
- Williams, C.M. (1981). "Bounded Straight-Line Approximation of Digital Planar Curves and Lines", *Computer Graphics and Image Processing*, 16, pp. 370-381.
- Wright, J.K. (1942). "Map Makers are Human; Comments on the Subjective in Maps", *The Geographical Review*, 32, pp. 527-544.
- Yoeli, P. (1975). "Methodology of Computer-Assisted Cartography", *Display & Analysis of Spatial Data*, John Wiley, pp. 130-136.
- Zoraster, S., Davis, D., and Hugus, M. (1984). *Manual and Automated Line Generalization and Feature Displacement*, Engineering Topographic Laboratory Report ETL-0359.

Appendix A

Coordinate Systems and Vectors

Coordinate systems provide a means of specifying position in a systematic way. A coordinate system in one dimension allows positions along a line to be described exactly. Such a system requires the selection of a starting point, or *origin*, at an arbitrary position along the line. The origin is typically assigned the number zero (0). In addition, a unit of distance must also be selected. The distance unit may also be arbitrary. However, certain units have been established as conventions for the measurement of length. Common units of distance are meters, kilometers, feet, and miles. Any point along the line can be uniquely specified by giving the distance from the origin. As a line is normally considered to be infinitely long, the distance may be either positive or negative, i.e., one direction from the origin must be designated as positive with the other direction being negative. A line with a coordinate system is shown in Figure A-1.

Because a plane is two-dimensional (a map can be considered a portion of a plane), positions on a plane can be described with reference to one-dimensional coordinate systems defined on any two non-parallel lines contained within the plane. Utilizing a planar coordinate system is usually simplest if the two lines are perpendicular and if the origin of the coordinate systems along both lines begin at the point of intersection of the lines (two non-parallel lines on a plane must intersect at a unique point). Further simplification occurs if the distance units along both lines are the same. Such a coordinate is usually called a *rectilinear* coordinate system or

Figure A-1. A line with a coordinate system. In this case, positive coordinates extend to the right and negative coordinates to the left. A coordinate of 3 indicates a location three distance units to the right of the origin, and a coordinate of -1.5 designates a position one and a half distance units to the left of the origin.

Figure A-2. A two-dimensional coordinate system with the x -coordinates increasing to the right and the y -coordinate increasing upwards. Two sample locations are shown with their coordinate representations.

a *Cartesian* coordinate system. The two lines defining the coordinate system are referred to as *axes*. Such a coordinate system is shown on Figure A-2.

Any point in the plane can be specified by listing the distance along one axis followed by the distance along the other axis. These two numbers are called a *coordinate pair*. In mathematical expressions, it is common to express the two

numbers forming a coordinate pair as (x, y) , where x is the coordinate along one axis and y is the coordinate along the other axis. Two such points are indicated in Figure A-2 in this format.

There are numerous other ways of specifying locations on a plane, some of which rely on geometrical figures other than a straight line. For example, in polar coordinates, a point is selected as a distance origin and a direction, or axis, from that point is chosen as an angular origin. Each point in the plane is a specific distance, or radius, from the distance origin. In addition, the line from the distance origin to the point forms a specific angle with the angular origin. These two coordinates, the radius and the angle, specify a unique location on the plane. If (1) the origin of the polar coordinate system is at the origin of a Cartesian system, (2) the angular origin of the polar system corresponds with the x -axis of the Cartesian system, and (3) the radial distance of the polar system has the same distance unit as the Cartesian system, then the coordinates in the two systems are simply related by:

$$r^2 = x^2 + y^2, \quad (\text{A} - 1\text{a})$$

$$\tan \theta = \frac{y}{x}, \quad (\text{A} - 1\text{b})$$

where r is the radial distance in the polar system from the origin to the point (x, y) and θ is the angle from the origin to (x, y) . Equation A-1a is a restatement of the Pythagorean theorem and Equation A-1b is the definition of the tangent of an angle. These relations are illustrated in Figure A-3.

Vectors

A vector can be considered to be a mathematical object which has a direction and a magnitude. This differentiates a vector from a scalar quantity which has only magnitude, typically described by a real number. A vector is represented

Figure A-3. (a) A polar coordinate system. (b) The relation between polar coordinates and Cartesian coordinates. In this example, $x = 2.83$ and $y = 2.83$, so $r = \sqrt{x^2 + y^2} = \sqrt{8 + 8} = 4$ and $\arctan(y/x) = \arctan 1 = 45^\circ$.

geometrically as an arrow which can be moved freely on a plane as long as its length and direction do not change. Algebraically, a vector, $\mathbf{u}$, is represented by a pair of components, (u_x, u_y) , which represent the lengths of the projections of the vector onto the x - and y -coordinate axes, respectively. This is illustrated in Figure A-4a. A vector can be defined by any two points in a coordinate system, i.e., given two points, p_1 and p_2 , located at (x_1, y_1) and (x_2, y_2) , respectively, the vector $\mathbf{u}$

from p_1 to p_2 can be described as follows:

$$\mathbf{u} = (u_x, u_y) = (x_2 - x_1, y_2 - y_1). \quad (\text{A} - 2)$$

An example is shown in Figure A-4b.

There are a number of common operations involving vectors. For example, the length of a vector is a scalar value, $(\|\mathbf{u}\|)$, which is given by the

Figure A-4. (a) A vector and its components. (b) A vector between two points in a coordinate system. Here $\mathbf{u} = (8-4, 14-3) = (4, 11)$.

situations. This length is computed from the vector components u_x and u_y via the Pythagorean theorem:

$$\begin{aligned} \|\mathbf{u}\| &= \sqrt{u_x^2 + u_y^2} \\ &= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \end{aligned} \quad (\text{A} - 3)$$

A vector can also be lengthened or shortened via multiplication by a scalar. This is accomplished by multiplying each of the vector components by the scalar, so

$$c\mathbf{u} = (cu_x, cu_y), \quad (\text{A} - 4)$$

where c is a scalar. It can be shown from Equations A-3 and A-4 that

$$\|c\mathbf{u}\| = c\|\mathbf{u}\|, \quad (\text{A} - 5)$$

which implies that $\mathbf{u}$ can be made longer or shorter by choosing $c > 1$ or $c < 1$, respectively. Two vectors can also be combined in various ways. One way is vector

Figure A-5. Vector addition.

addition. Two vectors, $\mathbf{u}$ and $\mathbf{v}$, are added by adding their respective components. That is,

$$\mathbf{u} + \mathbf{v} = (u_x + v_x, u_y + v_y). \quad (\text{A} - 6)$$

The geometric interpretation of vector addition is shown in Figure A-5.

Vectors can be multiplied in two ways, the dot product and the cross product. The dot product produces a scalar value, hence its alternate name of scalar product, and is expressed as follows:

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos \theta, \quad (\text{A} - 7)$$

where θ is the angle between the two vectors. The relation between the vectors and the angle is illustrated in Figure A-6. It should be pointed out that if $\mathbf{u} \cdot \mathbf{v} = 0$ when $\|\mathbf{u}\| \neq 0$ and $\|\mathbf{v}\| \neq 0$, then $\cos \theta = 0$ which means that the two vectors are perpendicular. It can be shown that if Equation A-6 is expressed in term of vector components, then the following is true:

$$\mathbf{u} \cdot \mathbf{v} = u_x v_x + u_y v_y. \quad (\text{A} - 8)$$

Figure A-6. Two vectors and their included angle, θ .

Figure A-7. The parallelogram associated with a cross product.

In the cross product, two vectors, $\mathbf{u}$ and $\mathbf{v}$, are combined to give another vector. The resultant vector, an alternative name is the vector product, is perpendicular to both $\mathbf{u}$ and $\mathbf{v}$. This means that the new vector is perpendicular to the plane defined by $\mathbf{u}$ and $\mathbf{v}$. The three-dimensional aspects of vector manipulation are not being considered here, so only the relevant formulae will be presented. First,

$$\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\|\|\mathbf{v}\| \sin \theta, \quad (\text{A} - 9)$$

which can be interpreted as the area of the parallelogram defined by the two vectors as shown in Figure A-7. In component form,

$$\|\mathbf{u} \times \mathbf{v}\| = u_x v_y - u_y v_x. \quad (\text{A} - 10)$$

These equations provide the basis for much of the computational geometry used in processing cartographic line data. For a fuller discussion, many references are available, e.g., Davis (1967).

Appendix B

Computer Storage Terminology

In order to discuss computer data structures effectively, it is necessary to introduce some of the terminology associated with computers and the information they store. The basic unit of data storage in all current computers is the *bit*. This is an acronym for *binary digit*, which is a component of the machine or a portion of a data storage medium which has only two states, normally referred to as 0 and 1. The state of a bit can be set, or *written*, by some signal, typically electrical or magnetic, based on processes taking place elsewhere in the computer or storage device. The state of a bit can also be sensed, or *read*, so that the information it represents can be used to control or take part in processes occurring in the computer or data storage device. In some cases, a bit can be written once and then read an indefinite number of times. Such a bit would be considered *read-only*. In other cases, a bit can be both written and read an indefinite number of times.

The actual mechanism for storing a bit of data varies greatly from computer to computer and differs even more between different types of data storage device. Most current computers have millions of bits available to store information and some data storage devices can maintain billions or trillions of bits. The individual bits are organized into larger structures to facilitate processing. For example, groups of eight bits are often manipulated as units called *bytes*. A byte can store 2^8 , or 256, different values ranging from all eight bits set to 0 through all the bits set to 1. Bits

are also grouped into large structures (e.g., 16, 32, or 64 bits) called *words*. Larger groupings also occur but vary widely from computer system to computer system.

Much of the power of modern computers stems from the ability to use such bits to stand for many different kinds of information. Perhaps the most straightforward use of these bits is to represent numbers. So the 256 values associated with a byte are commonly associated with the integers from 0 to 255. Using somewhat more involved interpretations of the bits, it is also possible to represent negative integers. Real numbers are usually represented (to a limited accuracy) in 16 or 32-bit words. The kinds of data that can be stored are not limited to numbers; almost any type of information can be represented by groups of bits. For example, it is very common

to associate the various characters that are available on a typewriter keyboard with specific values of a byte. So a byte with a value of 65 might represent the character 'A', the value 66 indicates 'B', etc. As a result, anything that can be written at a typewriter can be stored as a string of bytes within the computer or on a data storage device. As a geographic example, 256 different kinds of geographic features that might appear on a map could be distinguished using a code based on a single byte.

The process of encoding information is crucial to most applications of computers. Because any such encoding is ultimately based on binary digits, the encoding is referred to as a *digital* representation of the data. In addition to storing the data to be processed by a computer, a word in the computer can also contain processing instructions for the computer. For example, a word in one part of the computer can tell the computer to add the value in a second word to the value in a third word and to put the result in a fourth word. The flexibility that is the result of the ability to represent all kinds of data and to control the functions of the computer

from the computer's storage allows all manner of procedures to be performed by the computer.

Words are not the largest structure that can be stored and manipulated in a computer system. Larger structures can be build up out of bytes and words in a flexible way to meet the needs of an application. A sequence of computer words can be accessed as a unit called an array. Individual words in the array can referenced by an index which contains a count of the number of words beyond the beginning of the array where the individual word of interest resides. Simple lists of numbers or other data are often stored in such arrays, especially if the numbers are normally processed in sequence. More complex structures can be formed by making arrays of arrays. Each element of a sequence in memory is not a single word but an array, so that one array follows immediately after another. A simple array is considered to have a single dimension and requires one index to locate any given element. An array of arrays has two dimensions and uses two indices to locate an element. This concept can be extended further and is heavily used in some kinds of computer processing. Other kinds of higher level groupings are also common.

A sequence of word and bytes can be collected into a structure called a record where each entry, or *field*, in the record is related to the other entries in the same record. This is very common in many databases. For example, a corporation might build a record for each employee. This record would be made up of computer words and bytes that might store the employee's name, department, job classification, salary, and start date. A record with the same structure would be built for all employees. All the records could be listed sequentially forming an array of records. Records, arrays and combinations of these and other structures can be made arbi-

trarily complex in order to handle the infinitely varied kinds of information that can be processed on computers.

A larger structure is typically used to manipulate groupings of data on computer systems and storage devices. Any data can be grouped and saved in a *file*. Files may be very small, e.g., a file could contain a single byte storing one character, or extremely large, e.g., another file might contain all the geographic features in Asia at a scale of 1:3,000,000. A common operation on almost all computer systems is to copy or move files from one device to another, e.g., from a permanent storage system that responds slowly to a faster system that is only available for temporary use. While some databases are stored in a single file, many cartographic databases include such vast quantities of information that they must be broken down into many files that are small enough to be conveniently handled on typical computer systems.

Appendix C

Implementation Details

The software used in performing this study was developed by the author. It was executed on a VAX-8550TM, which is produced by Digital Equipment Corporation (DEC). This machine is rated as having an execution speed of 6–7 million instructions per second. The operating system supported on this machine was VMSTM, also a DEC product. The software was written in DEC's FORTRAN language. The graphical plots were produced on a QMS Lasergraphics 2400, a 300 dot per inch laser printer, using Grafpak-GKSTM produced by Advanced Technology Center. GKS (Graphical Kernel System) is an International Standards Organization and American National Standards Institute standard for two-dimensional graphics. This standard is largely oriented towards vector graphics, and the functionality provided by GKS is appropriate for many cartographic purposes. Much of the work required to perform this study, as well as to produce the final map products, required interactive graphical manipulation of the cartographic databases. The interactive graphics was performed using GKS to drive a Tektronix 4128TM graphics terminal with a mouse used for graphics input. The Tektronix 4128 has 1280 by 1024 pixels and supports 256 colors selected from a palette of over 16,000,000 colors. The map projection routines utilized had been developed by the author before this study based on equations in works by Snyder (1982, 1987). The line intersection routine used in implementing the area measure of line simplification algorithm performance had also been developed previously.

In some cases, problems in the graphics system affected details of the maps. The change in symbolization between the comparison of individual chains over modest scale change and the comparison over extreme scale change was necessitated by a limitation in both the graphics hardware and software used for this work. When plotting a line style (such as a dotted line) for a very complex line represented by many points, the graphics system failed completely. The use of solid lines was the only option given the available resources. Another case in which the limitations of the systems utilized affected this work occurred when the memory allocation on the computer was exceeded. This happened during the processing to create the topological structure for WDB II. As a result, a simple memory management had to be implemented which was adequate but very slow.

This document was produced using $\text{\TeX}$ running on the VAX-8550 and was printed on the same laser printer used for the maps. Most of the diagrams were produced using a Macintosh IITM.