STATISTICAL CHARACTERIZATION OF AREA AND DISTANCE IN ARC-NODE GEOGRAPHIC INFORMATION SYSTEMS

by

Stephen P. Prisley

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in
Forestry

APPROVED:

J. L. Smith, Chairman

T. G. Gregoire

J. A. Sullivani

J. J. Buhyoff

M. R. Reynolds

September, 1989

Blacksburg, Virginia

STATISTICAL CHARACTERIZATION OF AREA AND DISTANCE IN ARC-NODE GEOGRAPHIC INFORMATION SYSTEMS

by

Stephen P. Prisley

Committee Chairman: James L. Smith

Forestry

(ABSTRACT)

While Geographic Information Systems (GIS) have proven to be effective tools for the management and analysis of forest resources data, estimates of the reliability of area and distance measures computed in GIS have been lacking. Using fairly weak assumptions regarding the variability of point location errors, expressions for computing the mean, variance and covariance of polygon area, and an approximate distribution for distance are derived.

Assumptions about point location errors include unbiasedness, independence between X and Y coordinate errors, known and equal variance of errors in X and Y coordinates, and correlation between errors at adjacent points. For the derivation of distance from a point to a line, the assumption of normality of errors is added. Because the variance of polygon area that was derived depends on the location of the centroid, a centroid location which minimizes polygon variance was defined.

After the mean and variance of polygon area errors were obtained, polygon area was shown to be approximately normally distributed in a simulation of errors in regular polygons. Distance between a point and a line consists of two cases: distance from the point to a vertex of the line, and perpendicular distance to a line segment. The square of vertex distance was shown to be distributed as a non-central chi-square random variable when normal errors are assumed. The normal distribution was demonstrated to be a reasonable approximation for perpendicular distance under similar assumptions.

As an application of the polygon variance and covariance formulas, the variability of value of a tract of land was estimated, based upon fixed per-acre values and assumptions regarding variability of location errors. Under moderate assumptions of variability and correlation, the coefficient of variation of mean tract value was 8%. To demonstrate the application of the distribution of distance, a probabilistic point-in-polygon analysis was performed using timber cruise plot locations in a timber stand map. Over half of the plots were ambiguously located when evaluated using the most liberal set of assumptions tested. The advantages and disadvantages of the models developed herein are discussed.

ACKNOWLEDGEMENTS

This work is the result of contributions by a number of people. First, I would like to express my gratitude to my major professor, Dr. James L. Smith. Jim's encouragement brought me back to school to do something I have wanted to do for quite some time. I would also like to acknowledge the guidance and support of my graduate committee: Tim Gregoire, John Scrivani, Marion Reynolds, and Greg Buhyoff. I am especially indebted to Tim Gregoire, without whose considerable statistical expertise and careful review of numerous drafts this work would not have been completed. Beyond the graduate committee, I have benefitted greatly from the experience and insight of many members of the faculty of the Virginia Tech Department of Forestry.

There have been some very difficult times during the past forty months, and I thank my fellow graduate students, roommates, parents, and friends for their help along the way. But one person in particular has never failed to support, encourage, listen patiently, and understand. So it is to my wife, , the *real* forester in the family, that I dedicate this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 - INTRODUCTION

GEOGRAPHIC INFORMATION SYSTEMS IN FOREST MANAGEMENT

Forest resource management is, by its nature, concerned with spatially occurring phenomena. Data collected and used by resource managers are associated with earth locations. Decisions made by forest managers are typically implemented at specific sites. It comes as no surprise, therefore, that foresters have exhibited a great deal of interest in Geographic Information Systems (GIS).

The tremendous investment in GIS systems recently by public and private forestry organizations has helped to create a substantial market for computer mapping hardware, software, and services. Even a cursory review of current literature suggests that nearly all large forestry organizations (those managing perhaps 500,000 acres or more of land) rely upon GIS for spatial data management and analysis. If the present trend continues, it is not difficult to imagine that within the next decade, almost every acre of professionally managed forest land in North America may be represented in digital form in some GIS system.

While most foresters using GIS recognize that their data and results contain error, this acknowledgement is not communicated to users of these analyses unless specific statements are made as to accuracy and precision. Users of forest resource reports and inventory estimates may be quite accustomed to statements about accuracy and precision. These statements may come in the form of confidence intervals or standard deviations, and acknowledge the probabilistic

nature of the figures reported. However, this stochastic treatment of the *attributes* of spatial phenomena has not been extended to a stochastic treatment of the *location* of spatial phenomena. Recent GIS literature and studies have begun to call attention to this serious shortcoming.

Forestry represents only one aspect of GIS use. Concurrent with widespread application of GIS in forestry has been its growth in such fields as urban and regional planning, geologic and hydrologic investigations, automated cartography, land records management, and utility mapping. The forestry perspective on GIS often differs from some of these other applications. For example, forestry applications are characterized by large area coverage, small scale maps, quantification of many-variable phenomena, and an orientation towards analyses and provision of summary statistics, in addition to production of cartographic products. For this reason, many of the studies of GIS errors performed by practitioners in other fields may not be directly applicable in forest management situations.

CONCERNS ABOUT ERRORS IN GIS

Numerous recent articles have called for studies on the sources and effects of GIS errors. A new National Science Foundation consortium on GIS has selected accuracy analysis as its primary initiative for research. Most studies on error reported so far have considered a classification of the sources or types of errors, or have viewed error and accuracy from the perspective of the map producer. Some of the accuracy studies have resulted in qualitative descriptions or models of spatial error, which are not appropriate for quantitative analyses. Thus, while cartographers have been investigating spatial error from a map producer's

orientation, and in a somewhat qualitative manner, foresters are also concerned with quantification of error from a user's point-of-view; e.g., the likelihood that estimated stand acreages are correct.

The primary expression of spatial error that has been used to date is a positional accuracy statement such as that included in the United States National Map Accuracy Standards (Thompson, 1979):

> "For maps on publication scales larger than 1:20,000 not more than 10 percent of the points tested shall be in error by more than 1/30 inch, measured on the publication scale..."

However, such a statement does not provide a map user with any information on the impact of locational errors on resulting area or length estimates, and therefore does not communicate much information relevant to the forestry user's application.

An example of what is missing from GIS accuracy studies is a quantification of the variability of area estimates. Acreage figures are pervasive in the data used in resource management decisions. Yet when acreage figures are printed on a map or in a report, there is rarely any indication that these are estimates. When resource variables such as timber volume are expressed on a per-acre or per-hectare basis, the implication is that these numbers will eventually be multiplied by an estimate of area. Therefore, area errors will have a multiplicative effect in many analyses, and an error statement becomes important.

APPLICATIONS OF A SPATIAL ERROR MODEL

The application of GIS in forestry would be greatly enhanced by an explicit recognition of the errors in digital spatial databases and a more thorough understanding of the impacts of these errors on the resource management decisions which rely upon GIS analyses. First, since forestry GIS databases typically contain information from a variety of sources, scales, and accuracies, it would be helpful to incorporate information regarding the quality of the source data into the database itself. Next, once information is available regarding the variability of source data (or if assumptions can be made about this variability), a technique is needed to translate information about positional variability into information about area or distance variability. For example, a procedure to derive confidence intervals on the area of timber stands (given a map and some statement or assumptions about its locational precision) would be valuable. Such a procedure should be statistically sound, capable of being automated and included in typical GIS analyses, and sufficiently flexible to accommodate different assumptions about source data variability.

A model of spatial error in a GIS would contribute toward a better understanding of the relative magnitude of spatial errors and attribute errors. If estimates of area are found to be more variable than estimates of per-acre values, for example, a different allocation of inventory resources may be in order: more on deriving area estimates and less on per-acre value estimates. On the other hand, if area estimates are found to be less variable than previously assumed, perhaps more confidence in GIS analyses would be justified.

Another use of a spatial error model would be the analysis of error propagation through

the combination of various source maps in map overlay procedures. At present, concerned GIS users have expressed uneasiness about the reliability of overlay products. A quantification of possible ranges of error would be quite useful, and would require an error modeling capability.

STATEMENT OF PURPOSE

The purpose of this study is to develop a procedure for incorporating information or assumptions about the locational variability of data in GIS databases into analyses typically encountered in forestry applications of GIS. This will be done by creating a stochastic model of area and distance errors based upon assumptions about locational accuracy. Next, a method for expressing area and distance variability to a GIS user will be developed; this method will be suitable for automation in the type of GIS systems currently popular in forestry applications. Finally, the use of the model and its interpretation will be demonstrated through example applications typical of forest management decision processes.

Chapter 2 - LITERATURE REVIEW

This study will consider those aspects of errors in GIS that are pertinent to forestry GIS users. Therefore, an overview of the main features of GIS and a discussion of forest management applications is in order.

DEFINITION AND IMPORTANCE OF GIS

Cowen (1988) recently reviewed some definitions of GIS and concluded: "GIS is best defined as a decision support system involving the integration of spatially-referenced data in a problem-solving environment." As noted in the Forestry Handbook (Wenger, 1984) the purpose of Geographic Information Systems "is not so much to draw maps as to provide information on the spatial location of the resources". In the first comprehensive text on GIS, Burrough (1986) lists the five basic sub-systems necessary in GIS:

1) Data input and verification;
2) Data storage and database management;
3) Data output and presentation;
4) Data transformation;
5) Interaction with the user.

While all the above modules are necessary, the one that sets GIS apart from automated drafting systems is the connection between the spatial information and a database management system containing resource attribute data.

Geographic Information Systems have evolved over the recent years into effective means of accomodating exactly the types of spatially-referenced data that forest managers rely upon. Prior to the development of modern GIS systems, resource data were either analyzed apart from their spatial context, or combined through a tedious manual process of drafting and overlaying of translucent maps of the same scale (McHarg, 1971). The advent of computerized map analysis techniques has initiated a revolution in the way spatial analyses of resource data are performed. It is now possible to readily combine diverse maps of different sources and scales, and to perform complex analyses such as proximity analysis, spatial routing, and multiple map overlay (Johnston, 1987). Physical products such as colored thematic maps, perspective view diagrams, and spatially aggregated reports can now be generated in considerably less time than ever before. All these capabilities introduced by GIS have provided the resource manager with analytic powers that will change the way resource management decisions are made and implemented (Shumway, 1986).

The rapid infusion of GIS technology into resource management organizations is evidenced by the investment in such systems. The money spent by public and private forestry organizations on GIS hardware and software alone has helped create a market for GIS which could reach $500 million by 1991 (Lang, 1988). The initial purchase of a GIS system represents only a fraction of an investment in GIS. Creation of the digital spatial databases used in GIS requires an even larger expenditure than does the initial system acquisition. A huge market for GIS database "conversion", or digitizing, services has been spawned by this demand. Employment opportunities for those with GIS skills have abounded in both the companies providing systems and services, and the companies and agencies which are considered the "end users" of GIS technology.

FOREST MANAGEMENT APPLICATIONS OF GIS

Forestry applications of GIS often differ from cultural applications (urban planning, utility mapping) and from the growing land records management applications (which usually refer to a Land Information System - LIS, or a cadastre). Several contrasts between these applications can help identify the relevance of accuracy studies.

First, the scope of forestry operations in which GIS typically have been applied is larger than in many other fields. Forest products companies and public resource management agencies using GIS may be responsible for the management of hundreds of thousands of acres of forested land, often extending over numerous counties and multiple states. In contrast, most other non-forestry applications of GIS are concerned with cities, counties, development areas, or regions of smaller size. In public forest management, the scope of responsibility usually includes multiple resource concerns. For example, while private forestry companies may be interested in primarily timber, and perhaps wildlife resources, the U.S. Forest Service must also consider cultural resources (archeological sites and historic features), water resources (sensitive soils, watersheds, surface and subsurface water quality and quantity), mineral resources, recreation facilities and opportunities, and non-game wildlife species and habitats. This depth of interest creates a need for multiple "layers" of spatial data which cover the same areas on the earth. For example, in a workload analysis for a GIS to be installed at the George Washington National Forest in Virginia, 49 data layers were identified (Tomlinson Associates, 1985). While some non-forestry applications of GIS (such as regional planning) may also require attention to multiple resources, they rarely cover the areal extent that many forest management companies or agencies must deal with.

In keeping with the broad scope of forestry GIS, data are typically recorded at a smaller scale. While some municipalities utilize GIS databases at source scales as large as 1:1200 (Hanson, 1988), the broad scope of forest management GIS usually requires coverage at a much smaller scale. Scale is a critical consideration in error analysis since small scale maps are almost always more generalized than larger scale maps, and since a given error on a small scale map will represent a larger offset on the ground than an error of the same magnitude on a large scale map.

The forest resource, being a biological one, is often more difficult to delineate and measure as precisely as man-made features. For example, boundaries of forest stands or soils units cannot be mapped at the same level of precision at which cadastral mapping of ownership boundaries is performed. As noted by Goodchild and Dubuc (1987), natural resources data differ in character from socio-economic data, which may include primarily lines arbitrarily defined by man (such as administrative and political boundaries). The authors note: "Lines which follow streets are likely to have very different errors from lines which are defined to follow rivers.." Thus, the nature of the features being mapped imparts an important characteristic to mapping errors.

The orientation of forestry GIS is more towards measurement and analysis than towards production of a physical map product. In many GIS systems used by cartographers or planners, the GIS is a means to produce a physical map, which is the end product (Weibel and Buttenfield, 1988). In contrast, in most forestry GIS applications, the map is a means of representing spatial phenomena which are often combined with economic data and operational parameters, and an integrated analysis is the end product (Sieg, 1988). Consequently, a GIS which never produces a physical cartographic product may still be extremely useful to a forest

manager. Goodchild (1980a) notes "The most useful products of a geographical information system ... are measures of some kind, such as the area of a homogeneous patch of land of certain characteristics, the length of a line, or the distance between specified points." The orientation of any accuracy study in forestry will need to focus not so much on the planimetric accuracy of a map product as on the accuracy of the spatial extent (distance and area) of the phenomenon being mapped. Indeed, accurate location of a feature on a forester's map is rarely as critical as accurate expression of area or distance.

Some non-forestry applications of GIS require very high levels of accuracy and precision. For example, some land-records-information systems contain legal, authoritative descriptions of land ownership, and accuracy is critical (Sonnenburg, 1988). Forestry applications of GIS may vary widely in the required accuracy and precision. In many forest management situations, the GIS is used only as a first step in identifying areas for treatment or study. Rarely will such decisions be implemented on the ground based solely upon information from a GIS. A prudent manager would typically precede any costly management activity with field verification of information leading to the decision. For example, when selecting timber stands for fertilization, a query of the stand and soil attributes contained in the GIS might be performed to identify potentially responsive stands. These candidate stands would then be reviewed in the field prior to making a final selection. Thus, when absolute location of a feature is needed for implementation of a decision, it will usually be obtained from an on-site inspection, not from a map. In an article on a decision-support system involving GIS, Covington et. al. (1988) state that their system "is thus an interactive tool that ultimately depends on human judgement and expertise for a final decision". This "preliminary selection" type of GIS application does not require a high level of accuracy and precision. At the other extreme, payment for contracted silvicultural treatments is commonly made on a per-acre basis. Often, GIS is used in

conjunction with aerial photographs to determine acreage treated. In cases in which per-acre treatment costs are high, the financial impact of errors in area estimates can be significant.

DATA STRUCTURES

Because this study will deal directly with the representation of spatial features in a GIS, a discussion of the manner in which features are recorded is appropriate.

A GIS database typically contains spatial features which are identified by unique labels or code numbers, and which are recorded in one of several formats. Information about the attributes of a feature is then contained in a database management system (DBMS) and cross-referenced with the spatial feature by the label or identifier code (Parker, 1988). For example, a timber stand in a GIS may be identified by a stand number, which serves as an index to records in a DBMS which contain the attributes of the stand, such as height, age, site index, density, stocking, volume, etc. (Figure 1). The aggregation of data for all features of a given type in the spatial and attribute database comprises one "layer" of information, e.g. the timber stand layer. Additional layers pertaining to other feature types might be included in the same way to contain information on soils, wildlife, recreation, transportation, economic criteria, etc.

GIS can be divided into two broad categories based upon the data structure, or the way in which spatial features are represented in the computer. These are commonly termed "raster" and "vector" structures. Much has been written about the differences between them, and their relative advantages and disadvantages. Maffini (1987) states: "Both raster and vector data

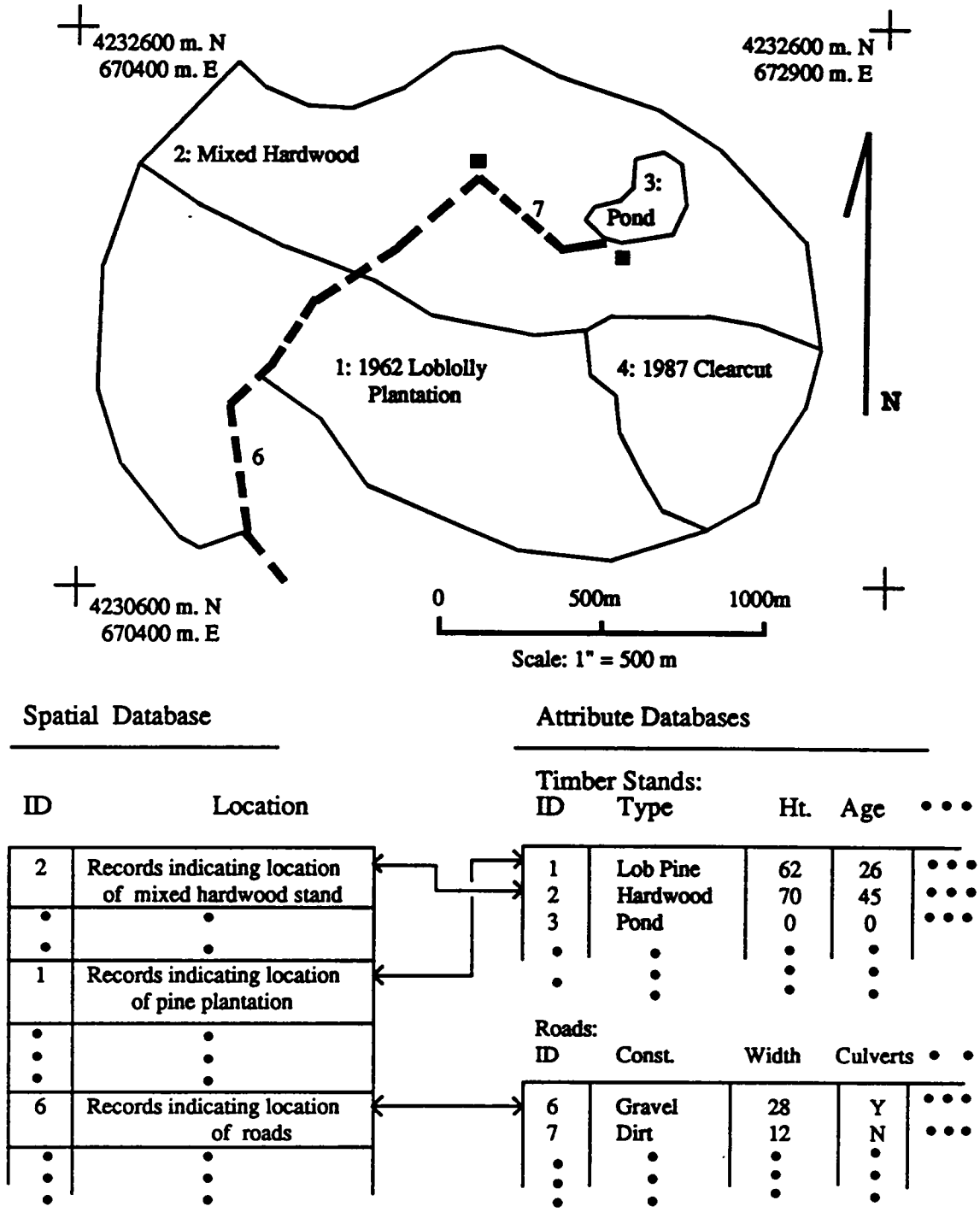


Figure 1. Representation of a map in a GIS with a DBMS linkage.

structures have a place in GIS and will continue to prevail for many more years". Peuker and Chrisman (1975), Monmonier (1982), Peuquet (1984), and Burrough (1986) are excellent sources for discussions of cartographic data structures.

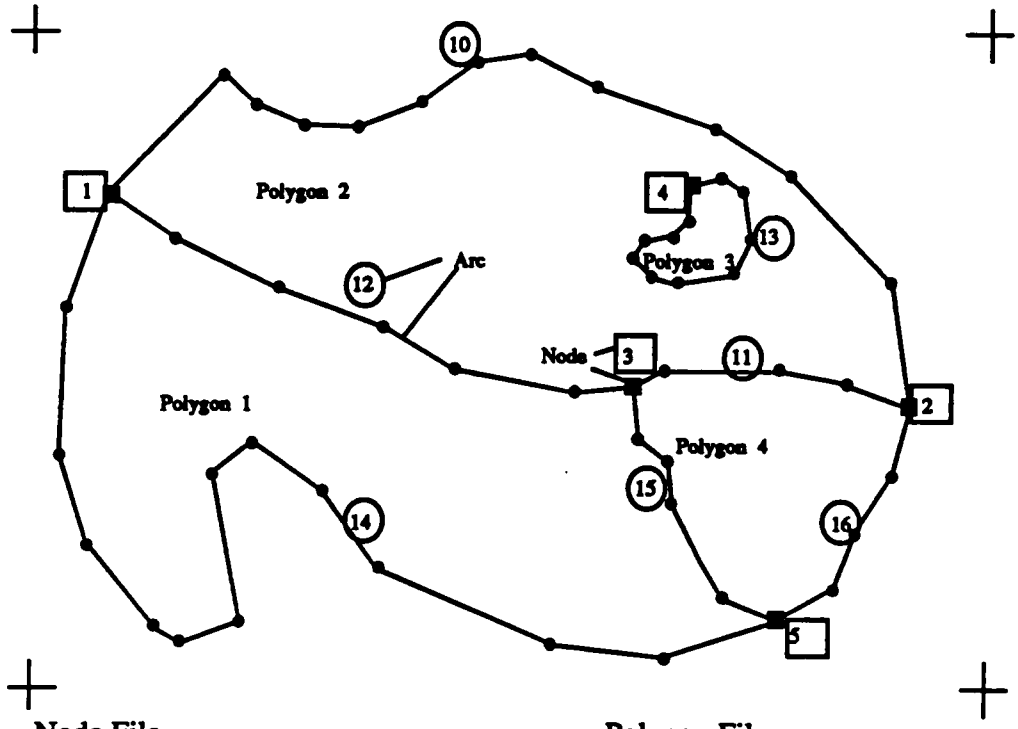
The older of these structures is the raster, or grid cell structure, in which the land surface is represented by a regular tessellation, usually rectangular grid cells, which are registered to earth coordinates. In this structure, ground attributes are recorded for each cell, or pixel, in the grid. The earth location of each item of information is implied by its position in the grid. Thus, a feature such as a timber stand is represented by a collection of adjacent grid cells, each encoded with a label corresponding to the stand. In the digital database of a raster system, a logical record typically consists of a concatenation of the numbers or letters representing cells in a row or column of the grid. The grid structure has been used extensively in conjunction with data derived from satellites, due to the raster format of digital remote sensing products. However, this structure requires a great amount of computer storage space when the resolution of the grid cells is fine. Raster storage is often inefficient for representation of sparse linear networks or for point features. Large grid cells are generally undesirable because of the degree of generalization involved when recording only one feature label per cell. Large cells also produce blocky, less appealing output products unless the scale of the product is small relative to the resolution of the grid. One advantage of the grid cell structure is that it remains the primary technique for computer representation of data that vary continuously across the landscape, such as elevation (Carter, 1988). In addition, almost all computer graphics display devices use a raster structure. For these reasons, the grid cell structure will continue to be used despite its shortcomings.

In the vector format, features on the earth are identified as 0-dimensional (points), 1-

dimensional (lines), or 2-dimensional (polygons). Cartesian coordinates, such as state plane zone or Universal Transverse Mercator (UTM) coordinates are recorded for points. These earth coordinates are derived from mathematical projections of the curved surface of the earth onto planes (see Snyder, 1982). State plane coordinates record the location of a point in feet, relative to an arbitrary origin defined for each zone. UTM coordinates typically represent the location of a point in meters relative to a UTM zone origin. Thus, a point is defined by a northing, or Y-axis distance, and an easting, or X-axis distance, in feet or meters. The X, Y coordinate of a point is therefore the fundamental element of which higher order features (lines and polygons) are composed.

There are at least two subtypes of vector structures. One is called the polygon format, in which lines are encoded as a series of points (sometimes called vertices), and polygons are encoded with a complete list of the points that make up the lines that bound them. When encoding adjacent polygons using this structure, all the boundary points that are shared by polygons are recorded twice: once for each polygon. While this technique allows all the points comprising a polygon to be located close together in physical computer storage (which speeds some processing steps), it involves a redundancy of all points shared by more than one polygon.

Another vector structure is called the arc-node format. This structure defines nodes as the endpoints of arcs, which are strings of points (see Figure 2). As in the polygon structure, points are represented by single coordinate pairs. Lines are created by linking connected arcs. Polygons are then identified by indicating a series of connected arcs which close. This data structure is more complex, involving numerous pointers which cross-reference nodes, arcs, and polygons, yet provides more efficiency in storage. It is this structure which has gained the most popularity in forest management applications in the recent years, and is the structure which



Node File

Node I.D.	Point I.D.	Arcs
1	1	10, 12, 14
2	13	10, 11, 16
3	17	11, 12, 15
4	23	13
5	46	14, 15, 16

Polygon File

Polygon I.D.	Arcs
1	12, 14, 15
2	10, 11, 12, 13
3	13
4	11, 15, 16

Arc File

Arc I.D.	From Node	To Node	Poly Left	Poly Right	Points List
10	1	2	0	2	1, 2, 3, . . . ,13
11	2	3	4	2	13, 14, 15, 16, 17
12	3	1	1	2	17, 18, . . . 21, 22, 1
13	4	4	2	3	23, 24, . . . 32, 33, 23
14	1	5	1	0	1, 34, 35, . . . 45, 46
15	5	3	1	4	46, 47, 48, 49, 50
16	2	5	0	4	13, 51, 52, 53, 46

Figure 2. Representation of polygon features in an arc-node GIS.

will be used in this study, for reasons which will be discussed later.

It is important to note that in neither of these vector structures are actual lines or vectors explicitly recorded. Rather, lines, arcs, and polygons are implied by the adjacency of point coordinates in computer storage. Any error model for vector structures must recognize that these structures really involve only points and that the relationships between them are implied.

Other data structures exist, such as quadtrees (Rosenfield, 1980), generalized balanced ternary structures (van Roessel, 1988), and vaster structures (Peuquet, 1984). However, these have not yet found wide acceptance and will not be considered here.

CALLS FOR ERROR STUDIES

As geographical information systems have developed technologically and expanded in application, more attention has been focused on the reliability of the results coming from these systems. Numerous authors have called for map-makers to begin providing error statements that are more directly interpretable by map users (Bennett, 1977; Aronoff, 1982a; Chrisman, 1984a; Bailey, 1988). Others have noted that more research is warranted in this area. Vitek *et al.* (1984) state "We believe that the next step in the refinement of geographic information systems is specifying the accuracy of the output products". As recently as 1987, Berry (1987) noted that this has not been accomplished: "... effective procedures to spatially characterize map variance and model uncertainty have yet to be developed".

The National Science Foundation recently established a National Center for Geographic Information Analysis. Accuracy analysis has been selected as the first research initiative of this center (Goodchild, 1988). An entire chapter in Burrough's textbook (1986) is devoted to error analysis and GIS accuracy. All of these sources indicate the strong interest in better models and procedures for dealing with the various aspects of errors in GIS.

ERRORS IN GIS - INTRODUCTION

Before discussing the literature on GIS error studies, the use of the term "error" must be placed into context. First, the terms "error" and "variability" are sometimes used interchangeably, and can lead to some confusion. Use of the word *error* may seem to some to imply that a mistake has been made, or that a recorded value is incorrect due to some flaw or defect in the data collection process. On the other hand, ***variability*** is concerned with deviations from a mean, which may or may not be the "true" value of interest. Thus, the term variability is often used in recognition of the fact that a "correct" or "true" value may not exist, or may be unknowable. This is often the case in natural resources, which contain a great deal of randomness, or in which attempts at precise quantification may be clouded by poor definition of terms. For example, the use of site index values to represent the potential height growth of a timber stand does not imply that all trees of a given species growing on a given site will achieve a specified height at a given age. Instead, site index is defined as an *average* height of dominant or codominant trees at a base age (Avery and Burkhart, 1983); this usage accommodates the natural variability of forest ecosystems. Thus, when measured heights of trees do not correspond to a height predicted from site index equations, the difference, or error, does not

necessarily imply that a mistake has been made. Similarly, the map location of a boundary between naturally occurring features should be taken to represent a *tendency* of objects of different types to occur on opposite sides of a possibly non-existent line (Averack and Goodchild, 1984). For example, an experienced forester is not at all surprised to note that hardwood trees will exist in a timber stand designated on a map as "pine".

The point here is that the term "error" will be used in this study to indicate that a there is a difference between features as they exist on the earth and the abstraction or representation of those features as they are portrayed on a map or in a digital data file. Switzer (1975) refers to reality as a "true map" and to any map at hand as an "estimated map". The difference between the "true map" and the "estimated map" constitutes error. Sometimes this difference is due to a flaw or defect, but often will be due simply to the inherent variability of the features being mapped.

Another point about error in maps and spatial databases is that errors are inescapable. The mapping process is one of creating a model of reality, and as such requires generalization (Peuquet, 1984). This generalization step necessarily involves the loss of some information, and results in a difference between the model and reality; it results in error. As Goodchild (1982) points out, the generalization process is one of "error that is deliberate". In addition to requiring generalization, mapping usually involves the representation of features on the curved surface of the earth by a plane, which immediately introduces errors in distances, areas, shapes, and/or relative angles. Thus, spatial errors can be acknowledged to be inevitable, and can be considered to include natural variation.

ERRORS IN GIS - CATEGORIZATIONS

Several recent works have provided classifications of GIS errors. Mead (1982) listed numerous important causal factors in GIS errors, including source map scale, age of data, completeness, degree of modification, and others. His purpose in categorizing such errors was to enable different source materials to be compared for possible inclusion in a database. Walsh *et al.* (1987) categorized errors as "inherent" (dealing with source materials), and "operational" (deriving from the process of being manipulated in a GIS). However, while their categorization pointed out the different stages in which errors may occur, it did not permit classification of individual errors into one of their categories. For example, according to their definitions, a source map error, if propagated through a GIS manipulation, becomes an operational error. Burrough (1986) describes the following three groups of error sources:

I) Obvious sources of error,

II) Errors resulting from natural variation or from original measurements, and

III) Errors arising through processing.

When discussing errors in topographic data, Vitek and Richards (1978) note that both horizontal and vertical errors are possible. However, it may be impossible to distinguish between them. For example, consider a grid representation of an elevation surface, as in Figure 3. The value in each grid cell represents the elevation in feet above mean sea level (MSL) of a given location on the earth's surface. If the true elevation of cell A were known to be 115 feet above MSL, we could say that the cell was in error. However, we could not say reliably whether a vertical error of 5 feet had occurred, or whether a horizontal error had occurred, in which case the cell directly to the east of cell A was positioned improperly. Therefore, in grid data sets

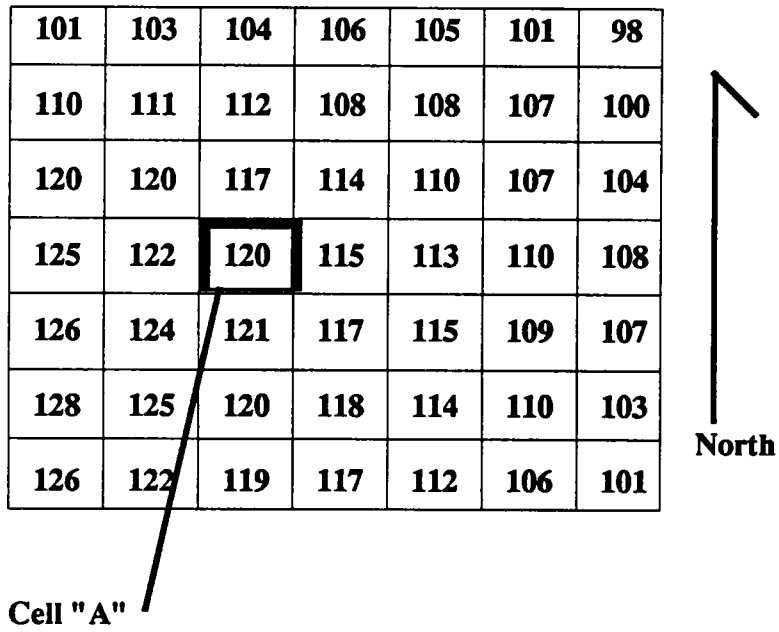


Figure 3. Example of a gridded elevation surface. Cell values represent feet above mean sea level (MSL).

such as this, while hard to categorize, error is relatively easy to diagnose: the value recorded for a cell does not match the "truth" for that location on the ground. This is not the case in vector systems, in which a boundary may be misplaced *and* the attributes of areas adjacent to the boundary may be in error.

The horizontal and vertical errors mentioned by Vitek and Richards correspond to the boundary and classification errors mentioned by Hord and Brooner (1976), and to the positional and attribute errors described by Chrisman (1987a). Chrisman notes that the inaccuracy with which a line is drawn is sometimes due to the indeterminacy of attribute definitions. As an example, he suggests that the problem with drawing wetland boundaries may lie more with the definition of what constitutes wetland than the ability to locate areas on the ground.

The above categorizations of GIS error serve mainly to organize our thinking about the diverse kinds of errors which contribute to the failure of a map to accurately depict reality. By refining the classification of Walsh, *et al.* (1987), it may be useful to distinquish between errors present in source materials, and those introduced in the process of digitizing and manipulating data in a GIS database. A review these types of errors will be helpful in understanding the structure of GIS errors in general.

ERRORS IN GIS - SOURCE DATA

Source data errors may be considered to be either locational or attribute errors (Chrisman, 1987a), acknowledging that it is sometimes impossible to distinguish between them.

Most data in forestry GIS databases come from physical maps. Thus, any errors existing in the source maps will likely remain in the digital data. Attribute data are typically derived from field inventory and entered into a DBMS which is integrated with the GIS. Thus, for the purposes of this review, locational and attribute source errors may be considered separately.

Locational source errors are introduced in the mapping process. The majority of natural resources maps used by foresters originate from aerial photographs. Even USGS topographic quadrangle maps, often considered as ultimately reliable base maps, are derived from photogrammetric processes. Thus, all the elements of the photogrammetric system used in creating source maps may contribute to GIS errors. Camera lens distortions, scale variation due to camera tip or tilt and topographic displacement will all contribute to map errors. Substantial attention has been paid to these sources of error in the photogrammetry literature (Wolf, 1974; Slama, 1980; Smith, 1987; Wiles, 1988). Human factors enter when aerial photographs are visually interpreted for delineation of features, and when the human hand is used to retrace lines. "Fuzzy", or indeterminate, boundaries (e.g., wetland borders) complicate the interpretation process and lend another level of variability. Maps derived from non-photogrammetric sources (field surveys, sketch maps, etc.) will also contain errors from such sources as faulty or misread instruments, failure to adjust for magnetic declination, procedural or arithmetic mistakes, etc. (Breed *et al.*, 1971).

Attribute errors include misclassification of features, errors in measurement of feature attributes (e.g., tree heights), modeling errors when attributes are estimated (e.g., timber volumes), sampling errors, and direct blunders in recording or transcribing data. In addition, generalization errors occur when a single value is recorded for an area of land, and an unwarranted degree of homogeneity is implied (Leung, 1987).

ERRORS IN GIS - PROCESSING

Processing errors may be defined as those errors that result from the entry and manipulation of data in a GIS database. These errors begin with digitizing. Manual digitizing is much like drafting, and includes both physiological and technological factors which contribute to error. Some of these factors have been discussed by Traylor (1979) and modeled by Keefer (1988). Implied in the term "digitizing" is the process of assigning discrete values to continuously occurring phenomena. This computerization of coordinates requires use of numbers with finite precision; thus, some rounding error is inevitable. In addition, when lines are digitized, a finite subset of points is selected from the infinity of points contained in a line. Thus, digitizing is a sampling process and introduces a sampling error. Furthermore, digitizing usually represents a redrafting of a line from some source map and is therefore essentially a generalization of an already generalized line.

An integral part of computerizing spatial data is "registration", or the referencing of points on maps to earth coordinate systems. A variety of techniques are available for registering maps, and with the exception of some work by Petersohn and Vanderohe (1982), little attention has been paid to the influence of the registration process on spatial errors in databases. There is a definite need for further study in the area of registration error, but it is beyond the scope of this study.

Once spatial features have been digitized into a vector system, they are represented by collections of coordinate pairs. Typical GIS algorithms such as proximity calculation, overlay analysis, scale change, coordinate translation, and plotting involve arithmetic manipulation of

these coordinates. Several authors have discussed the impact of finite computer precision on the accuracy of geographic data after such manipulation. Morrison (1980) suggested that computer precision represents the limiting resolution with which maps can accurately be portrayed, but this is refuted by Blakemore (1984) and Burrough (1986). One possible approach to resolving numeric precision problems is simply to allocate more storage for higher precision. Chrisman (1984b) has discussed some aspects of the finite precision of computer coordinates, and suggests "we should push for computer maps that show the graininess and imprecision of our basic information". He argues for judicial selection of a suitable level of precision and notes that using double precision (64-bit floating point) coordinates, "a whole world inventory can be carried to the incredible precision of locating individual viruses". Yet as Franklin (1984) points out, limited precision of coordinates can lead to undesirable results when algebraic manipulations are performed. As an example, he demonstrates a case in which an adjacent point and line are each rotated by a common angle, with the result that the rotated point moves to the opposite side of the rotated line. Thus, computer precision may play an important role in GIS processing errors.

A final type of processing error is the compounding of error in map overlay. This subject has intrigued a number of investigators. One of the earliest treatments of overlay error was by McAlpine and Cook (1971), who first noted the problem of map overlays which resulted in very large numbers of very small polygons which bore little or no agreement with the initial map descriptions. Goodchild (1978) followed the problem of these "spurious" polygons, and noted that the more vertices (points) used to define lines, the more spurious polygons were created. This result implied that the more detail used to draw a line, the more problems it would generate in a map overlay. MacDougall (1975) reported a pessimistic analysis of map overlay accuracy, concluding "... some overlay maps may indeed differ little from random maps

and ... contain more error than the compilers and users probably realize". Bailey (1988) seems to concur, stating "Map overlays may be so inaccurate or unable to capture significant units of productivity and ecological response as to be of questionable value for planning". Chrisman (1987a) argues for a more positive outlook: "While map error should not be ignored in map overlay, the estimates of MacDougall should be replaced by empirically derived test results. Combining information from diverse sources can actually strengthen the value of the information, not degrade it".

The problem of accuracy analysis in the process of multiple map overlay has been discussed from the perspective of a grid-based GIS by Newcomer and Szajgin (1984). They note that the accuracy of a product of map overlay is a function of the spatial coincidence of errors in the component maps. The highest accuracy that can be achieved in an overlay map is only as accurate as the most erroneous input map source. This level of accuracy is achieved when all the errors in all the component maps occur at the same places. The worst case of error occurs when none of the errors in the source maps occurs at the same place; in this case the accuracy of the overlay map is the mathematical product of the percentage accuracies of the component maps. Newcomer and Szajgin's analysis is very useful for grid-based GIS, but does not apply to vector systems.

Burrough (1986) provides an example of the compounding of attribute error in map overlay. Using the universal soil loss equation, which might be a typical application of GIS overlay analysis, he examined the impact of realistic errors in predictor variables on the variability in predicted soil loss and noted "95 percent of the cells having the climate/slope/soil regime specified here would have a soil loss ranging between 3 and 21 cm". The conclusion indicated here was that even moderate errors in single map layers may combine in an overlay

analysis to yield an unacceptable error in the resulting map.

GIS operations such as map overlay may have a drastic impact on the locational and attribute accuracy of maps produced in routine GIS analyses. Error propagation is a serious topic, and one continually being studied. It has been examined in a raster context (Newcomer and Szajgin, 1984), or with an emphasis on "sliver" or "spurious" polygons (Goodchild, 1980a), or with a focus on attribute errors (Burrough, 1986). However, a statistical analysis of the spatial errors resulting from map overlay in a vector system has not been reported. In any analysis of errors in forestry GIS (where overlay is common), consideration must be given to this subject. The closely related topic of detecting spurious polygons also merits further study. When a map overlay produces numerous small "sliver" polygons, the question arises: "Are these significant features on the landscape or are they artifacts of the precision and accuracy of the data and the computer algorithms used?" Some GIS software products (e.g., ESRI's ARC/INFO) provide an option for the analyst to specify a threshold parameter which prevents such spurious polygons from being created. However, published discussion of such capabilities has been concerned primarily with producing aesthetically pleasing map products, not reliable information.

METHODS FOR DEPICTING AND ANALYZING ATTRIBUTE ERRORS

Since the attributes recorded with spatial features are an integral part of GIS systems, some attention must be devoted to attribute errors. However, this is relatively easy, since attributes take the form of numbers and labels which are familiar, and for which methods of portraying uncertainty have been established. Attributes of spatial phenomena, like any

measures, may take the form of nominal, ordinal, interval or ratio values. Numeric variables (including interval and ratio data) may be dealt with in conventional statistical manners. Variability may be expressed as estimates of standard deviation or variance, or as confidence intervals. Derivation of these expressions of uncertainty are well explained in statistical texts and need not be repeated here. Examples in forestry situations are easy to find, and may include: estimates of timber volumes with accompanying variance figures, height measurements "plus or minus" a confidence interval half-width, and ranges on numbers of organisms per unit area.

Nominal data are common in GIS systems in the form of "thematic maps", sometimes referred to as "choropleth maps". These maps contain polygons that are labelled as belonging to categories, such as timber types or soil units. Attribute errors in nominal data take the form of misclassifications. While locational errors may also lead to misclassifications (as in instances of faulty registration), misclassification error will be considered here to be primarily an attribute error. This type of error has been commonly encountered in land cover mapping, and is discussed at length in the remote sensing literature (Campbell, 1987; Congalton *et al.*, 1983; Chrisman, 1980).

The conventional way to express misclassifications in thematic maps has been the contingency table or error matrix. Cross-tabulation of pixels for which both the classification and the reference ("truth") categories are known can provide a user with valuable indications of the suitability of the map for different purposes (Chrisman, 1982a; Prisley and Smith, 1987). Entries along the diagonal of an error matrix indicate the number of correctly classified pixels in various categories. Off-diagonal entries represent misclassifications. A variety of expressions, such as "producer's accuracy" and "consumer's accuracy" (Story and Congalton, 1986) can be

derived from contingency tables.

Aronoff (1982b) proposed another technique to use contingency tables to communicate misclassification rates to map users. His perspective is one of accepting or rejecting maps as if they were hypothesis tests. While this may not commonly occur in practice, it does encourage a statistical consideration of a map as an estimate. Aronoff differentiates between the concerns of the map producer and the map user, defining "producer's risk" (that of incorrectly rejecting an accurate map) and "consumer's risk" (that of accepting a map of insufficient accuracy); these risk classifications are analogous to Type I and Type II errors in hypothesis testing. While the accept/reject decision may not be applicable, Aronoff's suggestions for communicating information regarding map uncertainty to the map user are important. Most applications of contingency tables have been in conjunction with classifications of satellite imagery for land use or land cover mapping or other raster-format data (Greenland and Socher, 1985). There is no reason that contingency tables cannot be produced for vector-based maps; the concept is independent of any underlying structure to the data. However, in practice, reports including contingency tables have almost exclusively dealt with raster-format digital maps.

Attribute errors have been studied by Jenks and Caspall (1971) in the context of choroplethic maps. Their concern was that choroplethic (categorized) maps are a generalization of reality in that continuous spatial distributions are represented by discrete categories. In order to evaluate the error involved in this generalization, they considered several techniques for dividing a continuous variable (per-acre value of farm products in Illinois) into categories. A three-dimensional map of the state of Illinois which portrayed value of farm products by county was produced using each technique, and errors were calculated as the difference in the various representations. Their analysis suggested appropriate methods for separating spatially-

continuous phenomena into discrete categories; however, no attention was paid to errors in location, only in attribute representation.

An analysis of attribute errors by MacEachren (1982) involved a similar treatment of attribute errors. In this study, accuracy of a thematic map was related to the variability of the data being categorized, the size of the enumeration units, and the compactness of these units. Again, spatial location of enumeration unit boundaries were considered to be invariant, and only attribute errors were evaluated.

METHODS FOR DEPICTING AND ANALYZING SPATIAL ERRORS

A number of models have been developed to portray and/or analyze spatial errors in maps. A review of the more prominent and relevant ones is appropriate here.

Raster models

Since the raster data structure has been applied for a longer time than the vector structure, there have been more studies focussing on errors in a raster context. For example, some work by Frolov and Maling (1969) was concerned with the accuracy of area estimates based upon dot grid counts. Their results are directly applicable to raster-based GIS systems, and have been further discussed by Muller (1977), Goodchild (1980a), and Burrough (1986). Several of these authors have been interested in determining the variability of area estimates using grid structures, and agree that this variability is dependent on the resolution (dimension)

of the grid cells. Thompson (1981) related errors in area estimates not only to grid cell size, but also to shape of regions; he employed a shape factor based upon a ratio of perimeter to the square root of area. Switzer and Venetoulias (1987) also note that misclassification rates can be related to grid cell size. Detailed results of these raster studies are not directly applicable to vector data structures; however, some of the underlying concepts (such as the relationship of error to resolution) can be carried into the vector domain.

The Fuzzy Boundary Model

Some recent attempts at expressing the uncertainty of map information have utilized a mathematical framework called fuzzy set theory. Fuzzy set theory acknowledges that some concepts are "fuzzy" or indeterministic by nature, and cannot be dealt with appropriately by common mathematical or statistical techniques. Bouille (1982) suggests that most phenomena occurring in maps and spatial data bases fit this description, and would be suitable for consideration in a fuzzy-set context. In a comprehensive paper on the subject, Leung (1987) suggests "regional boundaries concern the degree of belongingness to regions and are thus fuzzy". His treatment of spatial boundaries using fuzzy set theory involves first a linguistic proposition that defines regions (e.g., the statement "timber stand X is pre-merchantable pine" may be a linguistic proposition which defines a region based upon its predominant timber type and size). The regions are characterized by a set of characteristics (e.g., species, age, stocking, etc.) and a membership function. The membership function indicates the degree to which an area with a set of characteristics corresponds to one of the defined regions. Next, Leung defines regional cores, boundaries, and edges as:

> CORE: "The core of a region Z is the point or area in space whose characteristics are most compatible to the linguistic proposition ...

characterizing Z."

BOUNDARY: "A boundary is in fact a zone including all points in space whose characteristics are more or less compatible to the regional characterization.."

EDGE: "The edge of a region ... consists of points in space which just fall short in having a positive degree of compatibility to the characterization.."

In an example of the above, Leung uses a classification of regions based upon climate. Regions may be classified according to temperature and precipitation, into categories whose labels may include: hot, warm, cool, cold, and abundant, substantial, and adequate precipitation. Rather than using arbitrary thresholds for temperature and rainfall to delineate these regions, a membership function was used to indicate the "degree of belongingness" to a given category. For example, Figure 4 shows the membership for the fuzzy sets "hot", "warm", "cool", and "cold", as a function of temperature. Areas in which the value of the membership function is equal to one are deemed the core of the regions. Areas in which the value of the membership function is above some threshold (Leung uses the notation "alpha level") are considered boundary regions. Edges are then defined as areas in which the membership function for all sets falls below the specified alpha level.

Fuzzy set theory provides an enlightening perspective on boundary imprecision. First, it recognizes that boundaries may be quite indeterminate, and occupy a broad area on the ground. In addition, the membership function shows the indeterminacy of the definitions of regions (Robinson and Frank, 1985). This concept is important in determining how well boundaries can be located. While Leung's application demonstrated the potential utility of mapping with fuzzy

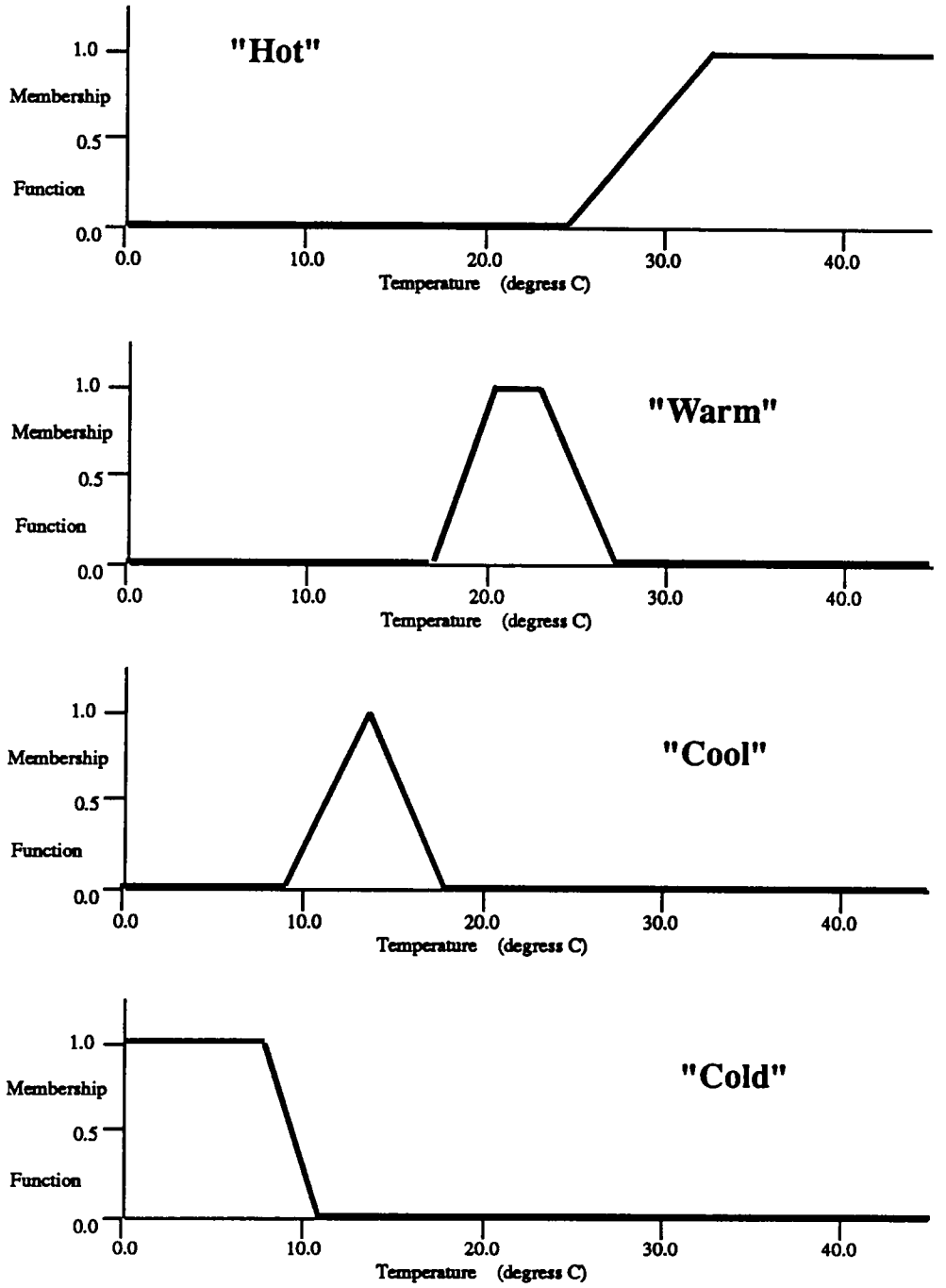


Figure 4. Examples of membership functions for fuzzy sets (from Leung, 1987).

boundaries, it also required extensive data (temperature and precipitation), and involved interpolation, which introduces another dimension of error which was not discussed. The use of fuzzy set theory in the classification of land cover was discussed by Robinson and Strahler (1984), and an application involving relational databases has been presented by Buckles and Petry (1982). Fuzzy set theory has been useful in these situations as a means of portraying the inexactness of the data being used.

The Fractal Model

Some attempts at characterizing the variability of lines have used the concept of fractional dimensionality developed by Mandelbrot (1977). Most of these efforts have focused on the relationship between line complexity or length and scale (Loehle, 1983). The fundamental premise is that the length of a line depends upon the scale at which it is measured. Plotting the logarithm of measured length against the logarithm of the measurement precision indicates a constant relationship. The slope of the plotted relationship is the fractional dimension. Shelberg and Moellering (1983) have developed a computer program to measure the fractional dimensionality of lines. Goodchild (1980b) has shown that Switzer's (1975) estimate of mismatch areas in a raster structure can be rewritten such that the mismatch area is a function of both the grid cell size and the fractal dimension D. However, no practical application has been published which demonstrates the usefulness of such a model to a map user interested in data reliability.

The Epsilon Model and other Error-Band Models

A number of authors have appreciated the fact that boundaries may be more

appropriately represented by a band of a certain width than by an infinitely thin line. Peuker (1976) postulates that a line can be characterized by:

a) a general direction,

b) a band width, and

c) a length.

This definition of a cartographic line has some utility in generalizing lines; that is, reducing the number of coordinate pairs required to adequately define the line. It also may help in determining line intersections, and in determining whether two different representations of a line are actually independent records of the same line. Peuker's discussion does not directly address error, but the concept he advances is useful in characterizing line errors.

Perkal (1966) originated the concept of an "epsilon band" which may be used to represent the probable location of a line. The application he discussed was an attempt to determine the length of a line while considering scale and generalization. Chrisman (1982b) developed the epsilon model into a useful framework for error analysis. He suggested "Given a cartographic line as a straight line approximation, it might be supposed that the true line lies within a constant tolerance, epsilon, of the measured line". In an example application of the epsilon band model, a U.S. Geological Survey Land Use/Land Cover map was noted to have seven percent of the total area contained in epsilon bands of 20 meters. The epsilon model was also used to develop bounds on area measurements for individual land use/land cover categories. Chrisman noted that "These bounds should be interpreted as standard deviations, not absolute limits". His treatment of bounds on area, and the epsilon model itself, was quantitative but not probabilistic; no statement (such as a statistical confidence interval) was provided which would indicate how often the true line could be expected to lie within the epsilon band, or how often the true area would be expected to be contained within the "probable bounds". Another

potential drawback to Chrisman's example is the assumption that the epsilon band is constant in a given map, regardless of the boundary being drawn. A contrary situation can readily be conceived; note that boundaries between similar features (such as a pine-hardwood timber stand and a pure hardwood stand) may be more difficult to delineate accurately than boundaries between quite distinct features (such as between a pine plantation and a water body). Such situations favor the use of different epsilon band widths for different boundary types. Thus, while Chrisman has proposed a reasonable and useful model, greater flexibility in assumptions and a more formal statistical treatment would be desirable enhancements.

Blakemore (1984) applied the epsilon model in an examination of the common "point-in-polygon" analysis. In his study, industrial establishments in England were represented as point data (coordinate pairs), and Employment Office Areas (EOA's) were represented as polygons in a vector database. The analysis was concerned with determining which industrial establishments were located in which EOA's. Using epsilon bands around the EOA boundaries, the uncertainty of the lines were acknowledged. Points (industrial establishments) were designated as being in one of the following categories:

a) possibly out (of an EOA),

b) possibly in,

c) unassignable,

d) ambiguous,

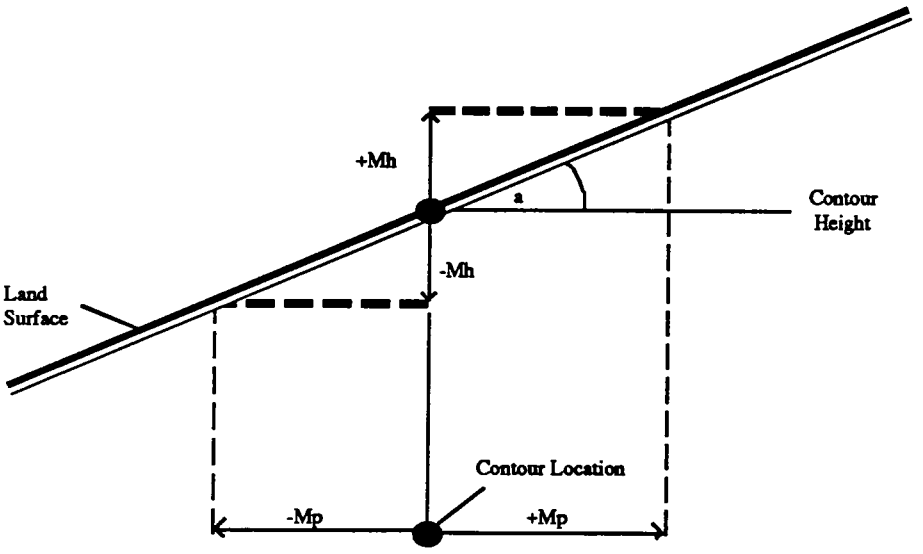
e) definitely in.

With an epsilon band width of 100 meters on the ground, 7% of the 780 test points were in areas of doubt. At an epsilon of 0.7 km, only 50% of the points were uniquely assignable to only one area.

Blakemore's analysis provides an excellent example of how a model of spatial uncertainty can be used in a GIS application to provide information about the reliability of the result. Note, however, that Blakemore's categories were qualitative and not related to specified probabilities. Also, the point locations were taken to be deterministic; variability was only assumed for the EOA boundary lines. Again, a quantitative, probabilistic treatment may enhance the epsilon model.

Another analysis which uses an "error band" model of a cartographic line was presented by Yoeli (1984). This application is interesting in that it relates attribute error to spatial error in a topographic data context. As noted earlier, an error in gridded elevation data is difficult to categorize as being an attribute or locational error. Recognizing this, Yoeli used statements concerning errors in elevation to depict possible locational errors. Figure 5 indicates how this is done. Given an elevation mean square error of M_h, and the slope of a parcel of land, it is possible to project the vertical mean square error to the ground surface, and thence to a planimetric surface. With a given vertical error, steep slopes will produce a smaller planimetric (horizontal) error than will relatively flat slopes. Yoeli produced maps depicting a zone of error around contour lines using this procedure. Unlike epsilon bands, his error bands are not necessarily symmetric about the estimated line, nor are they constant across a map. Rather, they change with the topography. Yoeli's analysis is intriguing, but it is not readily apparent how such a procedure could be applied to non-surficial data.

The above analyses using "error bands" around a cartographic line present the best opportunity for a statistical treatment of spatial error in a vector GIS. Most of the authors cited, especially Chrisman, were very close to the kind of model which has been needed by forestry users of GIS. A more formal statistical treatment can greatly expand the utility of



Mp = Mean square planimetric error

Mh = Mean square height error

a = Angle of slope of land surface

$Mp = Mh / \tan(a)$

Figure 5. The determination of planimetric contour error bands (from Yoeli, 1984).

these models, and strengthen the interpretation of results.

Digitizing Error Models

One component of digital map error that has attracted attention has been that of digitizing error. It is interesting that numerous users of GIS wonder about how closely a person digitizing a line can follow the map line, yet may never question how closely the original map draftsman was able to delineate lines that may have been barely perceptible in the first place. Since the digitizing of maps is a process that closely mimics the original drafting procedure, results from digitizing studies may contribute to an understanding of error processes involved in map creation. Thompson (1981) studied errors in digitizing directly from a photographic stereomodel. He reported an average digitizing error of 3.3 meters when digitizing from 1:80,000 scale photographs onto a map scale of 1:40,000. Chrisman (1982c) assumed that digitizing using scanning devices produced normal errors with a standard deviation of 8.3 meters.

Baugh and Boreham (1976) simulated errors in digital lines representing the coastline of Scotland. They noted that random errors in point coordinates produced a systematic error in length of lines. Keefer (1988) encountered this same result. Baugh and Boreham related this phenomena to what is termed the "Steinhaus paradox": the more accurately an empirical line is measured, the longer it gets. Baugh and Boreham's study set a precedent for using simulation of point errors to evaluate secondary errors, such as errors in line length.

One frequently-cited study of digitizing error was performed by Traylor (1979). Traylor measured and modeled digitizing error in an attempt to train digitizers to improve their

accuracy. Traylor's results helped establish some guidelines for modeling error by noting the presence of systematic patterns and differences between latitudinal and longitudinal errors. Jenks (1981) reviewed Traylor's work and emphasized the human element in digitizing errors.

Chrisman (1987b) examined digitizing error, but was primarily concerned with topological errors, such as missing or multiple labels for a polygon, polygons which did not close or contained "loops", or dangling or unconnected chains. Thus, while he noted that consistency between attributes (polygon or line labels) and the spatial features is critical, he did not examine the digitizing errors relating to correct placement of lines and points.

Otawa (1987) evaluated the impact of digitizing errors on polygon area. Fourteen students independently digitized the same square mile from a soils map, and Otawa reported only generalized results, noting that "The majority fell within plus or minus 7 percent from the mean regardless of polygon size". No tests were performed to detect differences between digitizers, and no attempt was made to relate the variability of polygon areas to either polygon size or complexity.

Keefer (1988) carried the work of many of the above authors to the modeling of spatial errors specifically to support interpretations important to map users. In his study, coordinate errors were simulated and the impact of these errors on line length and polygon area were recorded. Keefer's model included a provision for serial correlation between adjacent points. As noted by Baugh and Boreham (1976), errors in points comprising a line produce a bias in line length. However, no bias was noted in polygon area. Variation in length and area were related to the accuracy standard employed in the simulations, which controlled the magnitude of the

coordinate errors. Standard deviation of polygon area was noted to be related to polygon size, and variation in line length was related to line length. Keefer's work represents an important contribution to the interpretation of the impact of digitizing errors on map variables (length and area) of interest to a forestry GIS user.

Polygon Area Error Models

Two recent works have examined the effects of point location errors on areas of polygons. Bondesson (1986) derived estimates of the standard error of areas obtained by traversing compartments. He assumed normal errors in bearings and distances, and assumed that errors at adjacent points were uncorrelated. Using these assumptions, and standard surveying formulas for computing area, he obtained estimates for area variability. While Bondesson's work concerned points defined in a relative sense (bearings and distances from one point to another), his procedure is very similar to that followed here.

In a more recent effort, Chrisman and Yandell (1988) developed a model to describe the bias and precision of area estimates based upon X and Y coordinates. They assumed independent and identically distributed coordinate errors at well-defined points representing the vertices of a polygon. They reported that the resulting area estimates were unbiased, with variance a function of the coordinates. While their results may be applicable to cadastral data, the reliance on independent errors and well-defined points may limit the usefulness of their model for natural resources data, which are less likely to exhibit the assumed characteristics.

Chapter 3 - PROCEDURE

ASSUMPTIONS REGARDING POINT LOCATION ERRORS

The arc-node data structure described earlier was used as the data model for this work. Since point coordinates are the fundamental feature in the arc-node data structure, it is logical to begin any derivations with assumptions regarding errors in point locations. The notation introduced by Chrisman and Yandell (1988) will be used as a starting point. An X,Y coordinate is considered a multivariate random variable; the observed location of a point (X_i,Y_i) consists of the unknown true coordinate and an error term:

$$X_i = x_i + \epsilon_i \qquad Y_i = y_i + \eta_i \tag{3.1}$$

where: x_i, y_i = true location of point i

ϵ_i = error in location of x-coordinate of point i

η_i = error in location of y-coordinate of point i

The concept of the coordinates of a point being random variables is not new. Statements of map accuracy are typically based upon the recognition that points as represented on a map may be in error relative to their position on the earth. Thus, the coordinates of a point on a map may be considered to be random variables centered on the true location of the earth feature they represent, varying to some degree due to the errors accumulated in the mapping process. We express this concept by considering the observed point location to be an unbiased estimate of the true location:

$$E(X_i) = x_i \qquad E(Y_i) = y_i$$

The next assumption that must be made concerns the variability of the point coordinates. A similar error in both X and Y dimensions is plausible. This can be expressed by assigning an equivalent variance to both errors:

$$\mathrm{Var}(\epsilon_i) = \mathrm{Var}(\eta_i) = \sigma_i^2$$

In some specific cases, there may be reason to reject this assumption. For example, if it is known that the use of a certain map projection has distorted coordinates more in one direction than the other, unequal variances for X and Y errors may be more suitable. However, in developing a model to describe the general expected situation, equal variances do not seem unreasonable.

Next, we will assume that X and Y errors are uncorrelated:

$$\mathrm{E}(\epsilon_i \eta_i) = 0$$

This assumption may also be arguable. For example, Traylor (1979) provided evidence that X and Y digitizing errors *are not* independent, but rather are influenced by the direction that the digitizer cursor was moving when the point was sampled. Traylor noticed a greater tendency for longitudinal than latitudinal errors, which would imply that $\sigma_x \neq \sigma_y$ and that ρ_{xy} is a function of digitizing direction (such that the major axis of an isodensity ellipse would follow the direction in which the line was drafted or digitized). While Traylor's suggestion of dependence between error directions is convincing, his work concentrated on *digitizing* errors, which represent only one component of the overall coordinate error. Undoubtedly, some of the error components will exhibit this type of correlation (particularly those involved with human line-following processes), while others will not. The degree of correlation between X and Y errors at a point will depend upon the relative importance of the error components which exhibit such a dependency. If the model for point error were to include a correlation between X and Y errors, then the direction of digitizing, drafting, scribing, etc. at each point would have to be known or

arbitrarily assumed in order to determine the appropriate values for σ_x, σ_y, and ρ_{xy}. The model would require a different value of ρ_{xy} for every point; these values would presumably derive from the angular direction from the previous point. In order to avoid arbitrary selection of which adjacent point is to be deemed the previous point, a simpler model will be used here and X and Y errors will be assumed to be independent.

Next, we need to consider correlation between errors at adjacent points. Several authors (Traylor, 1979; Jenks, 1981; and Keefer, 1988) have indicated that digitizing produces correlated errors at adjacent points. Keefer *et al.* (1988) reported that an autoregressive process of order 1 was exhibited by the majority of the data they studied. This means that the error at a coordinate can be expressed as a function of the previous coordinate error and an independent random error:

$$\epsilon_i = \rho\epsilon_{i-1} + \epsilon_{i_0}$$

where: ϵ_i = error at point i

ϵ_{i-1} = error at point before point i

ρ = correlation coefficient

ϵ_{i_0} = independent random error

Many of the map-making processes (such as drafting and scribing) may be thought to produce error patterns similar to those from digitizing. In addition, processes such as map projection or coordinate rotation produce errors that are functions of the coordinates themselves, with adjacent coordinates behaving similarly. Thus, the model used here will account for correlation between errors at adjacent points by using a correlation coefficient, ρ. If ρ_i is the correlation between errors at points i and $i+1$,

$$\rho_i = \frac{\mathrm{Cov}(\epsilon_i,\epsilon_{i+1})}{\sigma_i\sigma_{i+1}} = \frac{\mathrm{Cov}(\eta_i,\eta_{i+1})}{\sigma_i\sigma_{i+1}} \quad .$$

Thus, we assume that X errors at adjacent points are correlated to the same extent as the Y errors at those points, but that X errors are not correlated with Y errors.

The assumptions regarding errors in point locations can be summarized with the following expressions of expectation:

$$E(\epsilon_i) = 0 \qquad E(\eta_i) = 0$$

$$E(\epsilon_i\epsilon_{i+1}) = \rho_i\sigma_i\sigma_{i+1} \qquad E(\eta_i\eta_{i+1}) = \rho_i\sigma_i\sigma_{i+1}$$

$$E(\epsilon_i^2) = \sigma_i^2 \qquad E(\eta_i^2) = \sigma_i^2$$

$$E(\epsilon_i\epsilon_k) = 0 \quad \text{for } |k\text{-}i| > 1 \qquad E(\eta_i\eta_k) = 0 \quad \text{for } |k\text{-}i| > 1$$

$$E(\epsilon_i\eta_k) = 0 \quad \forall\ i,k$$

POLYGON AREA ERRORS - DERIVATION

The area of a polygon in a vector GIS is calculated using some form of a standard algorithm which expresses area as a function of cartesian coordinates (Maling, 1989):

$$A_N = \tfrac{1}{2} * \sum_{i=1}^{n} \left(X_iY_{i+1} - X_{i+1}Y_i\right)$$

where: $X_1 = X_{n+1} \quad Y_1 = Y_{n+1}$

A_N = area of a polygon composed of n unique points

The above formulation yields a positive result for area when coordinates are indexed in a counter-clockwise manner; thus, when points are recorded in a clockwise direction, the absolute value is taken. This expression is sometimes referred to as the *herringbone method* (Maling, 1989). An alternative formulation involves centering the coordinates about a local origin. For a

polygon, a reasonable origin is the polygon centroid. If the coordinates of the polygon centroid (X_c,Y_c) are subtracted from all other coordinates, the resulting area is the same. This is equivalent to a coordinate translation from the origin of the coordinate system to the polygon centroid. In computer processing, this has the advantage of increasing the precision with which coordinates are manipulated, and thereby reducing rounding errors. If *centered* coordinates $(\tilde{X}_i,\tilde{Y}_i)$ are defined as:

$$\tilde{X}_i = X_i - X_c \qquad \tilde{Y}_i = Y_i - Y_c$$

$$\tilde{x}_i = x_i - X_c \qquad \tilde{y}_i = y_i - Y_c \tag{3.2}$$

then the alternate expression for polygon area is:

$$A_N = \tfrac{1}{2} * \sum_{i=1}^{n} \left(\tilde{X}_i \tilde{Y}_{i+1} - \tilde{X}_{i+1} \tilde{Y}_i \right) \tag{3.3}$$

which holds for arbitrary X_c, Y_c. As will be noted, an approximate expression for polygon area variance was derived which omitted minor covariance terms, and a dependency of area variance upon the centroid location was noted. Thus, a consistent method for defining the polygon centroid was needed. Rather than arbitrarily selecting one of the variety of centroid definitions in use (Monmonier, 1982), differential calculus was used to obtain the centroid location which *minimizes* polygon area variance. This centroid will be termed the *minimum-variance centroid* (MVC). It is hypothesized that inclusion of omitted covariance terms would obviate the necessity for use of the MVC.

The derivation of mean polygon area and polygon area variance proceeded in steps from the coordinates to the entire polygon. First, it was noted that each pair of points in the polygon boundary formed a triangle with the MVC. The area of one of these triangles is obtained by one element of the sum in (3.3):

$$A_i = \tfrac{1}{2} \left(\tilde{X}_i \tilde{Y}_{i+1} - \tilde{X}_{i+1} \tilde{Y}_{i+1} \right) \tag{3.4}$$

where: A_i = area of the triangle formed by points i, i+1, and the MVC

The sum of the areas of these triangles is then equivalent to the polygon area. Note that (3.4) may yield negative values, depending on the direction in which coordinates are indexed. Some triangles may effectively deduct area from the polygon (Figure 6). However, the overall polygon area will be positive if coordinates are indexed in a counter-clockwise direction.

The first step in deriving the mean and variance of polygon area was to obtain the mean and variance of area of a single triangle. This was done using simple statistical techniques for obtaining the expectation and variance of sums and products of random variables.

Next, it is apparent that adjacent triangles will have correlated areas. This is because each point on the polygon boundary is shared by two triangles, and an error in the point location will affect the area of both triangles. Thus, it was necessary to derive the covariance of area for adjacent triangles. This was also done using the methods for finding the expectation of a function of random variables, since, according to the definition of covariance:

$$\mathrm{Cov}(A_i,A_{i+1}) = \mathrm{E}(A_iA_{i+1}) - \mathrm{E}(A_i)\mathrm{E}(A_{i+1}).$$

Now, the expectation and variance of polygon area is found simply by taking the expectation and variance of the sum in (3.3). The expectation of polygon area is the sum of the expected triangular areas. The variance of polygon area is the sum of the triangle variances plus twice the sum of the triangle covariances:

$$\mathrm{E}(A_N) = \sum_{i=1}^{n}\mathrm{E}(A_i)$$

$$\mathrm{Var}(A_N) = \sum_{i=1}^{n}\mathrm{Var}(A_i) + 2\sum_{i=1}^{n}\mathrm{Cov}(A_i,A_{i+1})$$

where: $A_{n+1} = A_1$

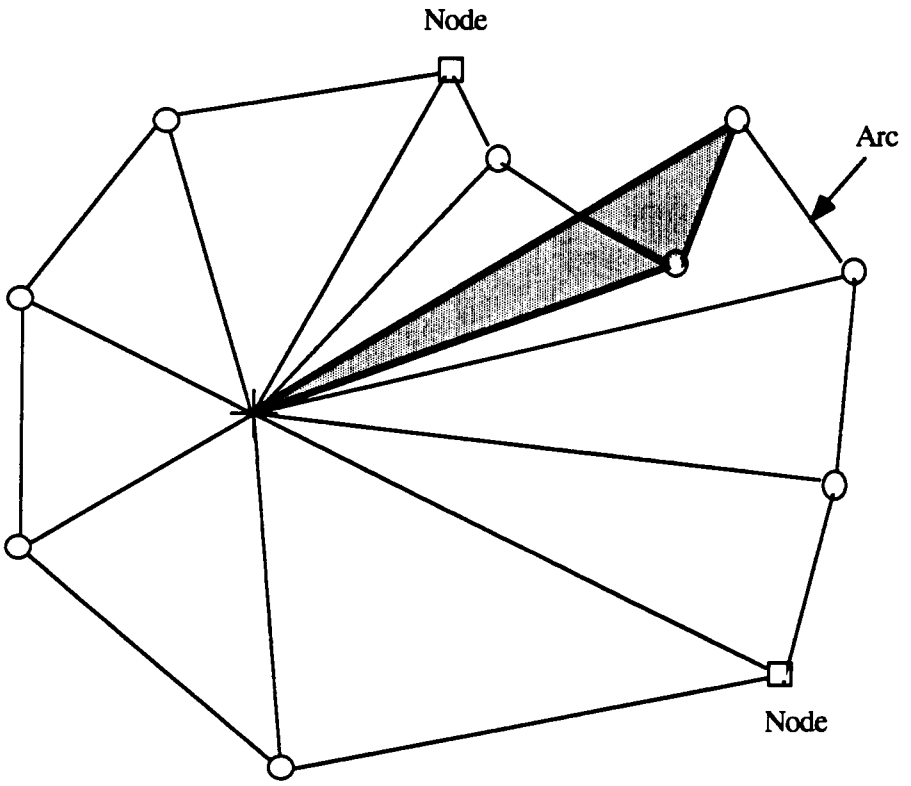


Figure 6. Polygon area as a summation of triangle areas. The area of the shaded triangle is deducted from polygon area.

Note that in the above equation, only pairwise covariances between adjacent triangles $\left(\mathrm{Cov}(A_i,A_{i+1})\right)$ are considered. This represents a first approximation, and omits the covariance between triangle A_i and all non-adjacent triangles within an arc. These additional covariance terms would be functions of order ρ^n, $n>1$. Thus, while not an exact expression, it is believed that this represents a sufficient approximation and enables significant simplification.

After obtaining an expression for polygon area variance, the covariance of area of adjacent polygons was considered. Due to the simultaneous dependency of a pair of polygons upon the arc which they share, there is a negative covariance of areas of adjacent polygons. Any error in an arc between two polygons will increase the area of one polygon while decreasing the area of the other. The covariance of polygon area is important for cartographic modeling applications, an example of which will be discussed later. To obtain the needed covariances, we begin again at the level of the triangles. Now, we note that every pair of points in an arc belongs to *two* triangles: one with the MVC of the polygon to its left, and one with the MVC of the polygon to its right (ignoring for the moment arcs along the exterior map boundary). Thus, every point along the arc is involved in *four* triangles; two to the left and two to the right (Figure 7). This means that triangle A_i on the left of an arc will exhibit covariance of area with triangles A_{i-1}, A_{i+1}, B_{i-1}, B_i, B_{i+1} (where A denotes triangles to the left of the arc and B denotes triangles to the right of the arc). To obtain covariance between adjacent polygons, we now must consider the covariances:

$$\mathrm{Cov}(A_i,B_{i-1}),\ \mathrm{Cov}(A_i,B_i),\ \text{and}\ \mathrm{Cov}(A_i,B_{i+1}).$$

These terms will be derived as before, using common statistical methods for obtaining expectations of sums and products.

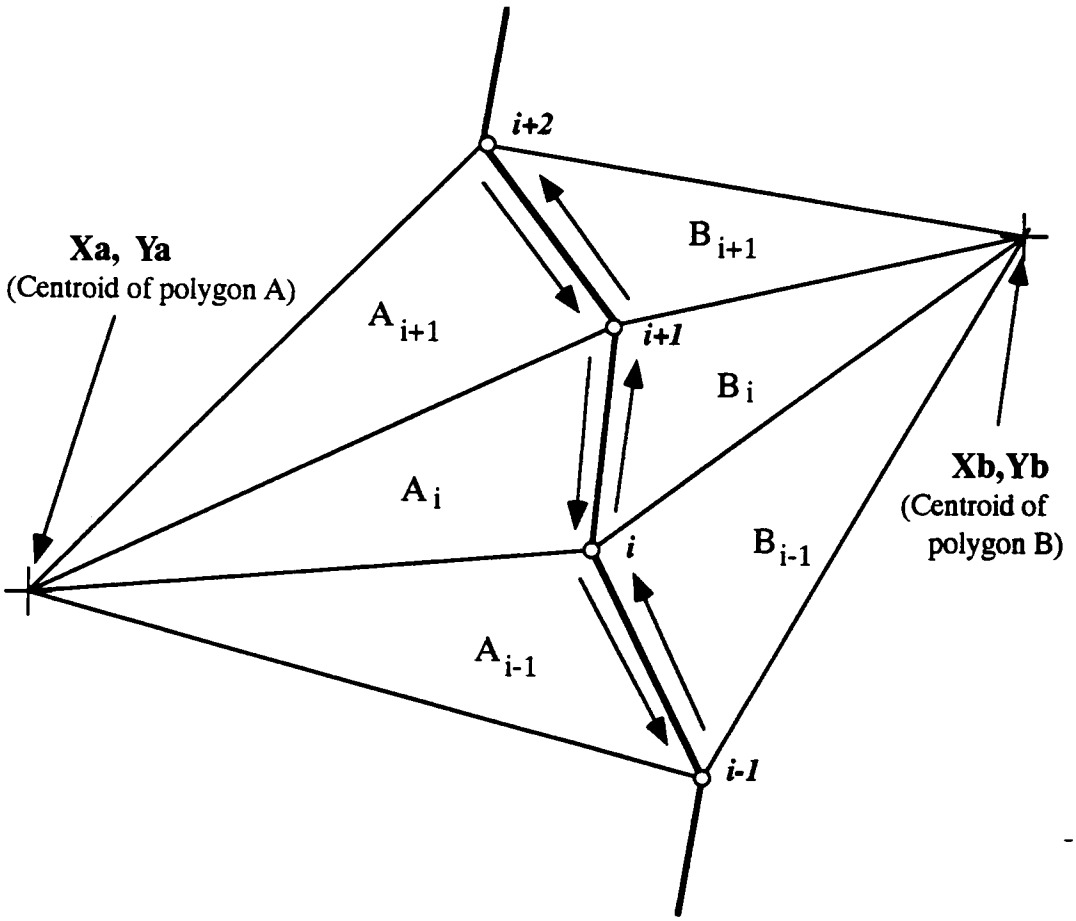


Figure 7. Diagram of the triangular areas involved in polygon covariance.

Once these covariances have been obtained, we combine them in a sum along an arc:

$$\mathrm{Cov}(A,B) = \sum_{i=1}^{m}\Big(\mathrm{Cov}(A_i,B_{i-1}) + \mathrm{Cov}(A_i,B_i) + \mathrm{Cov}(A_i,B_{i+1})\Big)$$

where: A = area of the polygon to the left of the arc

B = area of the polygon to the right of the arc

A_i = area of triangle i in the polygon to the left of the arc

B_i = area of triangle i in the polygon to the right of the arc

m = number of points in the arc

$\mathrm{Cov}(A_1,B_0) = \mathrm{Cov}(A_m,B_{m+1}) = 0$ by definition

(Here we assume the order of indexing of points is that which yields a positive area for the polygon to the left of the arc). Once again, the derivation will allow for a different σ and ρ for each point, while a practical application would likely use a single σ and ρ for the entire arc separating the two polygons.

POLYGON AREA ERRORS - VALIDATION OF THE DISTRIBUTION

Knowledge of the mean and variance of polygon area errors can be more useful if they can be shown to be parameters of some known distribution. Given a distribution for these errors, the probability of occurrence of certain events could then be inferred. For example, "sliver" polygons which arise from overlay and intersection of similar arcs present a problem in interpretation. Do such polygons represent significant features on the ground, or are they artifacts of the map overlay process? Generally, such polygons are small in size. In fact, some

GIS software modules provide for the arbitrary elimination of polygons smaller than some threshold area, on the assumption that they must be insignificant. If the distribution of polygon area were known, a *p*-value could be obtained which would indicate the probability of getting a sliver polygon of the observed size when, in fact, no such feature exists in the area being mapped. Such a statement could be useful in the determination of which sliver polygons to eliminate.

An assumption of normal errors in point location has been suggested by Chrisman (1982a), and seems quite reasonable. This is expressed as:

$$\epsilon_i \sim N(0,\sigma_i^2) \qquad \eta_i \sim N(0,\sigma_i^2)$$

Then, $$X_i \sim N(x_i,\sigma_i^2) \qquad Y_i \sim N(y_i,\sigma_i^2)$$

and $$(X_i\text{-}X_c) \sim N(\tilde{x}_i,\sigma_i^2) \qquad (Y_i\text{-}Y_c) \sim N(\tilde{y}_i,\sigma_i^2)$$

The formula for the area of a triangle (3.4) can be rewritten as a function of a random determinant:

$$A_i = \tfrac{1}{2} * \begin{vmatrix} (X_i - X_c) & (X_{i+1} - X_c) \\ (Y_i - Y_c) & (Y_{i+1} - Y_c) \end{vmatrix} = \tfrac{1}{2} * \begin{vmatrix} \tilde{X}_i & \tilde{X}_{i+1} \\ \tilde{Y}_i & \tilde{Y}_{i+1} \end{vmatrix}$$

If we assume that adjacent coordinates are independent and normally distributed, we can apply the findings of Nicholson (1958), who noted that the distribution of a random normal determinant could be approximated by a normal distribution. Indeed, if ρ=0, the expression for variance of a triangle obtained here agrees with the variance of a 2x2 random normal determinant described by Nicholson (1958). Thus, in the absence of correlation between coordinate errors, the polygon area is a sum of nearly-normal random variables. Because these triangles *are not* independent, the Central Limit Theorem is not strictly applicable. However, it appeared that a normal distribution would still be a reasonable approximation to the distribution of area estimates.

It was thought that as the number of vertices (and therefore triangles) in a polygon increased, the more the distribution of errors would tend towards normality. To test this idea, simulations of errors in polygon coordinates were performed. Six polygons were created by sampling points at regular intervals on circles of different radii such that the polygons all had the same area. The area was arbitrarily set at 8000 square units. Polygons were created with 3, 5, 7, 9, 11, and 15 vertices. (The polygons were therefore an equilateral triangle, a regular pentagon, heptagon, nonagon, etc.). The polygons were created with a single arc which closed on itself. It was noted that when the standard deviation of location errors is large relative to line segment length, pathological situations arise since adjacent points may actually *reverse order* when errors are added to point coordinates. A regular polygon with 15 vertices and an area of 8000 square units will have a boundary composed of 15 line segments which are each 21.29 units in length. The standard deviation for point errors was set at 2 units (roughly 10% of the line segment length) so as to effectively eliminate the potential for simulated errors causing a reversal of points in the polygon boundary. The correlation between adjacent X errors and between adjacent Y errors was arbitrarily set at ρ=0.5. Using the IMSL (1987) Fortran subroutine RNMVN, vectors of n correlated normal errors were created to represent X and Y coordinate errors at each vertex of each polygon such that:

$$\epsilon_i \sim \mathrm{N}(0,4) \qquad \eta_i \sim \mathrm{N}(0,4) \qquad \text{for } i = 1..n$$

$$\mathrm{E}(\epsilon_i \epsilon_{i+1}) = 2 \qquad \mathrm{E}(\eta_i \eta_{i+1}) = 2 \qquad \text{for } i = 1..n\text{-}1$$

One iteration of a simulation on one polygon involved creating the vectors of correlated X and Y errors, adding them to the coordinates, and recording the area of the resulting perturbed polygon. A simulation consisted of 80 iterations for each polygon. (The sample size was set at 80 because initially, the IMSL routine KSONE was used to obtain Kolmogorov-Smirnov test statistics, and KSONE provides exact probabilities only for $n \leq 80$). These 80

polygon areas were then sorted, and the resulting empirical distribution function was compared to the normal (generated by the IMSL program ANORDF). The Anderson-Darling A^2 statistic (Stephens, 1974) was calculated to test the null hypothesis of normality, with the normal parameters of mean and variance as specified by the expressions derived herein. Twenty simulations were performed, resulting in 20 values of A^2 for each of the six polygons.

POLYGON AREA ERRORS - EXAMPLE APPLICATION

An example application of the polygon area variance expression was performed to demonstrate its utility. The application consisted of an analysis of the variability of value of a tract of forested land, due to the variability of area estimates. As noted by Meyer (1963), variability of area estimates is often ignored when total volume (or value) estimates are calculated. The analysis conducted here was meant to demonstrate how area variability can be accounted for in performing such routine tasks as volume or value summaries. The variability of total volume (value) consists of two parts; that due to volume determination, and that due to acreage determination. The former is commonly considered using a standard error estimated from the sample of plot volumes, but will be ignored here in order to concentrate on the latter. However, both components could be considered jointly.

This analysis was performed using data from a recent land acquisition by a southeastern U.S. forest products company. The subject parcel will be referred to as the "Webster" tract. The 218-acre tract consists of five stands, which are comprised of 20 individual polygons (Figure 8). Two of the stands are non-forested. The remainder include a 110-acre stand of upland hardwood, a 53-acre stand of bottomland hardwood, and a 26-acre stand of loblolly pine planted

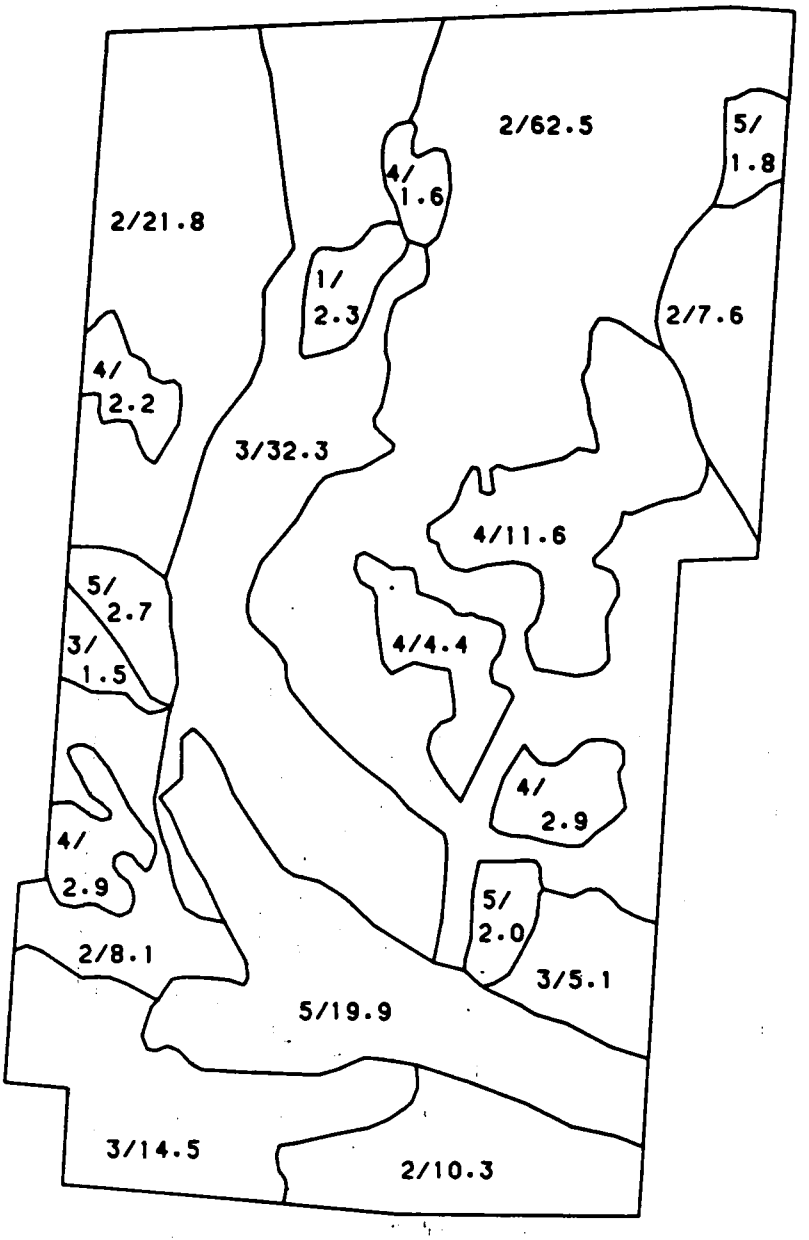


Figure 8. Timber stand map of the Webster tract. Labels inside the stands indicate stand number/acres. Stand descriptions are given in Table 1.

in 1961. These three stands were inventoried using a systematic sample of variable-radius plots (BAF 10) on a 4-chain by 5-chain grid.

After obtaining a survey of the tract boundary, 70mm natural color aerial photographs of the tract were interpreted to delineate stand boundaries. The photo scale was 1:15,840, or 1 inch to 20 chains. The map was digitized, and point coordinates were extracted for input into Fortran programs which calculated polygon area variance and covariance. From the inventory data, pine and hardwood pulpwood and sawtimber volumes were calculated using appropriate tree volume equations. Using current stumpage prices as reported by Timber-Mart South (1989), and volumes per acre from the inventory, a per-acre timber value was calculated for each stand (Table 1). For simplicity, no "bare-land" values were assumed. Thus, the total tract value is obtained from:

$$V = \sum_{i=1}^{n} v_i a_i$$

where:

V = total tract value

v_i = timber value per acre for stand i

a_i = acres in stand i

n = number of stands in the tract

The variance of tract value is:

$$\mathrm{Var}(V) = \sum_{i=1}^{n} v_i^2 * \mathrm{Var}(a_i) + 2* \sum_{i<j} v_i v_j \mathrm{Cov}(a_i a_j) \qquad (3.5)$$

The variance and covariance terms of (3.5) were obtained using the expressions developed herein, under 9 different sets of assumptions about locational errors (three σ's and three ρ's). The assumptions began with three choices for ρ which were expected to encompass

Table 1. Timber volumes and values for the Webster tract.

Stand	Description	Acres	Pine tons/ac.	Hardwood tons/ac.	Value $/ac.
1	Open-Scrub hardwood	2.3	0.0	0.0	0.00
2	Upland hardwood	110.3	5.3	74.3	341.46
3	Bottomland Hardwood	53.4	0.0	193.5	1248.08
4	Loblolly pine (1961)	25.6	87.6	8.3	528.26
5	Open-Abandoned field	26.4	0.0	0.0	0.00
	Total (Average):	218.0	(13.0)	(85.0)	(540.52)

the range of realistic values; these were 0.3, 0.6, and 0.9. A recent survey of the tract boundary indicated a standard deviation of error in monument location of approximately 2 feet. Thus, all arcs on the tract boundary were assigned a standard deviation for points of 2 feet. Even under the best possible conditions, the accuracy of timber stand mapping cannot be expected to be better than the width of the line with which stands are delineated; therefore, the lowest standard deviation used for point location errors was 25 feet (a 0.5mm line at a scale of 1:15840 represents a band 26 feet wide on the ground). A more realistic value was estimated to be about 50 feet, and an extreme value of 75 feet was included. The nine sets of assumptions are indicated in Table 2. Using the estimated values per acre for each stand and variances and covariances of polygon areas, variance of total tract value was obtained for each of the nine sets of assumptions.

DISTANCE ERRORS - INTRODUCTION & ASSUMPTIONS

Distances are calculated in GIS systems in two common forms: distance from one point to another, and distance from a point to a line (arc). As will be seen, the former can be treated as a special case of the latter; the distance from a point to a "degenerate" arc of zero length, consisting only of a single node. Thus, this section will be concerned with obtaining expressions to describe the statistical behavior of the distance from a point to an arc when errors in point location (of both the subject point and the points in the arc) are present.

First, the concept of distance from a point to a line must be clarified. There are an infinite number of such distances; the one treated here is the *minimum* distance from the subject point to some location along the arc. In vector data structures, there are two cases to consider.

Table 2. Nine sets of assumptions about point location errors for estimating polygon area variances and covariances for the Webster tract.

Assumption	σ for boundary arcs	σ for internal arcs	ρ
A	2.0	25.0	0.3
B	2.0	25.0	0.6
C	2.0	25.0	0.9
D	2.0	50.0	0.3
E	2.0	50.0	0.6
F	2.0	50.0	0.9
G	2.0	75.0	0.3
H	2.0	75.0	0.6
I	2.0	75.0	0.9

First, the smallest distance may be from the subject point to a *vertex* or *node* of the arc (Figure 9a). We may call this the *vertex distance.* This is quite distinct from the other case, in which the shortest distance is from the subject point perpendicular to some line segment in the arc (Figure 9b). We will call this the *perpendicular distance.* These cases require individual attention. As mentioned, Case 1 includes the situation of distance from one isolated point to another.

For both cases, we will adopt the same assumptions regarding point location errors as were used when considering polygon area errors. That is, X and Y errors are independent, with the same σ; X errors at adjacent points are correlated, as are Y errors at adjacent points; and the mean errors are zero. In addition, we have the subject point (X_s, Y_s) which is similarly represented as composed of the true location and an error:

$$X_s = x_s + \epsilon_s \qquad Y_s = y_s + \eta_s$$

where: $E(\epsilon_s) = 0 \qquad E(\eta_s) = 0$

$$E(\epsilon_s^2) = \sigma_s^2 \qquad E(\eta_s^2) = \sigma_s^2$$

$$E(\epsilon_s \eta_s) = 0$$

and the errors at the subject point are independent of any errors along the arc.

DISTANCE ERRORS - DERIVATION

Case 1: Vertex Distance

Let (X_v,Y_v) denote the coordinates of the vertex found to be the closest to the subject point, and ϵ_v and η_v the X and Y errors at that point (each with standard deviation σ_v). Then,

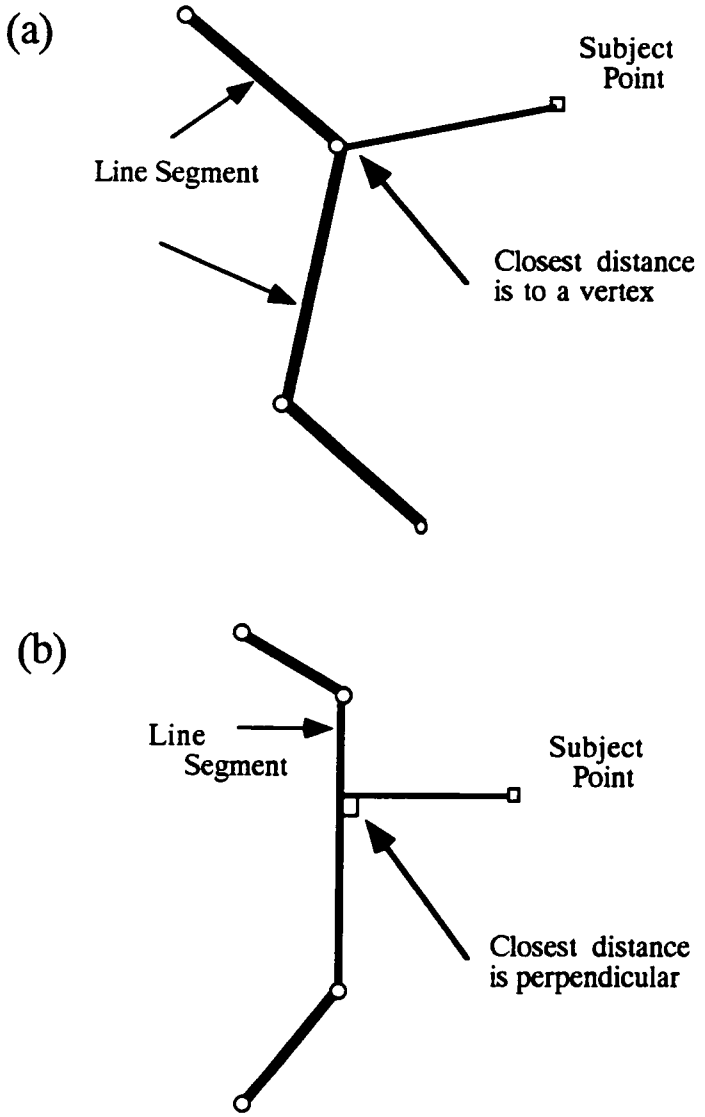


Figure 9. Two cases of distance from a point to a line. Case 1 (a) shows the closest distance being from the subject point to a vertex on the line. In Case 2 (b), the closest distance is along a perpendicular to the line segment from the subject point.

if D is the distance from (X_s,Y_s) to (X_v,Y_v), we have:

$$D = \sqrt{(X_s - X_v)^2 + (Y_s - Y_v)^2}$$

It becomes necessary at this point to adopt an assumption regarding the distribution of errors. The location of a point on a map coming from a GIS is the result of a number of steps, such as photointerpretation, transfer onto a map base, possible multiple plotting and redrafting of that map, digitizing of the map, storage and manipulation of computerized coordinates, and a final re-plotting of the map from the GIS. As discussed previously, each of these steps may contribute to the final error in the point location. Because of the Central Limit Theorem, an argument can be made that the distribution of the composite errors is normal. This argument has precedent: the normal distribution was chosen by Chrisman (1982c) for modeling point errors for the reason cited above. In addition, the symmetry and shape of the normal distribution seem appropriate as a projection of the density of point locations onto a set of coordinate axes. Thus, it might reasonably be expected that point errors will behave like bivariate normal random variables. We can express this as:

$$\epsilon_s \sim N(0,\sigma_s^2) \qquad \eta_s \sim N(0,\sigma_s^2)$$

$$\epsilon_v \sim N(0,\sigma_v^2) \qquad \eta_v \sim N(0,\sigma_v^2) \tag{3.6}$$

These assumptions allow the exact distribution function of D^2 to be obtained. While it would be desirable to know the distribution (or simply the mean and variance) of D, the derivation is intractable. However, in many applications, knowledge of the distribution of D^2 will provide most of the information needed.

Case 2: Perpendicular Distance

The distance from a point to a line segment presents a more complex situation. While the concept of a *point* varying randomly in two-dimensional space is relatively commonplace,

the case of a *line segment* whose endpoints are bivariate random variables was not discussed in any of the literature reviewed. Certain treatments of random lines in two-dimensional space have come close to modeling the situations described here (Solomon, 1978), but have not provided results which are useful for this application.

If we were to attempt to derive the expectation, variance, and distribution of distance as before, we would note that the perpendicular distance is a function of the three coordinate pairs: (X_s,Y_s), (X_i,Y_i), and (X_{i+1},Y_{i+1}). A reasonably simple expression for this function derives from noting that the distance we seek is the height (h) of a triangle whose base (b) is the length of the line segment, and whose area is similar to that defined in (3.4). Noting that $A_i = \frac{1}{2}bh$, we have:

$$h = D = \frac{2A_i}{b} = \frac{\left|(X_i - X_s)(Y_{i+1} - Y_s) - (X_{i+1} - X_s)(Y_i - Y_s)\right|}{\sqrt{(X_{i+1} - X_i)^2 + (Y_{i+1} - Y_i)^2}} \tag{3.7}$$

or:

$$D^2 = \frac{(2A_i)^2}{b^2} = \frac{\Big((X_i - X_s)(Y_{i+1} - Y_s) - (X_{i+1} - X_s)(Y_i - Y_s)\Big)^2}{(X_{i+1} - X_i)^2 + (Y_{i+1} - Y_i)^2}$$

While the numerator of (3.7) can be shown to be approximately normally distributed, we encounter a difficulty which will be discussed later in dealing with the distribution of distance in the denominator. Therefore, it was decided to find a reasonable approximation for the perpendicular distance using other methods.

Consider a line segment and a point with the properties assumed in the previous section. For convenience in the following discussion, we will redefine the coordinates of the point and line segment. Without loss of generality, we can establish a new coordinate system such that the new X-axis coincides with the true line segment. If coordinates in the new system are referred to

as (X', Y'), then let the endpoints of the line segment be at $(x'_1, 0)$, $(x'_2, 0)$, and the subject point be at (x'_s, y'_s) (Figure 10). Let us denote the point of intersection of the perpendicular with the line segment as (x'_p, y'_p). In the absence of any errors in the line segment or the subject point, $X'_p=x'_p=x'_s$ and $Y'_p=y'_p=0$. The errors in the new coordinate system will have the same means, variances, and correlations as before.

One useful expression which is more soluble than a direct derivation of the distribution of D in (3.7) is the distribution of distance *conditioned* on x'_s. By conditioning on x'_s, we will ignore ϵ_p and assume $X'_p=x'_s$ and we return to a univariate case of distance between (x'_s,Y'_p) and (x'_s,Y'_s); and we have $D|x'_s = Y'_s - Y'_p$.

In order to proceed further, we must consider the error at Y'_p (η'_p). Initially, it might appear reasonable that η'_p would be distributed identically to η'_1 and η'_2; this would create an isodensity region around a line segment which would appear as in Figure 11a. In the common epsilon-band models of line error, a constant width for the epsilon band implies exactly this. However, it was believed that the errors in Y' at locations along the line segment between points 1 and 2 would have a *lower* variance than the errors at the line segment endpoints. The reasoning for this is as follows. In order for a Y' error of a given magnitude (say, k) to be observed at some point along the line segment, one of two events must occur. Either *both* endpoints exhibit a Y' error at least as great as k $\left(\{\eta'_1 \geq k\} \cap \{\eta'_2 \geq k\}\right)$, or one endpoint has a smaller error $\left(\{\eta'_1 = u;\ u < k\}\right)$ and the other endpoint has a sufficiently larger error $\left(\{\eta'_2 > k + m(X'_2 - X'_p)$; where m is the slope of the line from one endpoint to the other$\}\right)$. Thus, the probability of an error of magnitude k at an intermediate point is a *joint* probability involving errors at both endpoints. Consequently, $\text{Prob}\{\eta'_p>k\} < \text{Prob}\{\eta'_1>k\}$ and $\text{Prob}\{\eta'_p>k\} < \text{Prob}\{\eta'_2>k\}$. This, in turn, implies that $\text{Var}(\eta'_p) < \text{Var}(\eta'_1)$ and $\text{Var}(\eta'_p) < \text{Var}(\eta'_2)$.

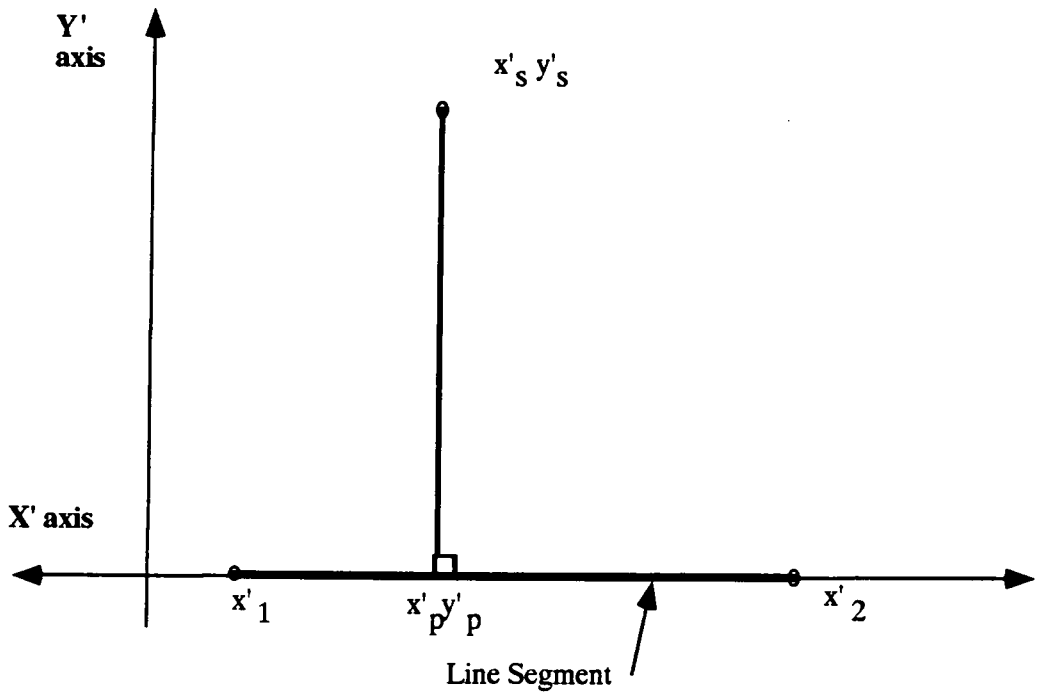
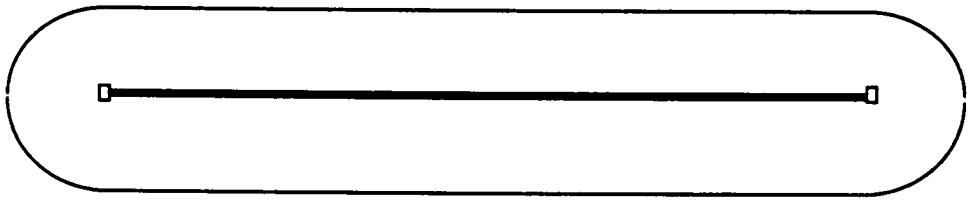
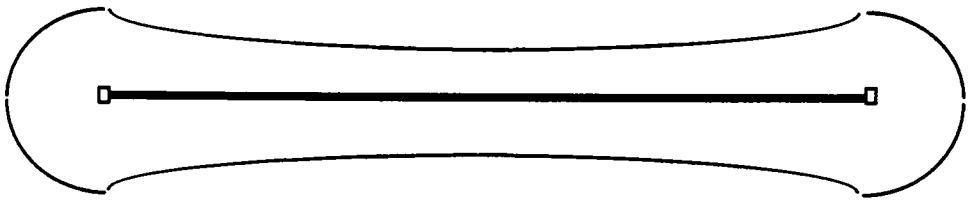


Figure 10. Diagram of a point and a line segment in X′, Y′ coordinates.



(a) Isodensity region is parallel about the line.



(b) Isodensity region is concave.

Figure 11. Isodensity regions indicating the "probable location" of a line segment. The first diagram (a) shows a region for which σ_p^2 is the same for all X'_p (comparable to the epsilon-band model). Diagram (b) shows a modified region, in which σ_p^2 is a function of X'_p.

Thus, it was hypothesized that the isodensity contours indicating the probable locations of a line segment with variable endpoints would be a figure which was circular at the endpoints but concave between (Figure 11b). Noting that at the endpoints, the distribution of Y′ errors was normal, it was further assumed that along the line segment, the *conditional* distribution of Y′ errors (given x′) was also normal:

$$\eta'_p|x'_p \sim N(0,\sigma_p^2) \qquad \text{where } \sigma_p^2 = f(\sigma_1^2,\sigma_2^2,\rho,X'_p) \tag{3.8}$$

The rationale for this assumption comes from simple geometry. The slope of the line segment from (X'_1,Y'_1) to (X'_2,Y'_2) is a ratio of two normal random variables:

$$m = \frac{(Y'_2 - Y'_1)}{(X'_2 - X'_1)}$$

where: $(Y'_2 - Y'_1) \sim N(y'_2\text{-}y'_1,\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$

$(X'_2 - X'_1) \sim N(x'_2\text{-}x'_1,\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$

To obtain the Y′ value at x'_p, we insert x'_p into the equation for the line segment:

$$Y'_p = m(x'_p - X'_1) + b$$

where $Y'_p = \eta'_p$ (since $y'_p = 0$)

b = the Y′-coordinate of the line segment at $X' = X'_1$

$= \eta'_1$

Thus, given x'_p, $(x'_p - X'_1)$ is normal, m is a ratio of normals, and b is normal. This suggests that Y'_p may be normal also.

The next step, therefore, was to obtain the expression which describes σ_p^2 in terms of the endpoint error variances, the correlation coefficient, and the location along the line segment. This function was first evaluated graphically. A program was written to simulate line segments

with normal errors at endpoints (with specified σ and ρ). For each line segment simulated, the errors in Y′ at different points along the line segment were calculated. The variance of these errors was plotted against the location along the line at which they were observed (Figure 12). By simulating different values of σ_1, σ_2, and ρ, the effects of these variables on the function were noted. Several expressions (quadratic and trigonometric) which were thought to provide reasonable estimates of the observed function were tested. The curves of σ_p^2 versus X'_p obtained through simulation were compared graphically with prospective estimates of the function until a "best fit" was found. It was not felt that thorough examination of goodness-of-fit (through regression analysis, for example) was called for, as the intention here was to develop a reasonable approximation, not a definitive expression.

Now, given the conditional distribution hypothesized in (3.8) and the assumption of a bivariate normal error at (X'_s,Y'_s), we note that:

$$D|X'_s = Y'_s - Y'_p$$

$$E(D|X'_s) = E(Y'_s) - E(Y'_p) = y'_s - 0 = y'_s$$

$$Var(D|X'_s) = Var(Y'_s) + Var(Y'_p) - 2Cov(Y'_s,Y'_p) = \sigma_s^2 + \sigma_p^2$$

so: $$D|X'_s \sim N(y'_s,\sigma_s^2 + \sigma_p^2) \tag{3.9}$$

Thus, we have hypothesized that distance from a point to a line segment may be modeled by a normal distribution with mean equal to the nominal distance and variance a function of the individual point variances, the location of the intersection of the line segment and the perpendicular, and the correlation coefficient.

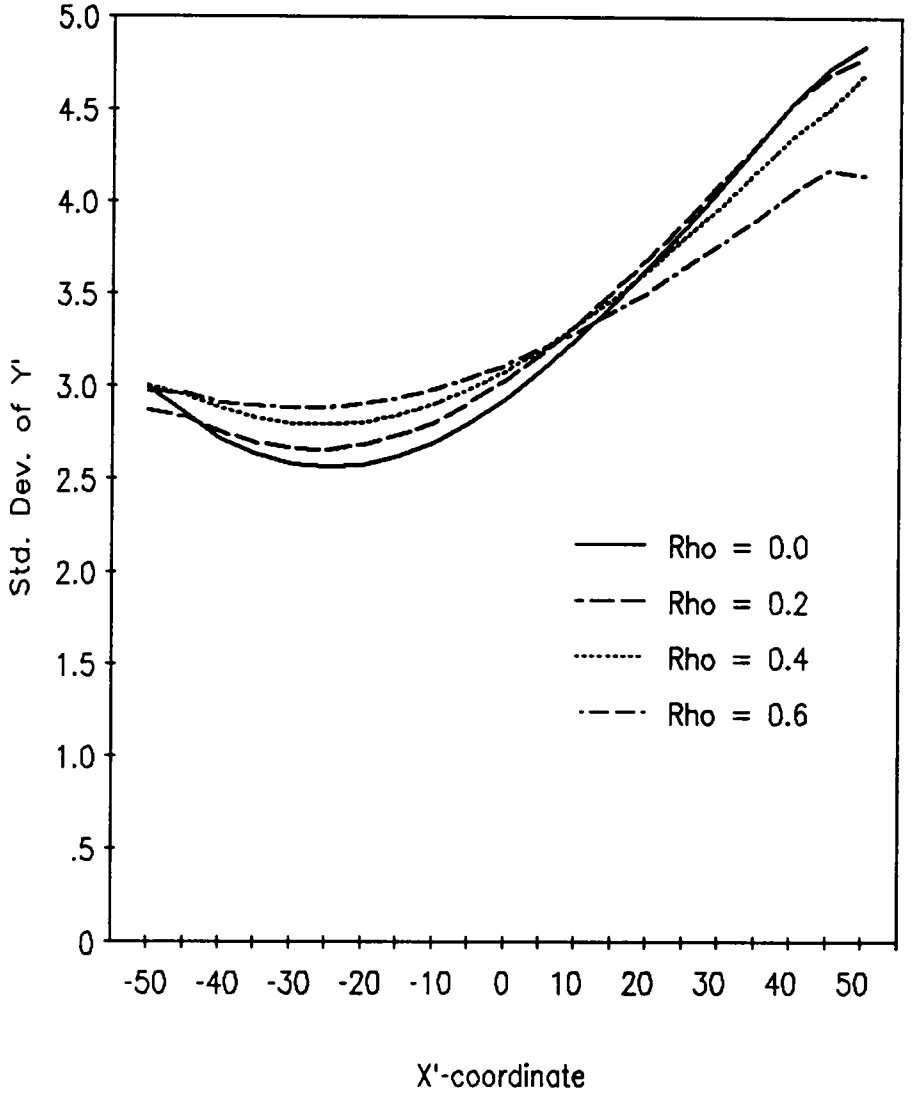


Figure 12. Example graph of σ_p versus X'_p. The curves represent the standard deviation of Y'_p versus X'_p for ρ=0.0, 0.2, 0.4, and 0.6. The data are from 1200 simulations of line segments with $\sigma_1 = 3.0$, and σ_2=5.0.

DISTANCE ERRORS - VALIDATION OF THE DISTRIBUTION

Since case 1 resulted in an exact solution for the distribution of D^2, no validation was necessary. Case 2, however, involved an approximation, and warranted some testing to determine the validity of the approximation.

To conduct a thorough test of the distribution hypothesized in (3.9), it would be necessary to consider ranges of values for X_1', X_2', σ_1^2, σ_2^2, ρ, X_s', Y_s', and σ_s^2. The purpose of this phase of the study was not to *prove* that the distribution of distance is normal, but rather to examine whether a normal distribution is a reasonable model for distance errors in a few limited situations. Of special interest are those cases in which a point is located close to a line segment (close in the sense of a short distance relative to the length and the variability of the line segment). Such cases might include sliver polygons (in which the sliver is triangular, and the distance from the longest side to the opposite vertex is near zero), and situations in which the location of a point with respect to a boundary is in question. Thus, for the purposes of this study, a limited test involving only a few values of σ_2^2, ρ, X_s' and Y_s' (with X_1', X_2', σ_1^2, and σ_s^2 fixed) was performed. The results from a restricted set of assumptions may be generalized to numerous other cases by scaling and by symmetry.

The line segment simulation program used for investigating σ_p^2 was extended to calculate distances between a specified point and line segment with specified error variances and correlations. The line segment was fixed at:

$$(X_1',Y_1') = (-50,0) \qquad (X_2',Y_2') = (50,0)$$

and σ_1^2 and σ_s^2 were fixed at 3.0. Fifty-four combinations of X_s', Y_s', σ_2^2, and ρ (3×3×3×2) were

evaluated:

$X'_s = \{-25.0, 0.0, 25.0\}$

$Y'_s = \{2.0, 8.0, 32.0\}$

$\sigma_2^2 = \{1.0, 3.0, 5.0\}$

$\rho = \{0.3, 0.7\}$

For each combination, 20 simulations of 80 iterations each were performed. Each iteration consisted of generating errors (ϵ_1, η_1, ϵ_2, η_2, ϵ_s, η_s) from the specified normal distributions using the IMSL (1987) subroutines RNMVN and RNNOA. Errors were added to the specified coordinates, and the distance from the subject point to the line segment was calculated. After 80 such iterations, the vector of distances was compared to the hypothesized distribution (3.9) using the Anderson-Darling statistic (A^2). The result (acceptance or rejection of the hypothesized distribution) was recorded. Twenty such simulations were performed for each of the 54 combinations of variables.

DISTANCE ERRORS - EXAMPLE APPLICATION

The most obvious use of a probablistic expression of distance from a point to a line is in point-in-polygon analysis, such as that conducted by Blakemore (1984). An example of possible interest in forest management is the determination of the probability that an inventory plot lies in a specified timber stand. At least one southeastern forest products company is considering digitizing plot locations in their GIS system in order to maintain a spatial identifier for the inventory information associated with a plot. As stand lines change through silvicultural manipulations and map updates, new polygons are created and the question will arise: "Which stand is this plot in?" One approach would consist of simply overlaying the plot location with

the timber stand data, resulting in a deterministic identification of the polygon containing the point. However, when points are near polygon boundaries, and both the point locations and the boundaries contain errors, the prudent analyst would recognize the possibility that the digitized point is not properly located with respect to the digitized stand line. A statement of probability would be useful.

To demonstrate this capability, an analysis was performed using the Webster tract described earlier. The purpose of the analysis was to determine the number of plots which could be located with at least 80% certainty in a stand. Plot locations were simulated according to the protocol described below. To perform the analysis, the closest stand line was determined for each plot, and the distance d from the plot to the stand line was calculated. A p-value was then calculated according to the situation; for vertex distances the value was $P(D^2 \geq d^2 | d=0)$, for perpendicular distances the value was $P(D \geq d | d=0)$. These values represented the probability of observing a distance at least as great as d, if in fact, the point was on the stand boundary (d=0). Low probabilities ($p<0.20$) indicated that the point was not likely to be on the boundary, and therefore could be assumed to be located inside a polygon with some reliability. The number of plots with p-values over 20% represented the number of plots whose polygon membership was not reliably defined; these plots are termed "ambiguous".

A grid of hypothetical plot locations at an exact 4-chain by 5-chain spacing was generated by a computer algorithm and overlaid on the Webster tract stands at a random orientation. The assumptions regarding point location errors were based on a subset of the nine sets of assumptions used previously (B, E, and H in Table 2), and two assumptions regarding variability of plot locations. When plot locations are established on a map or in a digital file *prior* to their visitation in the field (as is often the case), the variability of location no longer

depends upon mapping processes, but upon the process of locating mapped positions in the field. The best judgement of the most experienced inventory forester in the company providing the example data suggested that 90% of the time, he was able to locate plots on the ground within 1.5 chains of their mapped position. This translates roughly into a standard deviation of error in plot location of 60 feet. Thus, the two assumptions used to "bracket" this estimate were σ_s = 50 feet and σ_s = 70 feet. For each set of assumptions, the number of plots of questionable polygon membership were recorded.

Chapter 4 - RESULTS & DISCUSSION

POLYGON AREA ERRORS - DERIVATION

Derivation of Polygon Area Mean and Variance

The first step in obtaining the mean and variance of polygon area is to obtain the expectation of the area of a triangle. Recall from (3.4) that the area of the triangle formed by points X_i,Y_i and X_{i+1}, Y_{i+1}, and the MVC is given by:

$$A_i = \tfrac{1}{2} * (\tilde{X}_i\tilde{Y}_{i+1} - \tilde{X}_{i+1}\tilde{Y}_i)$$

Using the equalities in (3.1) and (3.2), this can be written as:

$$A_i = \tfrac{1}{2} * \Big((\tilde{x}_i + \epsilon_i)(\tilde{y}_{i+1} + \eta_{i+1}) - (\tilde{x}_{i+1} + \epsilon_{i+1})(\tilde{y}_i + \eta_i)\Big)$$

$$A_i = \tfrac{1}{2} * \Big((\tilde{x}_i\tilde{y}_{i+1} - \tilde{x}_{i+1}\tilde{y}_i) + (\tilde{x}_i\eta_{i+1} + \tilde{y}_{i+1}\epsilon_i - \tilde{x}_{i+1}\eta_i - \tilde{y}_i\epsilon_{i+1}) + (\epsilon_i\eta_{i+1} - \epsilon_{i+1}\eta_i)\Big) \qquad (4.1)$$

Now, note that the nominal area of the triangle (assuming no errors) is equal to the first two terms in (4.1):

$$a_i = \tfrac{1}{2} * (\tilde{x}_i\tilde{y}_{i+1} - \tilde{x}_{i+1}\tilde{y}_i)$$

We define the remaining six terms in (4.1) as follows:

$$t_1 = \tfrac{1}{2}(\tilde{x}_i\eta_{i+1}) \qquad t_4 = -\tfrac{1}{2}(\tilde{y}_i\epsilon_{i+1})$$

$$t_2 = \tfrac{1}{2}(\tilde{y}_{i+1}\epsilon_i) \qquad t_5 = \tfrac{1}{2}(\epsilon_i\eta_{i+1})$$

$$t_3 = -\tfrac{1}{2}(\tilde{x}_{i+1}\eta_i) \qquad t_6 = -\tfrac{1}{2}(\epsilon_{i+1}\eta_i)$$

And we note:

$$E(t_1) = \tfrac{1}{2}\tilde{x}_i E(\eta_{i+1}) = 0 \qquad E(t_4) = -\tfrac{1}{2}\tilde{y}_i E(\epsilon_{i+1}) = 0$$

$$E(t_2) = \tfrac{1}{2}\tilde{y}_{i+1} E(\epsilon_i) = 0 \qquad E(t_5) = \tfrac{1}{2}E(\epsilon_i\eta_{i+1}) = 0$$

$$E(t_3) = -\tfrac{1}{2}\tilde{x}_{i+1}E(\eta_i) = 0 \qquad E(t_6) = -\tfrac{1}{2}E(\epsilon_{i+1}\eta_i) = 0$$

Thus, taking the expectation of (4.1) yields:

$$E(A_i) = a_i + \Big(E(t_1) + E(t_2) + E(t_3) + E(t_4) + E(t_5) + E(t_6)\Big) = a_i$$

Evidently, the mean area of a triangle with coordinate errors coincides with its nominal area. If we are willing to assume that coordinate errors are zero, on average, then the estimated area equals the true area, on average. However, an individual area estimate will deviate from the true area, so a precision estimate is the next logical step. To get the variance of area of a triangle, we can take the variance of (4.1), which can be expressed as the sum of the variances of the individual terms plus twice the sum of the covariances:

$$Var(A_i) = \sum_{i=1}^{6} Var(t_i) + 2 * \sum_{i<j} Cov(t_i,t_j) \tag{4.2}$$

where:

$$Var(t_1) = \tfrac{1}{4}\tilde{x}_i^2\sigma_{i+1}^2 \qquad Var(t_4) = \tfrac{1}{4}\tilde{y}_i^2\sigma_{i+1}^2$$

$$Var(t_2) = \tfrac{1}{4}\tilde{y}_{i+1}^2\sigma_i^2 \qquad Var(t_5) = \tfrac{1}{4}\sigma_i^2\sigma_{i+1}^2$$

$$Var(t_3) = \tfrac{1}{4}\tilde{x}_{i+1}^2\sigma_i^2 \qquad Var(t_6) = \tfrac{1}{4}\sigma_{i+1}^2\sigma_i^2$$

So the sum of the variance terms is:

$$\sum_{i=1}^{6} Var(t_i) = \tfrac{1}{4} * \Big((\tilde{x}_i^2 + \tilde{y}_i^2)\sigma_{i+1}^2 + (\tilde{x}_{i+1}^2 + \tilde{y}_{i+1}^2)\sigma_i^2 + 2\sigma_i^2\sigma_{i+1}^2\Big) \tag{4.3}$$

For the covariances we have:

$$Cov(t_i,t_j) = E(t_it_j) - E(t_i)*E(t_j) = E(t_it_j)$$

So:

$$Cov(t_1,t_2) = \tfrac{1}{4}\tilde{x}_i\tilde{y}_{i+1}E(\eta_{i+1}\epsilon_i) = 0$$

$$Cov(t_1,t_3) = -\tfrac{1}{4}\tilde{x}_i\tilde{x}_{i+1}E(\eta_{i+1}\eta_i) = -\tfrac{1}{4}\tilde{x}_i\tilde{x}_{i+1}\sigma_i\sigma_{i+1}\rho_i$$

$$Cov(t_1,t_4) = -\tfrac{1}{4}\tilde{x}_i\tilde{y}_iE(\eta_{i+1}\epsilon_{i+1}) = 0$$

$$Cov(t_1,t_5) = \tfrac{1}{4}\tilde{x}_iE(\eta_{i+1}\eta_{i+1}\epsilon_i) = \tfrac{1}{4}\tilde{x}_i\sigma_{i+1}^2E(\epsilon_i) = 0$$

$$Cov(t_1,t_6) = -\tfrac{1}{4}\tilde{x}_iE(\eta_{i+1}\eta_i\epsilon_{i+1}) = -\tfrac{1}{4}\tilde{x}_i\sigma_i\sigma_{i+1}\rho_iE(\epsilon_{i+1}) = 0$$

$$\mathrm{Cov}(t_2,t_3) = -\tfrac{1}{4}\tilde{y}_{i+1}\tilde{x}_{i+1}E(\epsilon_i\eta_i) = 0$$

$$\mathrm{Cov}(t_2,t_4) = -\tfrac{1}{4}\tilde{y}_i\tilde{y}_{i+1}E(\epsilon_i\epsilon_{i+1}) = -\tfrac{1}{4}\tilde{y}_i\tilde{y}_{i+1}\sigma_i\sigma_{i+1}\rho_i$$

$$\mathrm{Cov}(t_2,t_5) = \tfrac{1}{4}\tilde{y}_{i+1}E(\epsilon_i\epsilon_i\eta_{i+1}) = \tfrac{1}{4}\tilde{y}_{i+1}\sigma_i^2E(\eta_{i+1}) = 0$$

$$\mathrm{Cov}(t_2,t_6) = -\tfrac{1}{4}\tilde{y}_{i+1}E(\epsilon_i\epsilon_{i+1}\eta_i) = -\tfrac{1}{4}\tilde{y}_{i+1}\sigma_i\sigma_{i+1}\rho_iE(\eta_i) = 0$$

$$\mathrm{Cov}(t_3,t_4) = \tfrac{1}{4}\tilde{x}_{i+1}\tilde{y}_iE(\eta_i\epsilon_{i+1}) = 0$$

$$\mathrm{Cov}(t_3,t_5) = -\tfrac{1}{4}\tilde{x}_{i+1}E(\eta_i\eta_{i+1}\epsilon_i) = -\tfrac{1}{4}\tilde{x}_{i+1}\sigma_i\sigma_{i+1}\rho_iE(\epsilon_i) = 0$$

$$\mathrm{Cov}(t_3,t_6) = \tfrac{1}{4}\tilde{x}_{i+1}E(\eta_i\eta_i\epsilon_{i+1}) = \tfrac{1}{4}\tilde{x}_{i+1}\sigma_i^2E(\epsilon_{i+1}) = 0$$

$$\mathrm{Cov}(t_4,t_5) = -\tfrac{1}{4}\tilde{y}_iE(\epsilon_i\epsilon_{i+1}\eta_{i+1}) = -\tfrac{1}{4}\tilde{y}_i\sigma_i\sigma_{i+1}\rho_iE(\eta_{i+1}) = 0$$

$$\mathrm{Cov}(t_4,t_6) = \tfrac{1}{4}\tilde{y}_iE(\epsilon_{i+1}\epsilon_{i+1}\eta_i) = \tfrac{1}{4}\tilde{y}_i\sigma_{i+1}^2E(\eta_i) = 0$$

$$\mathrm{Cov}(t_5,t_6) = -\tfrac{1}{4}E(\epsilon_i\epsilon_{i+1}\eta_i\eta_{i+1}) = -\tfrac{1}{4}E(\epsilon_i\epsilon_{i+1})E(\eta_i\eta_{i+1}) = -\tfrac{1}{4}\sigma_i^2\sigma_{i+1}^2\rho_i^2$$

And twice the sum of the covariances is:

$$2 * \sum_{i<j} \mathrm{Cov}(t_i,t_j) = -\tfrac{1}{2} * \left((\tilde{x}_i\tilde{x}_{i+1} + \tilde{y}_i\tilde{y}_{i+1})\sigma_i\sigma_{i+1}\rho_i + \sigma_i^2\sigma_{i+1}^2\rho_i^2\right) \tag{4.4}$$

Thus, substituting (4.4) and (4.3) into (4.2);

$$\mathrm{Var}(A_i) = \tfrac{1}{4} * \left(r_i^2\sigma_{i+1}^2 + r_{i+1}^2\sigma_i^2 - 2(\tilde{x}_i\tilde{x}_{i+1} + \tilde{y}_i\tilde{y}_{i+1})\sigma_i\sigma_{i+1}\rho_i + 2(1-\rho_i^2)\sigma_i^2\sigma_{i+1}^2\right) \tag{4.5}$$

Where: $r_i^2 = \tilde{x}_i^2 + \tilde{y}_i^2$ and $r_{i+1}^2 = \tilde{x}_{i+1}^2 + \tilde{y}_{i+1}^2$

Equation (4.5) gives the variance of area of a triangle formed by any pair of adjacent points in a polygon boundary. The area of a polygon is simply the sum of the areas of the n individual triangles:

$$A_N = \sum_{i=1}^{n} A_i = \sum_{i=1}^{n} \left(\tfrac{1}{2} * \left(\tilde{X}_i\tilde{Y}_{i+1} - \tilde{X}_{i+1}\tilde{Y}_i\right)\right)$$

where the sum is "circular", i.e.,

$$\tilde{X}_{n+1} = \tilde{X}_1 \qquad \tilde{Y}_{n+1} = \tilde{Y}_1$$

This expression yields a positive area when the coordinates are indexed in a counter-clockwise

direction. Individual triangles may be positive or negative in area, but the sum should be positive. The mean polygon area is:

$$E(A_N) = E\left(\sum_{i=1}^{n} \tfrac{1}{2} * \left(\tilde{X}_i \tilde{Y}_{i+1} - \tilde{X}_{i+1} \tilde{Y}_i\right)\right) = \sum_{i=1}^{n} a_i$$

which is the nominal polygon area. Because errors in area of triangles are *not* independent, covariance terms will be required for triangles in order to derive the variance of A_N. Ignoring covariance of non-adjacent triangles, the variance of A_N can be expressed as the variance of a sum:

$$Var(A_N) = Var\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} Var(A_i) + 2 * \sum_{i=2}^{n+1} Cov(A_{i-1}, A_i) \tag{4.6}$$

Where: $A_{n+1} = A_1$

The variance of the A_i is given in (4.5). To derive $Cov(A_{i-1}, A_i)$ we begin by noting:

$$Cov(A_{i-1}, A_i) = E(A_{i-1}A_i) - E(A_{i-1}) * E(A_i) = E(A_{i-1}A_i) - a_{i-1}a_i \tag{4.7}$$

And, $A_{i-1}A_i = \tfrac{1}{4} * \left((\tilde{X}_{i-1}\tilde{Y}_i - \tilde{X}_i\tilde{Y}_{i-1}) * (\tilde{X}_i\tilde{Y}_{i+1} - \tilde{X}_{i+1}\tilde{Y}_i)\right)$

$$= \tfrac{1}{4} * \left(\tilde{X}_{i-1}\tilde{X}_i\tilde{Y}_i\tilde{Y}_{i+1} - \tilde{X}_{i-1}\tilde{X}_{i+1}\tilde{Y}_i^2 - \tilde{X}_i^2\tilde{Y}_{i-1}\tilde{Y}_{i+1} + \tilde{X}_i\tilde{X}_{i+1}\tilde{Y}_{i-1}\tilde{Y}_i\right)$$

Taking these terms one by one, we define:

$$q_1 = \tilde{X}_{i-1}\tilde{X}_i\tilde{Y}_i\tilde{Y}_{i+1} = (\tilde{x}_{i-1} + \epsilon_{i-1})(\tilde{x}_i + \epsilon_i)(\tilde{y}_i + \eta_i)(\tilde{y}_{i+1} + \eta_{i+1})$$

$$= (\tilde{x}_{i-1}\tilde{x}_i + \tilde{x}_{i-1}\epsilon_i + \tilde{x}_i\epsilon_{i-1} + \epsilon_{i-1}\epsilon_i)(\tilde{y}_i\tilde{y}_{i+1} + \tilde{y}_i\eta_{i+1} + \tilde{y}_{i+1}\eta_i + \eta_i\eta_{i+1})$$

$$= \tilde{x}_{i-1}\tilde{x}_i\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_{i-1}\tilde{x}_i\tilde{y}_i\eta_{i+1} + \tilde{x}_{i-1}\tilde{x}_i\tilde{y}_{i+1}\eta_i + \tilde{x}_{i-1}\tilde{x}_i\eta_i\eta_{i+1} +$$

$$\tilde{x}_{i-1}\epsilon_i\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_{i-1}\epsilon_i\tilde{y}_i\eta_{i+1} + \tilde{x}_{i-1}\epsilon_i\tilde{y}_{i+1}\eta_i + \tilde{x}_{i-1}\epsilon_i\eta_i\eta_{i+1} +$$

$$\tilde{x}_i\epsilon_{i-1}\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_i\epsilon_{i-1}\tilde{y}_i\eta_{i+1} + \tilde{x}_i\epsilon_{i-1}\tilde{y}_{i+1}\eta_i + \tilde{x}_i\epsilon_{i-1}\eta_i\eta_{i+1} +$$

$$\epsilon_{i-1}\epsilon_i\tilde{y}_i\tilde{y}_{i+1} + \epsilon_{i-1}\epsilon_i\tilde{y}_i\eta_{i+1} + \epsilon_{i-1}\epsilon_i\tilde{y}_{i+1}\eta_i + \epsilon_{i-1}\epsilon_i\eta_i\eta_{i+1}$$

Similarly,

$$q_2 = \tilde{X}_{i-1}\tilde{X}_{i+1}\tilde{Y}_i^2 = (\tilde{x}_{i-1} + \epsilon_{i-1})(\tilde{x}_{i+1} + \epsilon_{i+1})(\tilde{y}_i + \eta_i)^2$$

$$= (\tilde{x}_{i-1}\tilde{x}_{i+1} + \tilde{x}_{i-1}\epsilon_{i+1} + \tilde{x}_{i+1}\epsilon_{i-1} + \epsilon_{i-1}\epsilon_{i+1})(\tilde{y}_i^2 + 2\tilde{y}_i\eta_i + \eta_i^2)$$

$$= \tilde{x}_{i-1}\tilde{x}_{i+1}\tilde{y}_i^2 + \tilde{x}_{i-1}\tilde{x}_{i+1}2\tilde{y}_i\eta_i + \tilde{x}_{i-1}\tilde{x}_{i+1}\eta_i^2 + \tilde{x}_{i-1}\epsilon_{i+1}\tilde{y}_i^2 +$$
$$\tilde{x}_{i-1}\epsilon_{i+1}2\tilde{y}_i\eta_i + \tilde{x}_{i-1}\epsilon_{i+1}\eta_i^2 + \tilde{x}_{i+1}\epsilon_{i-1}\tilde{y}_i^2 + \tilde{x}_{i+1}\epsilon_{i-1}2\tilde{y}_i\eta_i +$$
$$\tilde{x}_{i+1}\epsilon_{i-1}\eta_i^2 + \epsilon_{i-1}\epsilon_{i+1}\tilde{y}_i^2 + \epsilon_{i-1}\epsilon_{i+1}2\tilde{y}_i\eta_i + \epsilon_{i-1}\epsilon_{i+1}\eta_i^2$$

$$q_3 = \tilde{X}_i^2\tilde{Y}_{i-1}\tilde{Y}_{i+1} = (\tilde{x}_i + \epsilon_i)^2(\tilde{y}_{i-1} + \eta_{i-1})(\tilde{y}_{i+1} + \eta_{i+1})$$
$$= (\tilde{x}_i^2 + 2\tilde{x}_i\epsilon_i + \epsilon_i^2)(\tilde{y}_{i-1}\tilde{y}_{i+1} + \tilde{y}_{i-1}\eta_{i+1} + \tilde{y}_{i+1}\eta_{i-1} + \eta_{i-1}\eta_{i+1})$$
$$= \tilde{x}_i^2\tilde{y}_{i-1}\tilde{y}_{i+1} + \tilde{x}_i^2\tilde{y}_{i-1}\eta_{i+1} + \tilde{x}_i^2\tilde{y}_{i+1}\eta_{i-1} + \tilde{x}_i^2\eta_{i-1}\eta_{i+1} +$$
$$2\tilde{x}_i\epsilon_i\tilde{y}_{i-1}\tilde{y}_{i+1} + 2\tilde{x}_i\epsilon_i\tilde{y}_{i-1}\eta_{i+1} + 2\tilde{x}_i\epsilon_i\tilde{y}_{i+1}\eta_{i-1} + 2\tilde{x}_i\epsilon_i\eta_{i-1}\eta_{i+1} +$$
$$\epsilon_i^2\tilde{y}_{i-1}\tilde{y}_{i+1} + \epsilon_i^2\tilde{y}_{i-1}\eta_{i+1} + \epsilon_i^2\tilde{y}_{i+1}\eta_{i-1} + \epsilon_i^2\eta_{i-1}\eta_{i+1}$$

$$q_4 = \tilde{X}_i\tilde{X}_{i+1}\tilde{Y}_{i-1}\tilde{Y}_i = (\tilde{x}_i + \epsilon_i)(\tilde{x}_{i+1} + \epsilon_{i+1})(\tilde{y}_{i-1} + \eta_{i-1})(\tilde{y}_i + \eta_i)$$
$$= (\tilde{x}_i\tilde{x}_{i+1} + \tilde{x}_i\epsilon_{i+1} + \tilde{x}_{i+1}\epsilon_i + \epsilon_i\epsilon_{i+1})(\tilde{y}_{i-1}\tilde{y}_i + \tilde{y}_{i-1}\eta_i + \tilde{y}_i\eta_{i-1} + \eta_{i-1}\eta_i)$$
$$= \tilde{x}_i\tilde{x}_{i+1}\tilde{y}_{i-1}\tilde{y}_i + \tilde{x}_i\tilde{x}_{i+1}\tilde{y}_{i-1}\eta_i + \tilde{x}_i\tilde{x}_{i+1}\tilde{y}_i\eta_{i-1} + \tilde{x}_i\tilde{x}_{i+1}\eta_{i-1}\eta_i +$$
$$\tilde{x}_i\epsilon_{i+1}\tilde{y}_{i-1}\tilde{y}_i + \tilde{x}_i\epsilon_{i+1}\tilde{y}_{i-1}\eta_i + \tilde{x}_i\epsilon_{i+1}\tilde{y}_i\eta_{i-1} + \tilde{x}_i\epsilon_{i+1}\eta_{i-1}\eta_i +$$
$$\tilde{x}_{i+1}\epsilon_i\tilde{y}_{i-1}\tilde{y}_i + \tilde{x}_{i+1}\epsilon_i\tilde{y}_{i-1}\eta_i + \tilde{x}_{i+1}\epsilon_i\tilde{y}_i\eta_{i-1} + \tilde{x}_{i+1}\epsilon_i\eta_{i-1}\eta_i +$$
$$\epsilon_i\epsilon_{i+1}\tilde{y}_{i-1}\tilde{y}_i + \epsilon_i\epsilon_{i+1}\tilde{y}_{i-1}\eta_i + \epsilon_i\epsilon_{i+1}\tilde{y}_i\eta_{i-1} + \epsilon_i\epsilon_{i+1}\eta_{i-1}\eta_i$$

So: $A_{i-1}A_i = \frac{1}{4}*(q_1 - q_2 - q_3 + q_4)$

And: $E(A_{i-1}A_i) = \frac{1}{4}*\Big(E(q_1) - E(q_2) - E(q_3) + E(q_4)\Big)$ (4.8)

Again going term by term:

$$E(q_1) = \tilde{x}_{i-1}\tilde{x}_i\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_{i-1}\tilde{x}_i\tilde{y}_iE(\eta_{i+1}) + \tilde{x}_{i-1}\tilde{x}_i\tilde{y}_{i+1}E(\eta_i) + \tilde{x}_{i-1}\tilde{x}_iE(\eta_i\eta_{i+1}) +$$
$$\tilde{x}_{i-1}\tilde{y}_i\tilde{y}_{i+1}E(\epsilon_i) + \tilde{x}_{i-1}\tilde{y}_iE(\epsilon_i\eta_{i+1}) + \tilde{x}_{i-1}\tilde{y}_{i+1}E(\epsilon_i\eta_i) + \tilde{x}_{i-1}E(\epsilon_i\eta_i\eta_{i+1}) +$$
$$\tilde{x}_i\tilde{y}_i\tilde{y}_{i+1}E(\epsilon_{i-1}) + \tilde{x}_i\tilde{y}_iE(\epsilon_{i-1}\eta_{i+1}) + \tilde{x}_i\tilde{y}_{i+1}E(\epsilon_{i-1}\eta_i) + \tilde{x}_iE(\epsilon_{i-1}\eta_i\eta_{i+1}) +$$
$$\tilde{y}_i\tilde{y}_{i+1}E(\epsilon_{i-1}\epsilon_i) + \tilde{y}_iE(\epsilon_{i-1}\epsilon_i\eta_{i+1}) + \tilde{y}_{i+1}E(\epsilon_{i-1}\epsilon_i\eta_i) + E(\epsilon_{i-1}\epsilon_i\eta_i\eta_{i+1})$$

$$E(q_1) = \tilde{x}_{i-1}\tilde{x}_i\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_{i-1}\tilde{x}_i\sigma_i\sigma_{i+1}\rho_i + \tilde{x}_{i-1}E(\epsilon_i)E(\eta_i\eta_{i+1}) +$$
$$\tilde{x}_iE(\epsilon_{i-1})E(\eta_i\eta_{i+1}) + \tilde{y}_i\tilde{y}_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + \tilde{y}_iE(\epsilon_{i-1}\epsilon_i)E(\eta_{i+1}) +$$

$$\tilde{y}_{i+1}E(\epsilon_{i-1}\epsilon_i)E(\eta_i) + E(\epsilon_{i-1}\epsilon_i)E(\eta_i\eta_{i+1})$$

$$E(q_1) = \tilde{x}_{i-1}\tilde{x}_i\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_{i-1}\tilde{x}_i\sigma_i\sigma_{i+1}\rho_i + \tilde{y}_i\tilde{y}_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + \sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i \tag{4.9}$$

$$\begin{aligned}
E(q_2) &= \tilde{x}_{i-1}\tilde{x}_{i+1}\tilde{y}_i^2 + \tilde{x}_{i-1}\tilde{x}_{i+1}2\tilde{y}_iE(\eta_i) + \tilde{x}_{i-1}\tilde{x}_{i+1}E(\eta_i^2) + \\
&\quad \tilde{x}_{i-1}\tilde{y}_i^2E(\epsilon_{i+1}) + \tilde{x}_{i-1}2\tilde{y}_iE(\epsilon_{i+1}\eta_i) + \tilde{x}_{i-1}E(\epsilon_{i+1}\eta_i^2) + \\
&\quad \tilde{x}_{i+1}\tilde{y}_i^2E(\epsilon_{i-1}) + \tilde{x}_{i+1}2\tilde{y}_iE(\epsilon_{i-1}\eta_i) + \tilde{x}_{i+1}E(\epsilon_{i-1}\eta_i^2) + \\
&\quad \tilde{y}_i^2E(\epsilon_{i-1}\epsilon_{i+1}) + 2\tilde{y}_iE(\epsilon_{i-1}\epsilon_{i+1}\eta_i) + E(\epsilon_{i-1}\epsilon_{i+1}\eta_i^2) \\
E(q_2) &= \tilde{x}_{i-1}\tilde{x}_{i+1}\tilde{y}_i^2 + \tilde{x}_{i-1}\tilde{x}_{i+1}\sigma_i^2 + \tilde{x}_{i-1}E(\epsilon_{i+1})E(\eta_i^2) + \\
&\quad \tilde{x}_{i+1}E(\epsilon_{i-1})E(\eta_i^2) + 2\tilde{y}_iE(\epsilon_{i-1}\epsilon_{i+1})E(\eta_i) + E(\epsilon_{i-1}\epsilon_{i+1})E(\eta_i^2) \\
E(q_2) &= \tilde{x}_{i-1}\tilde{x}_{i+1}\tilde{y}_i^2 + \tilde{x}_{i-1}\tilde{x}_{i+1}\sigma_i^2
\end{aligned} \tag{4.10}$$

$$\begin{aligned}
E(q_3) &= \tilde{x}_i^2\tilde{y}_{i-1}\tilde{y}_{i+1} + \tilde{x}_i^2\tilde{y}_{i-1}E(\eta_{i+1}) + \tilde{x}_i^2\tilde{y}_{i+1}E(\eta_{i-1}) + \tilde{x}_i^2E(\eta_{i-1}\eta_{i+1}) + \\
&\quad 2\tilde{x}_i\tilde{y}_{i-1}\tilde{y}_{i+1}E(\epsilon_i) + 2\tilde{x}_i\tilde{y}_{i-1}E(\epsilon_i\eta_{i+1}) + 2\tilde{x}_i\tilde{y}_{i+1}E(\epsilon_i\eta_{i-1}) + 2\tilde{x}_iE(\epsilon_i\eta_{i-1}\eta_{i+1}) + \\
&\quad \tilde{y}_{i-1}\tilde{y}_{i+1}E(\epsilon_i^2) + \tilde{y}_{i-1}E(\epsilon_i^2\eta_{i+1}) + \tilde{y}_{i+1}E(\epsilon_i^2\eta_{i-1}) + E(\epsilon_i^2\eta_{i-1}\eta_{i+1}) \\
E(q_3) &= \tilde{x}_i^2\tilde{y}_{i-1}\tilde{y}_{i+1} + 2\tilde{x}_iE(\epsilon_i)E(\eta_{i-1}\eta_{i+1}) + \tilde{y}_{i-1}\tilde{y}_{i+1}\sigma_i^2 + \\
&\quad \tilde{y}_{i-1}E(\epsilon_i^2)E(\eta_{i+1}) + \tilde{y}_{i+1}E(\epsilon_i^2)E(\eta_{i-1}) + E(\epsilon_i^2)E(\eta_{i-1}\eta_{i+1}) \\
E(q_3) &= \tilde{x}_i^2\tilde{y}_{i-1}\tilde{y}_{i+1} + \tilde{y}_{i-1}\tilde{y}_{i+1}\sigma_i^2
\end{aligned} \tag{4.11}$$

$$\begin{aligned}
E(q_4) &= \tilde{x}_i\tilde{x}_{i+1}\tilde{y}_{i-1}\tilde{y}_i + \tilde{x}_i\tilde{x}_{i+1}\tilde{y}_{i-1}E(\eta_i) + \tilde{x}_i\tilde{x}_{i+1}\tilde{y}_iE(\eta_{i-1}) + \tilde{x}_i\tilde{x}_{i+1}E(\eta_{i-1}\eta_i) + \\
&\quad \tilde{x}_i\tilde{y}_{i-1}\tilde{y}_iE(\epsilon_{i+1}) + \tilde{x}_i\tilde{y}_{i-1}E(\epsilon_{i+1}\eta_i) + \tilde{x}_i\tilde{y}_iE(\epsilon_{i+1}\eta_{i-1}) + \tilde{x}_iE(\epsilon_{i+1}\eta_{i-1}\eta_i) + \\
&\quad \tilde{x}_{i+1}\tilde{y}_{i-1}\tilde{y}_iE(\epsilon_i) + \tilde{x}_{i+1}\tilde{y}_{i-1}E(\epsilon_i\eta_i) + \tilde{x}_{i+1}\tilde{y}_iE(\epsilon_i\eta_{i-1}) + \tilde{x}_{i+1}E(\epsilon_i\eta_{i-1}\eta_i) + \\
&\quad \tilde{y}_{i-1}\tilde{y}_iE(\epsilon_i\epsilon_{i+1}) + \tilde{y}_{i-1}E(\epsilon_i\epsilon_{i+1}\eta_i) + \tilde{y}_iE(\epsilon_i\epsilon_{i+1}\eta_{i-1}) + E(\epsilon_i\epsilon_{i+1}\eta_{i-1}\eta_i) \\
E(q_4) &= \tilde{x}_i\tilde{x}_{i+1}\tilde{y}_{i-1}\tilde{y}_i + \tilde{x}_i\tilde{x}_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + \tilde{x}_iE(\epsilon_{i+1})E(\eta_{i-1}\eta_i) + \\
&\quad \tilde{x}_{i+1}E(\epsilon_i)E(\eta_{i-1}\eta_i) + \tilde{y}_{i-1}\tilde{y}_i\sigma_i\sigma_{i+1}\rho_i + \tilde{y}_{i-1}E(\epsilon_i\epsilon_{i+1})E(\eta_i) + \\
&\quad \tilde{y}_iE(\epsilon_i\epsilon_{i+1})E(\eta_{i-1}) + E(\epsilon_i\epsilon_{i+1})E(\eta_{i-1}\eta_i)
\end{aligned}$$

$$E(q_4) = \tilde{x}_i\tilde{x}_{i+1}\tilde{y}_{i-1}\tilde{y}_i + \tilde{x}_i\tilde{x}_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + \tilde{y}_{i-1}\tilde{y}_i\sigma_i\sigma_{i+1}\rho_i + \sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i \qquad (4.12)$$

Substituting the expressions in (4.9), (4.10), (4.11), and (4.12) into (4.8) we have:

$$\begin{aligned} E(A_{i-1}A_i) = \tfrac{1}{4}*\Big(&\tilde{x}_{i-1}\tilde{x}_i\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_{i-1}\tilde{x}_i\sigma_i\sigma_{i+1}\rho_i + \tilde{y}_i\tilde{y}_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + \\ &\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i - \tilde{x}_{i-1}\tilde{x}_{i+1}\tilde{y}_i^2 - \tilde{x}_{i-1}\tilde{x}_{i+1}\sigma_i^2 - \\ &\tilde{x}_i^2\tilde{y}_{i-1}\tilde{y}_{i+1} - \tilde{y}_{i-1}\tilde{y}_{i+1}\sigma_i^2 + \tilde{x}_i\tilde{x}_{i+1}\tilde{y}_{i-1}\tilde{y}_i + \\ &\tilde{x}_i\tilde{x}_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + \tilde{y}_{i-1}\tilde{y}_i\sigma_i\sigma_{i+1}\rho_i + \sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i\Big) \end{aligned} \qquad (4.13)$$

Note that:

$$\begin{aligned} a_{i-1}a_i &= \tfrac{1}{4}*(\tilde{x}_{i-1}\tilde{y}_i - \tilde{x}_i\tilde{y}_{i-1})(\tilde{x}_i\tilde{y}_{i+1} - \tilde{x}_{i+1}\tilde{y}_i) \\ &= \tfrac{1}{4}*(\tilde{x}_{i-1}\tilde{x}_i\tilde{y}_i\tilde{y}_{i+1} - \tilde{x}_{i-1}\tilde{x}_{i+1}\tilde{y}_i^2 - \tilde{x}_i^2\tilde{y}_{i-1}\tilde{y}_{i+1} + \tilde{x}_i\tilde{x}_{i+1}\tilde{y}_{i-1}\tilde{y}_i) \end{aligned} \qquad (4.14)$$

Substituting (4.13) and (4.14) into (4.7) gives:

$$\begin{aligned} \mathrm{Cov}(A_{i-1},A_i) = \tfrac{1}{4}*\Big(&\tilde{x}_{i-1}\tilde{x}_i\sigma_i\sigma_{i+1}\rho_i + \tilde{y}_i\tilde{y}_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + \\ &\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i - \tilde{x}_{i-1}\tilde{x}_{i+1}\sigma_i^2 - \tilde{y}_{i-1}\tilde{y}_{i+1}\sigma_i^2 + \\ &\tilde{x}_i\tilde{x}_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + \tilde{y}_{i-1}\tilde{y}_i\sigma_i\sigma_{i+1}\rho_i + \sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i\Big) \end{aligned}$$

Combining terms we get:

$$\begin{aligned} \mathrm{Cov}(A_{i-1},A_i) = \tfrac{1}{4}*\Big(&(\tilde{x}_{i-1}\tilde{x}_i + \tilde{y}_{i-1}\tilde{y}_i)\sigma_i\sigma_{i+1}\rho_i + (\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_i\tilde{x}_{i+1})\sigma_{i-1}\sigma_i\rho_{i-1} + \\ &2\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i - (\tilde{x}_{i-1}\tilde{x}_{i+1} + \tilde{y}_{i-1}\tilde{y}_{i+1})\sigma_i^2\Big) \end{aligned}$$

This, then, is the expression for covariance between areas of adjacent triangles. Substituting into (4.6), we are now ready to complete the summation:

$$\mathrm{Var}(A_N) = \tfrac{1}{4}* \sum_{i=1}^{n}\Big(r_i^2\sigma_{i+1}^2 + r_{i+1}^2\sigma_i^2 - 2(\tilde{x}_i\tilde{x}_{i+1} + \tilde{y}_i\tilde{y}_{i+1})\sigma_i\sigma_{i+1}\rho_i + 2(1-\rho_i^2)\sigma_i^2\sigma_{i+1}^2\Big)$$

$$\begin{aligned} + \tfrac{1}{2}* \sum_{i=2}^{n+1}\Big(&(\tilde{x}_{i-1}\tilde{x}_i + \tilde{y}_{i-1}\tilde{y}_i)\sigma_i\sigma_{i+1}\rho_i + (\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_i\tilde{x}_{i+1})\sigma_{i-1}\sigma_i\rho_{i-1} + \\ &2\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i - (\tilde{x}_{i-1}\tilde{x}_{i+1} + \tilde{y}_{i-1}\tilde{y}_{i+1})\sigma_i^2\Big) \end{aligned}$$

To simplify this expression, we can define the following:

$$w_i = \tilde{x}_i\tilde{x}_{i+1} + \tilde{y}_i\tilde{y}_{i+1}$$

$$z_i = \tilde{x}_{i-1}\tilde{x}_{i+1} + \tilde{y}_{i-1}\tilde{y}_{i+1}$$

$$s_i = \sigma_i \sigma_{i+1} \rho_i$$

which yields:

$$\mathrm{Var}(A_N) = \tfrac{1}{4} * \sum_{i=1}^{n} \Big(r_i^2 \sigma_{i+1}^2 + r_{i+1}^2 \sigma_i^2 - 2w_i s_i + 2\sigma_i^2 \sigma_{i+1}^2 - 2s_i^2 + 2w_{i-1} s_i + 2w_i s_{i-1} + 4s_{i-1} s_i - 2z_i \sigma_i^2 \Big) \qquad (4.15)$$

Using the fact that this sum is "circular", we can rewrite this as:

$$\mathrm{Var}(A_N) = \tfrac{1}{4} * \sum_{i=1}^{n} \Big(r_i^2 (\sigma_{i-1}^2 + \sigma_{i+1}^2) + 2w_i (s_{i-1} - s_i + s_{i+1}) + 2\sigma_i^2 \sigma_{i+1}^2 - 2s_i^2 + 4s_{i-1} s_i - 2z_i \sigma_i^2 \Big)$$

Note that this expression for variance of area of a polygon, in terms of the coordinates, the point variances (σ_i's) and the correlations between errors at adjacent points (ρ_i's), does not explicitly include representation of the arcs which comprise the polygon. Instead, individual points which make the polygon boundary are used. The identification of arcs is not necessary; the results are the same. However, in a computer implementation of this formulation, some efficiencies will be noted if arcs are considered.

Derivation of the Minimum-Variance Centroid

At this point, we may consider the problem of consistently defining what is meant by the polygon centroid. It was noted that the location of the centroid used to center the coordinates prior to variance calculations affects the value of the variance obtained, possibly due to the omission of covariance of non-adjacent triangles. Therefore, it became necessary to establish a consistent method for determining the location of the centroid. (Here we are concerned with a centroid for use in variance calculations only; this centroid need not be in a polygon interior, and should not affect any other processing steps). For the purposes of estimating polygon area variance, a useful approach is to select the centroid location which *minimizes* the variance, such as Bondesson (1986) used when evaluating the variance of area obtained from traversing.

To determine the location of the minimum-variance centroid (MVC), the formula for polygon variance is written as a function of the arc variances (σ_i's), the correlations associated with the arcs (ρ_i's), the point coordinates (x_i's and y_i's), and the centroid coordinates (X_c and Y_c). Taking the derivative of variance with respect to X_c and Y_c, and solving for X_c and Y_c will yield the centroid coordinates which will minimize or maximize variance. Rewriting the formula for the variance of a polygon made of n points (4.15),

$$V = \tfrac{1}{4} * \sum_{i=1}^{n} \Big(\tilde{x}_i{}^2\sigma_{i+1}^2 + \tilde{y}_i{}^2\sigma_{i+1}^2 + \tilde{x}_{i+1}{}^2\sigma_i^2 + \tilde{y}_{i+1}{}^2\sigma_i^2 + 2\sigma_i^2\sigma_{i+1}^2(1-\rho_i^2) - 2\sigma_i\sigma_{i+1}\rho_i(\tilde{x}_i\tilde{x}_{i+1} + \tilde{y}_i\tilde{y}_{i+1}) \Big) + \tfrac{1}{4} * \sum_{i=1}^{n} \Big(4\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i + 2\sigma_i\sigma_{i+1}\rho_i(\tilde{x}_{i-1}\tilde{x}_i + \tilde{y}_{i-1}\tilde{y}_i) + 2\sigma_i\sigma_{i-1}\rho_{i-1}(\tilde{x}_i\tilde{x}_{i+1} + \tilde{y}_i\tilde{y}_{i+1}) - 2\sigma_i^2(\tilde{x}_{i-1}\tilde{x}_{i+1} + \tilde{y}_{i-1}\tilde{y}_{i+1}) \Big) \qquad (4.16)$$

Now, replacing the centered coordinates in equation 4.16 ($\tilde{x}_i$ and $\tilde{y}_i$) with the original coordinates x_i and y_i, where $\tilde{x}_i = x_i - X_c$ and $\tilde{y}_i = y_i - Y_c$; we obtain the following:

$$\begin{aligned} V = \tfrac{1}{4} * \sum_{i=1}^{n} \Big(& \sigma_{i+1}^2\Big((x_i-X_c)^2 + (y_i-Y_c)^2\Big) + \sigma_i^2\Big((x_{i+1}-X_c)^2 + (y_{i+1}-Y_c)^2\Big) \\ & + 2\sigma_i^2\sigma_{i+1}^2(1-\rho_i^2) - 2\sigma_i\sigma_{i+1}\rho_i\Big((x_i-X_c)(x_{i+1}-X_c) + (y_i-Y_c)(y_{i+1}-Y_c)\Big) \\ & + 4\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i + 2\sigma_i\sigma_{i+1}\rho_i\Big((x_{i-1}-X_c)(x_i-X_c) + (y_{i-1}-Y_c)(y_i-Y_c)\Big) \\ & + 2\sigma_{i-1}\sigma_i\rho_{i-1}\Big((x_i-X_c)(x_{i+1}-X_c) + (y_i-Y_c)(y_{i+1}-Y_c)\Big) \\ & - 2\sigma_i^2\Big((x_{i-1}-X_c)(x_{i+1}-X_c) + (y_{i-1}-Y_c)(y_{i+1}-Y_c)\Big) \Big) \end{aligned}$$

This can be expressed in three parts:

$$V = \phi_x + \phi_y + \phi$$

where:

$$\begin{aligned} \phi_x = {} & \sigma_{i+1}^2(x_i^2 - 2x_iX_c + X_c^2) + \sigma_i^2(x_{i+1}^2 - 2x_{i+1}X_c + X_c^2) \\ & - 2\sigma_i\sigma_{i+1}\rho_i(x_ix_{i+1} - x_iX_c - x_{i+1}X_c + X_c^2) \\ & + 2\sigma_i\sigma_{i+1}\rho_i(x_{i-1}x_i - x_iX_c - x_{i-1}X_c + X_c^2) \\ & + 2\sigma_{i-1}\sigma_i\rho_{i-1}(x_ix_{i+1} - x_iX_c - x_{i+1}X_c + X_c^2) \\ & - 2\sigma_i^2(x_{i-1}x_{i+1} - x_{i-1}X_c - x_{i+1}X_c + X_c^2) \end{aligned}$$

$$\phi_y = \sigma_{i+1}^2(y_i^2 - 2y_iY_c + Y_c^2) + \sigma_i^2(y_{i+1}^2 - 2y_{i+1}Y_c + Y_c^2)$$
$$- 2\sigma_i\sigma_{i+1}\rho_i(y_iy_{i+1} - y_iY_c - y_{i+1}Y_c + Y_c^2)$$
$$+ 2\sigma_i\sigma_{i+1}\rho_i(y_{i-1}y_i - y_iY_c - y_{i-1}Y_c + Y_c^2)$$
$$+ 2\sigma_{i-1}\sigma_i\rho_{i-1}(y_iy_{i+1} - y_iY_c - y_{i+1}Y_c + Y_c^2)$$
$$- 2\sigma_i^2(y_{i-1}y_{i+1} - y_{i-1}Y_c - y_{i+1}Y_c + Y_c^2)$$
$$\phi = 2\sigma_i^2\sigma_{i+1}^2(1-\rho_i^2) + 4\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i$$

Note that by separating terms,

$$\frac{\partial V}{\partial X_c} = \frac{\partial \phi_x}{\partial X_c} \qquad \text{and} \qquad \frac{\partial V}{\partial Y_c} = \frac{\partial \phi_y}{\partial Y_c}$$

Thus,

$$\frac{\partial V}{\partial X_c} = \tfrac{1}{4} * \sum_{i=1}^{n} \Big(\sigma_{i+1}^2(-2x_i + 2X_c) + \sigma_i^2(-2x_{i+1} + 2X_c)$$
$$+ 2\sigma_i\sigma_{i+1}\rho_i(x_i + x_{i+1} - 2X_c) - 2\sigma_i\sigma_{i+1}\rho_i(x_i + x_{i-1} - 2X_c)$$
$$- 2\sigma_{i-1}\sigma_i\rho_{i-1}(x_i + x_{i+1} - 2X_c) + 2\sigma_i^2(x_{i-1} + x_{i+1} - 2X_c) \Big) \qquad (4.17)$$

$$= \tfrac{1}{4} * \sum_{i=1}^{n} \Big(-2\sigma_{i+1}^2x_i + 2\sigma_{i+1}^2X_c - 2\sigma_i^2x_{i+1} + 2\sigma_i^2X_c + 2\sigma_i\sigma_{i+1}\rho_ix_i$$
$$+ 2\sigma_i\sigma_{i+1}\rho_ix_{i+1} - 4\sigma_i\sigma_{i+1}\rho_iX_c - 2\sigma_i\sigma_{i+1}\rho_ix_i - 2\sigma_i\sigma_{i+1}\rho_ix_{i-1}$$
$$+ 4\sigma_i\sigma_{i+1}\rho_iX_c - 2\sigma_{i-1}\sigma_i\rho_{i-1}x_i - 2\sigma_{i-1}\sigma_i\rho_{i-1}x_{i+1} + 4\sigma_{i-1}\sigma_i\rho_{i-1}X_c$$
$$+ 2\sigma_i^2x_{i-1} + 2\sigma_i^2x_{i+1} - 4\sigma_i^2X_c \Big)$$

And, combining terms:

$$\frac{\partial V}{\partial X_c} = \tfrac{1}{2} * \sum_{i=1}^{n} \Big(-\sigma_{i+1}^2x_i + \sigma_{i+1}^2X_c + \sigma_i^2x_{i-1} - \sigma_i^2X_c + \sigma_i\sigma_{i+1}\rho_ix_{i+1}$$
$$- \sigma_i\sigma_{i+1}\rho_ix_{i-1} - \sigma_{i-1}\sigma_i\rho_{i-1}x_i - \sigma_{i-1}\sigma_i\rho_{i-1}x_{i+1} + 2\sigma_{i-1}\sigma_i\rho_{i-1}X_c \Big)$$

Setting the derivative equal to zero and moving the terms involving X_c to the left side of the equation we get:

$$\tfrac{1}{2}*\sum_{i=1}^{n} X_c\Big(-\sigma_{i+1}^2+\sigma_i^2-2\sigma_{i-1}\sigma_i\rho_{i-1}\Big) = \tfrac{1}{2}*\sum_{i=1}^{n} x_i\Big(-\sigma_{i+1}^2-\sigma_{i-1}\sigma_i\rho_{i-1}\Big)$$

$$+ \tfrac{1}{2} * \sum_{i=1}^{n} x_{i-1}\left(\sigma_i^2 - \sigma_i \sigma_{i+1} \rho_i\right)$$

$$+ \tfrac{1}{2} * \sum_{i=1}^{n} x_{i+1}\left(\sigma_i \sigma_{i+1} \rho_i - \sigma_{i-1} \sigma_i \rho_{i-1}\right)$$

Next, we multiply both sides by two. Then, we note that the sums are "circular" (i.e., $x_0 = x_n$ and $x_{n+1} = x_1$) and we can make the following substitutions:

$$\sum_{i=1}^{n} x_{i-1} \sigma_i^2 = \sum_{i=1}^{n} x_i \sigma_{i+1}^2$$

$$\sum_{i=1}^{n} x_{i-1} \sigma_i \sigma_{i+1} \rho_i = \sum_{i=1}^{n} x_i \sigma_{i+1} \sigma_{i+2} \rho_{i+1}$$

$$\sum_{i=1}^{n} x_{i+1} \sigma_i \sigma_{i+1} \rho_i = \sum_{i=1}^{n} x_i \sigma_{i-1} \sigma_i \rho_{i-1}$$

$$\sum_{i=1}^{n} x_{i+1} \sigma_{i-1} \sigma_i \rho_{i-1} = \sum_{i=1}^{n} x_i \sigma_{i-2} \sigma_{i-1} \rho_{i-2}$$

And we get:

$$X_c * \left(\sum_{i=1}^{n} -\sigma_{i+1}^2 + \sum_{i=1}^{n} \sigma_i^2 - \sum_{i=1}^{n} 2\sigma_{i-1} \sigma_i \rho_{i-1}\right) =$$

$$\sum_{i=1}^{n} -x_i \sigma_{i+1}^2 + \sum_{i=1}^{n} -x_i \sigma_{i-1} \sigma_i \rho_{i-1} + \sum_{i=1}^{n} x_i \sigma_{i+1}^2 + \sum_{i=1}^{n} -x_i \sigma_{i+1} \sigma_{i+2} \rho_{i+1}$$

$$+ \sum_{i=1}^{n} x_i \sigma_{i-1} \sigma_i \rho_{i-1} + \sum_{i=1}^{n} -x_i \sigma_{i-2} \sigma_{i-1} \rho_{i-2}$$

Now, we note that on the left side of the equation,

$$\sum_{i=1}^{n} -\sigma_{i+1}^2 = -\sum_{i=1}^{n} \sigma_i^2$$

And on the right side of the equation, the first and third terms cancel and the second and fifth terms cancel, leaving:

$$X_c * \left(\sum_{i=1}^{n} - 2\ \sigma_i \sigma_{i+1} \rho_i \right) = \sum_{i=1}^{n} -x_i \left(\sigma_{i+1} \sigma_{i+2} \rho_{i+1} + \sigma_{i-2} \sigma_{i-1} \rho_{i-2} \right)$$

Thus, we have:

$$X_c = \frac{\sum_{i=1}^{n} x_i \left(\sigma_{i-2} \sigma_{i-1} \rho_{i-2} + \sigma_{i+1} \sigma_{i+2} \rho_{i+1} \right)}{\sum_{i=1}^{n} \left(2\ \sigma_i \sigma_{i+1} \rho_i \right)} \tag{4.18}$$

The procedure for ϕ_y follows identically, yielding:

$$Y_c = \frac{\sum_{i=1}^{n} y_i \left(\sigma_{i-2} \sigma_{i-1} \rho_{i-2} + \sigma_{i+1} \sigma_{i+2} \rho_{i+1} \right)}{\sum_{i=1}^{n} \left(2 \sigma_i \sigma_{i+1} \rho_i \right)} \tag{4.19}$$

To verify that this indeed is a minimum, we can take the second derivative of equation (4.17):

$$\frac{\partial^2 V}{\partial X_c^2} = \tfrac{1}{2} * \sum_{i=1}^{n} \left(\sigma_{i+1}^2 - \sigma_i^2 + 2 \sigma_{i-1} \sigma_i \rho_{i-1} \right) = \sum_{i=1}^{n} \left(\sigma_{i-1} \sigma_i \rho_{i-1} \right)$$

It is reasonable to believe that ρ will usually assume positive values. Positive ρ implies that errors at adjacent points are more likely to be in the same direction than in opposite directions; this seems to be the case for most mapping processes which produce errors. When $\rho > 0$, the above expression (and the similar one for the Y terms) is positive, indicating that the solution is a minimum. Thus, the coordinates which minimize the estimated variance of a polygon are weighted averages of the polygon coordinates, in which the weights are the products of σ's and ρ's associated with adjacent coordinates.

Derivation of Covariance between Polygons

After the derivation of polygon variance, the next step is to obtain an expression for the

covariance of area between polygons which share an arc. We will proceed to derive polygon covariance as we did for polygon variance, beginning with covariance between triangles.

First, consider polygon A as a polygon with centroid (X_a, Y_a). It shares an arc with polygon B, whose centroid is at (X_b,Y_b) (Figure 7). We will consider the triangles involved in a sequence of four points on the arc: (X_{i-1},Y_{i-1}), (X_i,Y_i), (X_{i+1},Y_{i+1}), (X_{i+2},Y_{i+2}). Since the sequence of indexing depends on the direction (relative to a centroid), assume the direction of indexing is that which will yield positive areas for polygon A (note direction of arrows in Figure 7). Thus, for polygon A, triangle i, the area is:

$$A_i = \tfrac{1}{2} * \Big((X_i - X_a)(Y_{i+1} - Y_a) - (X_{i+1} - X_a)(Y_i - Y_a)\Big) \tag{4.20}$$

This implies that for polygon B, the direction is reversed; i.e. for polygon B, triangle i, the area is:

$$B_i = \tfrac{1}{2} * \Big((X_{i+1} - X_b)(Y_i - Y_b) - (X_i - X_b)(Y_{i+1} - Y_b)\Big).$$

There are three cases to consider: these involve the covariance between triangle i in A and the three triangles in B with which there is a dependency: triangles B_{i+1}, B_i, and B_{i-1}. Thus, individual expressions are derived for:

$$\mathrm{Cov}(A_i,B_{i-1}), \qquad \mathrm{Cov}(A_i,B_i), \qquad \text{and } \mathrm{Cov}(A_i,B_{i+1})$$

Case 1: $Cov(A_i,B_{i-1})$

By definition, $\mathrm{Cov}(A_i,B_{i-1}) = E(A_iB_{i-1}) - E(A_i)E(B_{i-1})$ (4.21)

We start by obtaining the appropriate expectations. To begin with, equation (4.20) expands to:

$$A_i = \tfrac{1}{2} * \Big((x_i + \epsilon_i - X_a)(y_{i+1} + \eta_{i+1} - Y_a) - (x_{i+1} + \epsilon_{i+1} - X_a)(y_i + \eta_i - Y_a)\Big)$$

$$A_i = \tfrac{1}{2} * \Big(x_i y_{i+1} + x_i \eta_{i+1} - x_i Y_a + y_{i+1}\epsilon_i + \epsilon_i \eta_{i+1} - Y_a \epsilon_i - X_a y_{i+1} - X_a \eta_{i+1} + X_a Y_a$$
$$- x_{i+1} y_i - x_{i+1}\eta_i - x_{i+1} Y_a - y_i \epsilon_{i+1} - \epsilon_{i+1}\eta_i + Y_a \epsilon_{i+1} + y_i X_a + X_a \eta_i - X_a Y_a\Big)$$

Rearranging, we get:

$$A_i = \tfrac{1}{2} * (x_i y_{i+1} - x_i Y_a - X_a y_{i+1} + X_a Y_a - x_{i+1} y_i - x_{i+1} Y_a + y_i X_a - X_a Y_a) + \tfrac{1}{2} * (x_i \eta_{i+1}$$
$$+ y_{i+1}\epsilon_i + \epsilon_i \eta_{i+1} - Y_a \epsilon_i - X_a \eta_{i+1} - x_{i+1}\eta_i - y_i \epsilon_{i+1} - \epsilon_{i+1}\eta_i + Y_a \epsilon_{i+1} + X_a \eta_i\Big)$$

Substituting $a_i = \tfrac{1}{2} * \Big((x_i - X_a)(y_{i+1} - Y_a) - (x_{i+1} - X_a)(y_i - Y_a)\Big)$, we get

$$A_i = a_i + \tfrac{1}{2} * \Big(x_i \eta_{i+1} + y_{i+1}\epsilon_i + \epsilon_i \eta_{i+1} - Y_a \epsilon_i - X_a \eta_{i+1} - x_{i+1}\eta_i$$
$$- y_i \epsilon_{i+1} - \epsilon_{i+1}\eta_i + Y_a \epsilon_{i+1} + X_a \eta_i\Big) \qquad (4.22)$$

Now, define:

s_1= the second term in (4.22)

$$= \tfrac{1}{2} * \Big(x_i \eta_{i+1} + y_{i+1}\epsilon_i + \epsilon_i \eta_{i+1} - Y_a \epsilon_i - X_a \eta_{i+1} - x_{i+1}\eta_i - y_i \epsilon_{i+1}$$
$$- \epsilon_{i+1}\eta_i + Y_a \epsilon_{i+1} + X_a \eta_i\Big)$$

Then: $E(A_i) = E(a_i + s_1) = a_i + E(s_1)$

And: $E(s_1) = \tfrac{1}{2} * E\Big(x_i \eta_{i+1} + y_{i+1}\epsilon_i + \epsilon_i \eta_{i+1} - Y_a \epsilon_i - X_a \eta_{i+1} - x_{i+1}\eta_i - y_i \epsilon_{i+1}$
$$- \epsilon_{i+1}\eta_i + Y_a \epsilon_{i+1} + X_a \eta_i\Big)$$
$$= \tfrac{1}{2} * \Big(x_i E(\eta_{i+1}) + y_{i+1} E(\epsilon_i) + E(\epsilon_i \eta_{i+1}) - Y_a E(\epsilon_i) - X_a E(\eta_{i+1}) - x_{i+1} E(\eta_i)$$
$$- y_i E(\epsilon_{i+1}) - E(\epsilon_{i+1}\eta_i) + Y_a E(\epsilon_{i+1}) + X_a E(\eta_i)\Big)$$
$$= 0$$

So $E(A_i) = a_i$

For B_{i-1} we use:

$$B_{i-1} = \tfrac{1}{2} * \Big((X_i - X_b)(Y_{i-1} - Y_b) - (X_{i-1} - X_b)(Y_i - Y_b)\Big)$$
$$B_{i-1} = \tfrac{1}{2} * \Big((x_i + \epsilon_i - X_b)(y_{i-1} + \eta_{i-1} - Y_b) - (x_{i-1} + \epsilon_{i-1} - X_b)(y_i + \eta_i - Y_b)\Big)$$
$$= \tfrac{1}{2} * \Big(x_i y_{i-1} + x_i \eta_{i-1} - x_i Y_b + y_{i-1}\epsilon_i + \epsilon_i \eta_{i-1} - Y_b \epsilon_i - X_b y_{i-1} - X_b \eta_{i-1} + X_b Y_b$$
$$- x_{i-1} y_i - x_{i-1}\eta_i + x_{i-1} Y_b - y_i \epsilon_{i-1} - \epsilon_{i-1}\eta_i + Y_b \epsilon_{i-1} + X_b y_i + X_b \eta_i - X_b Y_b\Big)$$

Rearranging terms:

$$B_{i-1} = \tfrac{1}{2}*(x_iy_{i-1} - x_iY_b - X_by_{i-1} + X_bY_b - x_{i-1}y_i + x_{i-1}Y_b + y_iX_b - X_bY_b) + \tfrac{1}{2}*(x_i\eta_{i-1}$$
$$+ y_{i-1}\epsilon_i + \epsilon_i\eta_{i-1} - Y_b\epsilon_i - X_b\eta_{i-1} - x_{i-1}\eta_i - y_i\epsilon_{i-1} - \epsilon_{i-1}\eta_i + Y_b\epsilon_{i-1} + X_b\eta_i)$$

And substituting $b_{i-1} = \frac{1}{2} * \Big((x_i - X_b)(y_{i-1} - Y_b) - (x_{i-1} - X_b)(y_i - Y_b)\Big)$, we get:

$$B_{i-1} = b_{i-1} + \tfrac{1}{2}*(x_i\eta_{i-1} + y_{i-1}\epsilon_i + \epsilon_i\eta_{i-1} - Y_b\epsilon_i - X_b\eta_{i-1} - x_{i-1}\eta_i - y_i\epsilon_{i-1} - \epsilon_{i-1}\eta_i$$
$$+ Y_b\epsilon_{i-1} + X_b\eta_i) \qquad (4.23)$$

Now, we will define:

s_2 = the second term in (4.23)

$$= \tfrac{1}{2}*\Big(x_i\eta_{i-1} + y_{i-1}\epsilon_i + \epsilon_i\eta_{i-1} - Y_b\epsilon_i - X_b\eta_{i-1} - x_{i-1}\eta_i - y_i\epsilon_{i-1}$$
$$- \epsilon_{i-1}\eta_i + Y_b\epsilon_{i-1} + X_b\eta_i\Big)$$

Then: $E(B_{i-1}) = E(b_{i-1} + s_2) = b_{i-1} + E(s_2)$

And: $E(s_2) = \frac{1}{2} * E\Big(x_i\eta_{i-1} + y_{i-1}\epsilon_i + \epsilon_i\eta_{i-1} - Y_b\epsilon_i - X_b\eta_{i-1} - x_{i-1}\eta_i - y_i\epsilon_{i-1}$
$$- \epsilon_{i-1}\eta_i + Y_b\epsilon_{i-1} + X_b\eta_i\Big)$$
$$= \tfrac{1}{2} * \Big(x_iE(\eta_{i-1}) + y_{i-1}E(\epsilon_i) + E(\epsilon_i\eta_{i-1}) - Y_bE(\epsilon_i) - X_bE(\eta_{i-1}) - x_{i-1}E(\eta_i)$$
$$- y_iE(\epsilon_{i-1}) - E(\epsilon_{i-1}\eta_i) + Y_bE(\epsilon_{i-1}) + X_bE(\eta_i)\Big)$$
$$= 0$$

So: $E(B_{i-1}) = b_{i-1}$

Then, $A_iB_{i-1} = a_ib_{i-1} + a_is_2 + b_{i-1}s_1 + s_1s_2$

and $E(A_iB_{i-1}) = a_ib_{i-1} + a_iE(s_2) + b_{i-1}E(s_1) + E(s_1s_2) = a_ib_{i-1} + E(s_1s_2)$ (4.24)

Now, we substitute (4.24) into (4.21) and get:

$$Cov(A_i,B_{i-1}) = E(A_iB_{i-1}) - E(A_i)E(B_{i-1}) = a_ib_{i-1} + E(s_1s_2) - a_ib_{i-1} = E(s_1s_2)$$

So, solving $E(s_1s_2)$:

$$E(s_1s_2) = \tfrac{1}{4} * E\Bigg(\Big(x_i\eta_{i+1} + y_{i+1}\epsilon_i + \epsilon_i\eta_{i+1} - Y_a\epsilon_i - X_a\eta_{i+1} - x_{i+1}\eta_i - y_i\epsilon_{i+1}$$
$$- \epsilon_{i+1}\eta_i + Y_a\epsilon_{i+1} + X_a\eta_i\Big)*\Big(x_i\eta_{i-1} + y_{i-1}\epsilon_i + \epsilon_i\eta_{i-1} - Y_b\epsilon_i - X_b\eta_{i-1}$$
$$- x_{i-1}\eta_i - y_i\epsilon_{i-1} - \epsilon_{i-1}\eta_i + Y_b\epsilon_{i-1} + X_b\eta_i\Big)\Bigg) \qquad (4.25)$$

Equation (4.25) expands to 100 terms. When passing the expectation through, the only terms

which remain are those involving the following products, whose expectation is non-zero:

$$\epsilon_i\epsilon_{i+1}, \quad \epsilon_i\epsilon_{i-1}, \quad \eta_i\eta_{i+1}, \quad \eta_i\eta_{i-1},$$
$$\epsilon_i^2, \quad \epsilon_{i-1}^2, \quad \epsilon_{i+1}^2,$$
$$\eta_i^2, \quad \eta_{i-1}^2, \quad \eta_{i+1}^2.$$

Any term involving the product of three of the random variables (ϵ_i's or η_i's) will be zero. $\Big($For example, $E(\epsilon_i\epsilon_{i+1}\eta_i) = E(\epsilon_i\epsilon_{i+1})E(\eta_i) = (\sigma_i\sigma_{i+1}\rho_i)(0) = 0\Big)$. Thus, (4.25) reduces to:

$$\begin{aligned}E(s_1s_2) = \tfrac{1}{4} * \Big(&-x_{i-1}x_iE(\eta_i\eta_{i+1}) + X_bx_iE(\eta_i\eta_{i+1}) + y_{i-1}y_{i+1}E(\epsilon_i^2) - Y_by_{i+1}E(\epsilon_i^2)\\ &- y_iy_{i+1}E(\epsilon_{i-1}\epsilon_i) + Y_by_{i+1}E(\epsilon_{i-1}\epsilon_i) - E(\epsilon_{i-1}\epsilon_i\eta_i\eta_{i+1}) - Y_ay_{i-1}E(\epsilon_i^2)\\ &+ Y_aY_bE(\epsilon_i^2) + Y_ay_iE(\epsilon_i\epsilon_{i-1}) - Y_aY_bE(\epsilon_{i-1}\epsilon_i) + X_ax_{i-1}E(\eta_i\eta_{i+1})\\ &- X_aX_bE(\eta_i\eta_{i+1}) - x_ix_{i+1}E(\eta_{i-1}\eta_i) + X_bx_{i+1}E(\eta_{i-1}\eta_i) + x_{i-1}x_{i+1}E(\eta_i^2)\\ &- X_bx_{i+1}E(\eta_i^2) - y_{i-1}y_iE(\epsilon_i\epsilon_{i+1}) + Y_by_iE(\epsilon_i\epsilon_{i+1}) - E(\epsilon_i\epsilon_{i+1}\eta_{i-1}\eta_i)\\ &+ Y_ay_{i-1}E(\epsilon_i\epsilon_{i+1}) - Y_aY_bE(\epsilon_i\epsilon_{i+1}) + X_ax_iE(\eta_i\eta_{i-1}) - X_aX_bE(\eta_{i-1}\eta_i)\\ &- X_ax_{i-1}E(\eta_i^2) + X_aX_bE(\eta_i^2)\Big)\end{aligned}$$

And taking the expectations:

$$\begin{aligned}E(s_1s_2) = \tfrac{1}{4} * \Big(&-x_{i-1}x_i\sigma_i\sigma_{i+1}\rho_i + X_bx_i\sigma_i\sigma_{i+1}\rho_i + y_{i-1}y_{i+1}\sigma_i^2 - Y_by_{i+1}\sigma_i^2\\ &- y_iy_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + Y_by_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} - \sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i\\ &- Y_ay_{i-1}\sigma_i^2 + Y_aY_b\sigma_i^2 + Y_ay_i\sigma_{i-1}\sigma_i\rho_{i-1} - Y_aY_b\sigma_{i-1}\sigma_i\rho_{i-1}\\ &+ X_ax_{i-1}\sigma_i\sigma_{i+1}\rho_i - X_aX_b\sigma_i\sigma_{i+1}\rho_i - x_ix_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1}\\ &+ X_bx_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + x_{i-1}x_{i+1}\sigma_i^2 - X_bx_{i+1}\sigma_i^2 - y_{i-1}y_i\sigma_i\sigma_{i+1}\rho_i\\ &+ Y_by_i\sigma_i\sigma_{i+1}\rho_i - \sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i + Y_ay_{i-1}\sigma_i\sigma_{i+1}\rho_i - Y_aY_b\sigma_i\sigma_{i+1}\rho_i\\ &+ X_ax_i\sigma_{i-1}\sigma_i\rho_{i-1} - X_aX_b\sigma_{i-1}\sigma_i\rho_{i-1} - X_ax_{i-1}\sigma_i^2 + X_aX_b\sigma_i^2\Big)\end{aligned}$$

Combining terms, we have:

$$\begin{aligned}E(s_1s_2) = \tfrac{1}{4} * \Big(&\sigma_i^2(y_{i-1}y_{i+1} - Y_by_{i+1} - Y_ay_{i-1} + Y_aY_b)\\ &+\sigma_i^2(x_{i-1}x_{i+1} - X_bx_{i+1} - X_ax_{i-1} + X_aX_b)\\ &- \sigma_{i-1}\sigma_i\rho_{i-1}(y_iy_{i+1} - Y_by_{i+1} - Y_ay_i + Y_aY_b)\end{aligned}$$

$$- \sigma_{i-1}\sigma_i\rho_{i-1}(x_ix_{i+1} - X_bx_{i+1} - X_ax_i + X_aX_b)$$
$$- \sigma_i\sigma_{i+1}\rho_i(x_{i-1}x_i - X_bx_i - X_ax_{i-1} + X_aX_b)$$
$$- \sigma_i\sigma_{i+1}\rho_i(y_{i-1}y_i - Y_by_i - Y_ay_{i-1} + Y_aY_b)$$
$$-2\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i\Big)$$

Therefore,

$$\begin{aligned}\mathrm{Cov}(A_i,B_{i-1}) = \tfrac{1}{4} * \Big(& \sigma_i^2\Big((y_{i+1} - Y_a)(y_{i-1} - Y_b) + (x_{i+1} - X_a)(x_{i-1} - X_b)\Big) \\ & - \sigma_{i-1}\sigma_i\rho_{i-1}\Big((y_{i+1} - Y_a)(y_i - Y_b) + (x_{i+1} - X_a)(x_i - X_b)\Big) \\ & - \sigma_i\sigma_{i+1}\rho_i\Big((y_i - Y_a)(y_{i-1} - Y_b) + (x_i - X_a)(x_{i-1} - X_b)\Big) \\ & - 2\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i\Big)\end{aligned} \tag{4.26}$$

Case 2: $Cov(A_i,B_i)$

$$\text{By definition,}\quad \mathrm{Cov}(A_i,B_i) = E(A_iB_i) - E(A_i)E(B_i) \tag{4.27}$$

Equation (4.22) gives us an expression for A_i. To get B_i, we begin with:

$$\begin{aligned}B_i &= \tfrac{1}{2} * \Big((X_{i+1} - X_b)(Y_i - Y_b) - (X_i - X_b)(Y_{i+1} - Y_b)\Big) \\ &= \tfrac{1}{2} * \Big((x_{i+1} + \epsilon_{i+1} - X_b)(y_i + \eta_i - Y_b) - (x_i + \epsilon_i - X_b)(y_{i+1} + \eta_{i+1} - Y_b)\Big) \\ &= \tfrac{1}{2} * \Big(x_{i+1}y_i + x_{i+1}\eta_i - x_{i+1}Y_b + y_i\epsilon_{i+1} + \epsilon_{i+1}\eta_i - Y_b\epsilon_{i+1} - X_by_i - X_b\eta_i + X_bY_b \\ &\quad - x_iy_{i+1} - x_i\eta_{i+1} + x_iY_b - y_{i+1}\epsilon_i - \epsilon_i\eta_{i+1} + \epsilon_iY_b + X_by_{i+1} + X_b\eta_{i+1} - X_bY_b\Big)\end{aligned}$$

Rearranging terms,

$$\begin{aligned}B_i = \tfrac{1}{2}*(x_{i+1}y_i - x_{i+1}Y_b - X_by_i + X_bY_b - x_iy_{i+1} + x_iY_b + y_{i+1}X_b - X_bY_b) + \tfrac{1}{2}*(x_{i+1}\eta_i \\ + y_i\epsilon_{i+1} + \epsilon_{i+1}\eta_i - Y_b\epsilon_{i+1} - X_b\eta_i - x_i\eta_{i+1} - y_{i+1}\epsilon_i - \epsilon_i\eta_{i+1} + Y_b\epsilon_i + X_b\eta_{i+1})\end{aligned}$$

And substituting $\quad b_i = \tfrac{1}{2} * \Big((x_{i+1} - X_b)(y_i - Y_b) - (x_i - X_b)(y_{i+1} - Y_b)\Big)$,

We get:

$$\begin{aligned}B_i = b_i + \tfrac{1}{2}*(x_{i+1}\eta_i + y_i\epsilon_{i+1} + \epsilon_{i+1}\eta_i - Y_b\epsilon_{i+1} - X_b\eta_i - x_i\eta_{i+1} - y_{i+1}\epsilon_i - \epsilon_i\eta_{i+1} \\ + Y_b\epsilon_i + X_b\eta_{i+1})\end{aligned} \tag{4.28}$$

Now, let us define:

s_3 = the second term in (4.28)

$$= \tfrac{1}{2}*\Big(x_{i+1}\eta_i + y_i\epsilon_{i+1} + \epsilon_{i+1}\eta_i - Y_b\epsilon_{i+1} - X_b\eta_i - x_i\eta_{i+1} - y_{i+1}\epsilon_i - \epsilon_i\eta_{i+1} + Y_b\epsilon_i + X_b\eta_{i+1}\Big)$$

Then, $E(B_i) = E(b_i + s_3) = b_i + E(s_3)$

And
$$E(s_3) = \tfrac{1}{2} * E\Big(x_{i+1}\eta_i + y_i\epsilon_{i+1} + \epsilon_{i+1}\eta_i - Y_b\epsilon_{i+1} - X_b\eta_i - x_i\eta_{i+1} - y_{i+1}\epsilon_i - \epsilon_i\eta_{i+1} + Y_b\epsilon_i + X_b\eta_{i+1}\Big)$$

$$= \tfrac{1}{2} * \Big(x_{i+1}E(\eta_i) + y_iE(\epsilon_{i+1}) + E(\epsilon_{i+1}\eta_i) - Y_bE(\epsilon_{i+1}) - X_bE(\eta_i) - x_iE(\eta_{i+1}) - y_{i+1}E(\epsilon_i) - E(\epsilon_i\eta_{i+1}) + Y_bE(\epsilon_i) + X_bE(\eta_{i+1})\Big)$$

$$= 0$$

So $E(B_i) = b_i$

Then, $A_iB_i = a_ib_i + a_is_3 + b_is_1 + s_1s_3$

and
$$E(A_iB_i) = a_ib_i + a_iE(s_3) + b_iE(s_1) + E(s_1s_3) = a_ib_i + E(s_1s_3) \tag{4.29}$$

Now we substitute (4.29) into (4.27) and get:

$$Cov(A_i,B_i) = E(A_iB_i) - E(A_i)E(B_i) = a_ib_i + E(s_1s_3) - a_ib_i = E(s_1s_3)$$

And,
$$E(s_1s_3) = \tfrac{1}{4} * E\Big(\Big(x_i\eta_{i+1} + y_{i+1}\epsilon_i + \epsilon_i\eta_{i+1} - Y_a\epsilon_i - X_a\eta_{i+1} - x_{i+1}\eta_i - y_i\epsilon_{i+1} - \epsilon_{i+1}\eta_i + Y_a\epsilon_{i+1} + X_a\eta_i\Big)*\Big(x_{i+1}\eta_i + y_i\epsilon_{i+1} + \epsilon_{i+1}\eta_i - Y_b\epsilon_{i+1} - X_b\eta_i - x_i\eta_{i+1} - y_{i+1}\epsilon_i - \epsilon_i\eta_{i+1} + Y_b\epsilon_i + X_b\eta_{i+1}\Big)\Big) \tag{4.30}$$

Equation (4.30) expands to 100 terms. As before, passing the expectation through eliminates all terms except those whose expectation is non-zero:

$\epsilon_i\epsilon_{i+1}$, $\epsilon_i\epsilon_{i-1}$, $\eta_i\eta_{i+1}$, $\eta_i\eta_i$,

ϵ_i^2, ϵ_{i-1}^2, ϵ_{i+1}^2,

η_i^2, η_{i-1}^2, η_{i+1}^2.

Thus, (4.30) simplifies to:

$$E(s_1s_3) = \tfrac{1}{4} * \Big(x_ix_{i+1}E(\eta_i\eta_{i+1}) - X_bx_iE(\eta_i\eta_{i+1}) - X_i^2E(\eta_{i+1}^2) + X_bx_iE(\eta_{i+1}^2)$$

$$+ y_i y_{i+1} E(\epsilon_i \epsilon_{i+1}) - Y_b y_{i+1} E(\epsilon_i \epsilon_{i+1}) - y_{i+1}^2 E(\epsilon_i^2) + Y_b y_{i+1} E(\epsilon_i^2)$$
$$+ E(\epsilon_i \epsilon_{i+1} \eta_i \eta_{i+1}) - E(\epsilon_i^2 \eta_{i+1}^2) - Y_a y_i E(\epsilon_i \epsilon_{i+1}) + Y_a Y_b E(\epsilon_i \epsilon_{i+1})$$
$$+ Y_a y_{i+1} E(\epsilon_i^2) - Y_a Y_b E(\epsilon_i^2) - X_a x_{i+1} E(\eta_i \eta_{i+1}) + X_a X_b E(\eta_i \eta_{i+1})$$
$$+ X_a x_i E(\eta_{i+1}^2) - X_a X_b E(\eta_{i+1}^2) - X_{i+1}^2 E(\eta_i^2) + X_b x_{i+1} E(\eta_i^2)$$
$$+ x_i x_{i+1} E(\eta_i \eta_{i+1}) - X_b x_{i+1} E(\eta_i \eta_{i+1}) - y_i^2 E(\epsilon_{i+1}^2) + Y_b y_i E(\epsilon_{i+1}^2)$$
$$+ y_i y_{i+1} E(\epsilon_i \epsilon_{i+1}) - Y_b y_i E(\epsilon_i \epsilon_{i+1}) - E(\epsilon_{i+1}^2 \eta_i^2) + E(\epsilon_i \epsilon_{i+1} \eta_i \eta_{i+1})$$
$$+ Y_a y_i E(\epsilon_{i+1}^2) - Y_a Y_b E(\epsilon_{i+1}^2) - Y_a y_{i+1} E(\epsilon_i \epsilon_{i+1}) + Y_a Y_b E(\epsilon_i \epsilon_{i+1})$$
$$+ X_a x_{i+1} E(\eta_i^2) - X_a X_b E(\eta_i^2) - X_a x_i E(\eta_i \eta_{i+1}) + X_a X_b E(\eta_i \eta_{i+1}) \Big)$$

Taking the expectation:

$$E(s_1 s_3) = \tfrac{1}{4} * \Big(x_i x_{i+1} \sigma_i \sigma_{i+1} \rho_i - X_b x_i \sigma_i \sigma_{i+1} \rho_i - X_i^2 \sigma_{i+1}^2 + X_b x_i \sigma_{i+1}^2$$
$$+ y_i y_{i+1} \sigma_i \sigma_{i+1} \rho_i - Y_b y_{i+1} \sigma_i \sigma_{i+1} \rho_i - y_{i+1}^2 \sigma_i^2 + Y_b y_{i+1} \sigma_i^2$$
$$+ \sigma_i^2 \sigma_{i+1}^2 \rho_i^2 - \sigma_i^2 \sigma_{i+1}^2 - Y_a y_i \sigma_i \sigma_{i+1} \rho_i + Y_a Y_b \sigma_i \sigma_{i+1} \rho_i$$
$$+ Y_a y_{i+1} \sigma_i^2 - Y_a Y_b \sigma_i^2 - X_a x_{i+1} \sigma_i \sigma_{i+1} \rho_i + X_a X_b \sigma_i \sigma_{i+1} \rho_i$$
$$+ X_a x_i \sigma_{i+1}^2 - X_a X_b \sigma_{i+1}^2 - X_{i+1}^2 \sigma_i^2 + X_b x_{i+1} \sigma_i^2$$
$$+ x_i x_{i+1} \sigma_i \sigma_{i+1} \rho_i - X_b x_{i+1} \sigma_i \sigma_{i+1} \rho_i - y_i^2 \sigma_{i+1}^2 + Y_b y_i \sigma_{i+1}^2$$
$$+ y_i y_{i+1} \sigma_i \sigma_{i+1} \rho_i - Y_b y_i \sigma_i \sigma_{i+1} \rho_i - \sigma_{i+1}^2 \sigma_i^2 + \sigma_i^2 \sigma_{i+1}^2 \rho_i^2$$
$$+ Y_a y_i \sigma_{i+1}^2 - Y_a Y_b \sigma_{i+1}^2 - Y_a y_{i+1} \sigma_i \sigma_{i+1} \rho_i + Y_a Y_b \sigma_i \sigma_{i+1} \rho_i$$
$$+ X_a x_{i+1} \sigma_i^2 - X_a X_b \sigma_i^2 - X_a x_i \sigma_i \sigma_{i+1} \rho_i + X_a X_b \sigma_i \sigma_{i+1} \rho_i \Big)$$

Combining terms, we have:

$$E(s_1 s_3) = \tfrac{1}{4} * \Big(\sigma_i^2 (- y_{i+1}^2 + Y_b y_{i+1} + Y_a y_{i+1} - Y_a Y_b - X_{i+1}^2 + X_b x_{i+1} + X_a x_{i+1} - X_a X_b)$$
$$+ \sigma_{i+1}^2 (-X_i^2 + X_b x_i + X_a x_i - X_a X_b - y_i^2 + Y_b y_i + Y_a y_i - Y_a Y_b)$$
$$+ \sigma_i \sigma_{i+1} \rho_i (x_i x_{i+1} - X_b x_i - X_a x_{i+1} + X_a X_b + x_i x_{i+1} - X_b x_{i+1} - X_a x_i + X_a X_b)$$
$$+ \sigma_i \sigma_{i+1} \rho_i (y_i y_{i+1} - Y_b y_{i+1} - Y_a y_i + Y_a Y_b + y_i y_{i+1} - Y_b y_i - Y_a y_{i+1} + Y_a Y_b)$$
$$- \sigma_{i+1}^2 \sigma_i^2 + \sigma_i^2 \sigma_{i+1}^2 \rho_i^2 + \sigma_i^2 \sigma_{i+1}^2 \rho_i^2 - \sigma_i^2 \sigma_{i+1}^2 \Big)$$

And simplifying, this becomes:

$$\begin{aligned}\mathrm{Cov}(A_i,B_i) = \tfrac{1}{4} * \Big(& -\sigma_i^2\Big((y_{i+1} - Y_b)(y_{i+1} - Y_a) + (x_{i+1} - X_b)(x_{i+1} - X_a)\Big) \\ & - \sigma_{i+1}^2\Big((x_i - X_b)(x_i - X_a) + (y_i - Y_b)(y_i - Y_a)\Big) \\ & + \sigma_i\sigma_{i+1}\rho_i\Big((x_i - X_a)(x_{i+1}- X_b) + (x_i - X_b)(x_{i+1} - X_a)\Big) \\ & + \sigma_i\sigma_{i+1}\rho_i\Big((y_i- Y_b)(y_{i+1}- Y_a) + (y_{i+1} - Y_b)(y_i - Y_a)\Big) \\ & - 2\sigma_{i+1}^2\sigma_i^2(1 - \rho_i^2)\Big)\end{aligned} \tag{4.31}$$

Case 3: $Cov(A_i,B_{i+1})$

By definition, $\mathrm{Cov}(A_i,B_{i+1}) = E(A_iB_{i+1}) - E(A_i)E(B_{i+1})$ (4.32)

Again, equation (4.22) provides an expression for A_i. To get B_{i+1} we use:

$$\begin{aligned}B_{i+1} &= \tfrac{1}{2} * \Big((X_{i+2} - X_b)(Y_{i+1}- Y_b) - (X_{i+1} - X_b)(Y_{i+2} - Y_b)\Big) \\ &= \tfrac{1}{2} * \Big((x_{i+2} + \epsilon_{i+2} - X_b)(y_{i+1} + \eta_{i+1} - Y_b) - (x_{i+1} + \epsilon_{i+1} - X_b)(y_{i+2} + \eta_{i+2} - Y_b)\Big) \\ &= \tfrac{1}{2} * \Big(x_{i+2}y_{i+1} + x_{i+2}\eta_{i+1} - x_{i+2}Y_b + y_{i+1}\epsilon_{i+2} + \epsilon_{i+2}\eta_{i+1} - Y_b\epsilon_{i+2} - X_by_{i+1} \\ &\qquad - X_b\eta_{i+1} + X_bY_b - x_{i+1}y_{i+2} - x_{i+1}\eta_{i+2} + x_{i+1}Y_b - y_{i+2}\epsilon_{i+1} - \epsilon_{i+1}\eta_{i+2} \\ &\qquad + Y_b\epsilon_{i+1} + X_by_{i+2} + X_b\eta_{i+2} - X_bY_b\Big)\end{aligned}$$

Rearranging terms, we get:

$$\begin{aligned}B_{i+1} = {} & \tfrac{1}{2}*(x_{i+2}y_{i+1} - x_{i+2}Y_b - X_by_{i+1} + X_bY_b - x_{i+1}y_{i+2} + x_{i+1}Y_b + y_{i+2}X_b - X_bY_b) \\ & + \tfrac{1}{2}*(x_{i+2}\eta_{i+1} + y_{i+1}\epsilon_{i+2} + \epsilon_{i+2}\eta_{i+1} - Y_b\epsilon_{i+2} - X_b\eta_{i+1} - x_{i+1}\eta_{i+2} \\ & - y_{i+2}\epsilon_{i+1} - \epsilon_{i+1}\eta_{i+2} + Y_b\epsilon_{i+1} + X_b\eta_{i+2})\end{aligned}$$

And substituting $b_{i+1} = \tfrac{1}{2} * \Big((x_{i+2} - X_b)(y_{i+1} - Y_b) - (x_{i+1} - X_b)(y_{i+2} - Y_b)\Big)$,

We have

$$\begin{aligned}B_{i+1} = b_{i+1} + {} & \tfrac{1}{2}*(x_{i+2}\eta_{i+1} + y_{i+1}\epsilon_{i+2} + \epsilon_{i+2}\eta_{i+1} - Y_b\epsilon_{i+2} - X_b\eta_{i+1} - x_{i+1}\eta_{i+2} \\ & - y_{i+2}\epsilon_{i+1} - \epsilon_{i+1}\eta_{i+2} + Y_b\epsilon_{i+1} + X_b\eta_{i+2})\end{aligned} \tag{4.33}$$

Now, let us define:

s_4 = the second term in (4.33):

$$= \tfrac{1}{2}*(x_{i+2}\eta_{i+1} + y_{i+1}\epsilon_{i+2} + \epsilon_{i+2}\eta_{i+1} - Y_b\epsilon_{i+2} - X_b\eta_{i+1} - x_{i+1}\eta_{i+2} - y_{i+2}\epsilon_{i+1} - \epsilon_{i+1}\eta_{i+2} + Y_b\epsilon_{i+1} + X_b\eta_{i+2})$$

Then, $E(B_{i+1}) = E(b_{i+1} + s_4) = b_{i+1} + E(s_4)$

and
$$E(s_4) = \tfrac{1}{2} * E\Big(x_{i+2}\eta_{i+1} + y_{i+1}\epsilon_{i+2} + \epsilon_{i+2}\eta_{i+1} - Y_b\epsilon_{i+2} - X_b\eta_{i+1} - x_{i+1}\eta_{i+2} - y_{i+2}\epsilon_{i+1} - \epsilon_{i+1}\eta_{i+2} + Y_b\epsilon_{i+1} + X_b\eta_{i+2}\Big)$$

$$= \tfrac{1}{2} * \Big(x_{i+2}E(\eta_{i+1}) + y_{i+1}E(\epsilon_{i+2}) + E(\epsilon_{i+2}\eta_{i+1}) - Y_bE(\epsilon_{i+2}) - X_bE(\eta_{i+1}) - x_{i+1}E(\eta_{i+2}) - y_{i+2}E(\epsilon_{i+1}) - E(\epsilon_{i+1}\eta_{i+2}) + Y_bE(\epsilon_{i+1}) + X_bE(\eta_{i+2})\Big)$$

$$= 0$$

So $E(B_{i+1}) = b_{i+1}$

Then, $A_iB_{i+1} = a_ib_{i+1} + a_is_4 + b_{i+1}s_1 + s_1s_4$

and
$$E(A_iB_{i+1}) = a_ib_{i+1} + a_iE(s_4) + b_{i+1}E(s_1) + E(s_1s_4) = a_ib_{i+1} + E(s_1s_4) \tag{4.34}$$

Now, we substitute (4.34) into (4.32) and get:

$$Cov(A_i,B_{i+1}) = E(A_iB_{i+1}) - E(A_i)E(B_{i+1}) = a_ib_{i+1} + E(s_1s_4) - a_ib_{i+1} = E(s_1s_4)$$

And,
$$E(s_1s_4) = \tfrac{1}{4} * E\Big(\Big(x_i\eta_{i+1} + y_{i+1}\epsilon_i + \epsilon_i\eta_{i+1} - Y_a\epsilon_i - X_a\eta_{i+1} - x_{i+1}\eta_i - y_i\epsilon_{i+1} - \epsilon_{i+1}\eta_i + Y_a\epsilon_{i+1} + X_a\eta_i\Big)*\Big(x_{i+2}\eta_{i+1} + y_{i+1}\epsilon_{i+2} + \epsilon_{i+2}\eta_{i+1} - Y_b\epsilon_{i+2} - X_b\eta_{i+1} - x_{i+1}\eta_{i+2} - y_{i+2}\epsilon_{i+1} - \epsilon_{i+1}\eta_{i+2} + Y_b\epsilon_{i+1} + X_b\eta_{i+2}\Big)\Big) \tag{4.35}$$

Equation (4.35) expands to 100 terms. Passing the expectation through as before leaves only the terms which involving the following products, whose expectation is non-zero:

$\epsilon_i\epsilon_{i+1}$, $\epsilon_i\epsilon_{i-1}$, $\eta_i\eta_{i+1}$, $\eta_i\eta_{i-1}$,

ϵ_i^2, ϵ_{i-1}^2, ϵ_{i+1}^2,

η_i^2, η_{i-1}^2, η_{i+1}^2.

Any term involving the product of three of the random variables (ϵ_i's or η_i's) will be zero. Thus, (4.35) simplifies to:

$$E(s_1s_4) = \tfrac{1}{4} * \Big(x_ix_{i+2}E(\eta_{i+1}^2) - x_iX_bE(\eta_{i+1}^2) - x_ix_{i+1}E(\eta_{i+1}\eta_{i+2}) + x_iX_bE(\eta_{i+1}\eta_{i+2})$$

$$
\begin{aligned}
&- y_{i+1}y_{i+2}E(\epsilon_i\epsilon_{i+1}) + y_{i+1}Y_bE(\epsilon_i\epsilon_{i+1}) - E(\epsilon_i\epsilon_{i+1}\eta_{i+1}\eta_{i+2}) \\
&+ Y_ay_{i+2}E(\epsilon_i\epsilon_{i+1}) - Y_aY_bE(\epsilon_i\epsilon_{i+1}) - X_ax_{i+2}E(\eta_{i+1}^2) + X_aX_bE(\eta_{i+1}^2) \\
&+ X_ax_{i+1}E(\eta_{i+1}\eta_{i+2}) - X_aX_bE(\eta_{i+1}\eta_{i+2}) - x_{i+1}x_{i+2}E(\eta_i\eta_{i+1}) \\
&+ x_{i+1}X_bE(\eta_i\eta_{i+1}) - y_iy_{i+1}E(\epsilon_{i+1}\epsilon_{i+2}) + y_iY_bE(\epsilon_{i+1}\epsilon_{i+2}) \\
&+ y_iy_{i+2}E(\epsilon_{i+1}^2) - y_iY_bE(\epsilon_{i+1}^2) - E(\epsilon_{i+1}\epsilon_{i+2}\eta_i\eta_{i+1}) + Y_ay_{i+1}E(\epsilon_{i+1}\epsilon_{i+2}) \\
&- Y_aY_bE(\epsilon_{i+1}\epsilon_{i+2}) - Y_ay_{i+2}E(\epsilon_{i+1}^2) + Y_aY_bE(\epsilon_{i+1}^2) + X_ax_{i+2}E(\eta_i\eta_{i+1}) \\
&- X_aX_bE(\eta_i\eta_{i+1})\Big)
\end{aligned}
$$

Which is:

$$
\begin{aligned}
E(s_1s_4) = \tfrac{1}{4} * \Big(&x_ix_{i+2}\sigma_{i+1}^2 - x_iX_b\sigma_{i+1}^2 - x_ix_{i+1}\sigma_{i+1}\sigma_{i+2}\rho_{i+1} + x_iX_b\sigma_{i+1}\sigma_{i+2}\rho_{i+1} \\
&- y_{i+1}y_{i+2}\sigma_i\sigma_{i+1}\rho_i + y_{i+1}Y_b\sigma_i\sigma_{i+1}\rho_i - \sigma_i\sigma_{i+1}^2\sigma_{i+2}\rho_i\rho_{i+1} \\
&+ Y_ay_{i+2}\sigma_i\sigma_{i+1}\rho_i - Y_aY_b\sigma_i\sigma_{i+1}\rho_i - X_ax_{i+2}\sigma_{i+1}^2 + X_aX_b\sigma_{i+1}^2 \\
&+ X_ax_{i+1}\sigma_{i+1}\sigma_{i+2}\rho_{i+1} - X_aX_b\sigma_{i+1}\sigma_{i+2}\rho_{i+1} - x_{i+1}x_{i+2}\sigma_i\sigma_{i+1}\rho_i \\
&+ x_{i+1}X_b\sigma_i\sigma_{i+1}\rho_i - y_iy_{i+1}\sigma_{i+1}\sigma_{i+2}\rho_{i+1} + y_iY_b\sigma_{i+1}\sigma_{i+2}\rho_{i+1} \\
&+ y_iy_{i+2}\sigma_{i+1}^2 - y_iY_b\sigma_{i+1}^2 - \sigma_i\sigma_{i+1}^2\sigma_{i+2}\rho_i\rho_{i+1} + Y_ay_{i+1}\sigma_{i+1}\sigma_{i+2}\rho_{i+1} \\
&- Y_aY_b\sigma_{i+1}\sigma_{i+2}\rho_{i+1} - Y_ay_{i+2}\sigma_{i+1}^2 + Y_aY_b\sigma_{i+1}^2 + X_ax_{i+2}\sigma_i\sigma_{i+1}\rho_i \\
&- X_aX_b\sigma_i\sigma_{i+1}\rho_i\Big)
\end{aligned}
$$

Combining terms, we have:

$$
\begin{aligned}
E(s_1s_4) = \tfrac{1}{4} * \Big(&\sigma_{i+1}^2(x_ix_{i+2} - x_iX_b - X_ax_{i+2} + X_aX_b + y_iy_{i+2} - y_iY_b - Y_ay_{i+2} + Y_aY_b) \\
&-\sigma_i\sigma_{i+1}\rho_i(y_{i+1}y_{i+2} - y_{i+1}Y_b - Y_ay_{i+2} + Y_aY_b + x_{i+1}x_{i+2} - x_{i+1}X_b - X_ax_{i+2} + X_aX_b) \\
&-\sigma_{i+1}\sigma_{i+2}\rho_{i+1}(x_ix_{i+1} - x_iX_b - X_ax_{i+1} + X_aX_b + y_iy_{i+1} - y_iY_b - Y_ay_{i+1} + Y_aY_b) \\
&- 2\sigma_i\sigma_{i+1}^2\sigma_{i+2}\rho_i\rho_{i+1}\Big)
\end{aligned}
$$

Which simplifies to:

$$
\begin{aligned}
Cov(A_i,B_{i+1}) = \tfrac{1}{4} * \Big(&\sigma_{i+1}^2\Big((x_i - X_a)(x_{i+2} - X_b) + (y_i - Y_a)(y_{i+2} - Y_b)\Big) \\
&-\sigma_i\sigma_{i+1}\rho_i\Big((y_{i+1} - Y_a)(y_{i+2} - Y_b) + (x_{i+1} - X_a)(x_{i+2} - X_b)\Big) \\
&-\sigma_{i+1}\sigma_{i+2}\rho_{i+1}\Big((x_i - X_a)(x_{i+1} - X_b) + (y_i - Y_a)(y_{i+1} - Y_b)\Big)
\end{aligned}
$$

$$- 2\sigma_i\sigma_{i+1}^2\sigma_{i+2}\rho_i\rho_{i+1})\Big)\Big) \tag{4.36}$$

Combining Triangles Along an Arc

We have developed expressions for the three cases of covariance between a triangle in one polygon and the connected triangles in an adjacent polygon. The next step is to sum the covariances for all triangles formed by an arc. Assume that an arc which separates polygons A and B has $m+1$ points. There will be m triangles in the arc-sector in polygon A (A_i, i=1..m) and m triangles in the arc-sector in polygon B (B_j, j=1...m). Then, the covariance between polygons A and B is the sum of all appropriate triangle covariances:

$$\begin{aligned}
\mathrm{Cov}(A,B) = {} & \mathrm{Cov}(A_1,B_1) + \mathrm{Cov}(A_1,B_2) + \\
& \mathrm{Cov}(A_2,B_1) + \mathrm{Cov}(A_2,B_2) + \mathrm{Cov}(A_2,B_3) + \\
& \mathrm{Cov}(A_3,B_2) + \mathrm{Cov}(A_3,B_3) + \mathrm{Cov}(A_3,B_4) + \\
& \quad \cdots \qquad \cdots \qquad \cdots \\
& \mathrm{Cov}(A_i,B_{i-1}) + \mathrm{Cov}(A_i,B_i) + \mathrm{Cov}(A_i,B_{i+1}) + \\
& \quad \cdots \qquad \cdots \qquad \cdots \\
& \mathrm{Cov}(A_{m-1},B_{m-2}) + \mathrm{Cov}(A_{m-1},B_{m-1}) + \mathrm{Cov}(A_{m-1},B_m) + \\
& \mathrm{Cov}(A_m,B_{m-1}) + \mathrm{Cov}(A_m,B_m).
\end{aligned}$$

or, if we define $\mathrm{Cov}(A_1,B_0) = 0$ and $\mathrm{Cov}(A_m,B_{m+1}) = 0$, we can use the summation:

$$\mathrm{Cov}(A,B) = \sum_{i=1}^{m}\Big(\mathrm{Cov}(A_i,B_{i-1}) + \mathrm{Cov}(A_i,B_i) + \mathrm{Cov}(A_i,B_{i+1})\Big)$$

Now, we substitute the expressions we have derived for these individual triangle covariances (4.26, 4.31, and 4.36), and obtain:

$$\begin{aligned}
\mathrm{Cov}(A,B) = \tfrac{1}{4} * \sum_{i=1}^{m}\Big(\sigma_i^2(\tilde{x}_{a_{i+1}}\tilde{x}_{b_{i-1}} + \tilde{y}_{a_{i+1}}\tilde{y}_{b_{i-1}}) - \sigma_{i-1}\sigma_i\rho_i(\tilde{x}_{a_{i+1}}\tilde{x}_{b_i} + \tilde{y}_{a_{i+1}}\tilde{y}_{b_i}) \\
- \sigma_i\sigma_{i+1}\rho_{i+1}(\tilde{x}_{a_i}\tilde{x}_{b_{i-1}} + \tilde{y}_{a_i}\tilde{y}_{b_{i-1}}) - 2\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_i\rho_{i+1}\Big) -
\end{aligned}$$

$$\Big(\sigma_i^2(\tilde{x}_{a_{i+1}}\tilde{x}_{b_{i+1}} + \tilde{y}_{a_{i+1}}\tilde{y}_{b_{i+1}}) + \sigma_{i+1}^2(\tilde{x}_{a_i}\tilde{x}_{b_i} + \tilde{y}_{a_i}\tilde{y}_{b_i}) -$$
$$\sigma_i\sigma_{i+1}\rho_i(\tilde{x}_{a_i}\tilde{x}_{b_{i+1}} + \tilde{y}_{a_i}\tilde{y}_{b_{i+1}}) - \sigma_i\sigma_{i+1}\rho_i(\tilde{x}_{a_{i+1}}\tilde{x}_{b_i} + \tilde{y}_{a_{i+1}}\tilde{y}_{b_i}) +$$
$$2\sigma_i^2\sigma_{i+1}^2(1-\rho_i^2)\Big) + \Big(\sigma_{i+1}^2(\tilde{x}_{a_i}\tilde{x}_{bi+2} + \tilde{y}_{a_i}\tilde{y}_{bi+2}) - \sigma_i\sigma_{i+1}\rho_i(\tilde{x}_{a_{i+1}}\tilde{x}_{bi+2} +$$
$$\tilde{y}_{a_i}\tilde{y}_{bi+2}) - \sigma_{i+1}\sigma_{i+2}\rho_{i+1}(\tilde{x}_{a_i}\tilde{x}_{b_{i+1}} + \tilde{y}_{a_i}\tilde{y}_{b_{i+1}}) - 2\sigma_i\sigma_{i+1}^2\sigma_{i+2}\rho_i\rho_{i+1}\Big)\Big)$$

$$= \tfrac{1}{4}*\sum_{i=1}^{m}\Bigg(\sigma_i^2\Big(\tilde{x}_{a_{i+1}}(\tilde{x}_{b_{i-1}} - \tilde{x}_{b_{i+1}}) + \tilde{y}_{a_{i+1}}(\tilde{y}_{b_{i-1}} - \tilde{y}_{b_{i+1}})\Big)$$
$$+ \sigma_{i+1}^2\Big(\tilde{x}_{a_i}(\tilde{x}_{bi+2} - \tilde{x}_{b_i}) + \tilde{y}_{a_i}(\tilde{y}_{bi+2} - \tilde{y}_{b_i})\Big)$$
$$- s_{i-1}(\tilde{x}_{a_{i+1}}\tilde{x}_{b_i} + \tilde{y}_{a_{i+1}}\tilde{y}_{b_i}) - s_{i+1}(\tilde{x}_{a_i}\tilde{x}_{b_{i+1}} + \tilde{y}_{a_i}\tilde{y}_{b_{i+1}})$$
$$- s_i\Big(\tilde{x}_{a_i}(\tilde{x}_{b_{i-1}} - \tilde{x}_{b_{i+1}}) + \tilde{y}_{a_i}(\tilde{y}_{b_{i-1}} - \tilde{y}_{b_{i+1}})\Big)$$
$$- s_i\Big(\tilde{x}_{a_{i+1}}(\tilde{x}_{bi+2} - \tilde{x}_{b_i}) + \tilde{y}_{a_{i+1}}(\tilde{y}_{bi+2} - \tilde{y}_{b_i})\Big)\Bigg)$$
$$- \tfrac{1}{2}*\sum_{i=1}^{m}\Big(s_i(s_{i-1} - s_i + s_{i+1}) + \sigma_i^2\sigma_{i+1}^2\Big) \qquad (4.37)$$

where:

$$\tilde{x}_{a_{i-1}} = x_{i-1} - X_a \qquad \tilde{x}_{b_{i-1}} = x_{i-1} - X_b$$
$$\tilde{x}_{a_i} = x_i - X_a \qquad \tilde{x}_{b_i} = x_i - X_b$$
$$\tilde{y}_{a_{i-1}} = y_{i-1} - Y_a \qquad \tilde{y}_{b_{i-1}} = y_{i-1} - Y_b$$
$$\tilde{y}_{a_i} = y_i - Y_a \qquad \tilde{y}_{b_i} = y_i - Y_b$$

etc...

and where s_i is as in (4.15).

DISCUSSION - DETERMINING MODEL PARAMETERS

The expressions for polygon variance and covariance depend upon the coordinates and the σ's and ρ's which indicate the variability and correlation of points in an arc. Several possibilites exist for selecting values of σ. In Chrisman's (1982c) work in this area, close

examination of the steps involved in producing a USGS land cover map (scale 1:250,000) provided deductive estimates of individual error components, which were combined to arrive at an overall estimate of positional accuracy. Chrisman considered three components: line width, digitizing (by scanning), and rounding errors, under two assumptions he termed "conservative" and "less conservative". For example, under conservative assumptions, the standard deviation of errors caused by line width was assumed to be 14.4 meters, that of scanning errors was assumed to be 16.6 meters, and rounding errors exhibited a standard deviation of 5.8 meters. Adding the variances of these errors resulted in an overall standard deviation of 22.8 meters. This approach for estimating overall positional error has been endorsed in the Proposed Digital Cartographic Data Standard (Morrison, 1988).

An alternative for estimating the variance parameter is based upon map accuracy statements. As discussed by Keefer (1988), typical map accuracy statements include a band width and an alpha level. For example, the National Map Accuracy Standard quoted earlier states:

> "For maps on publication scales larger than 1:20,000, not more than 10% of the points tested shall be in error by more than 1/30th inch, measured on the publication scale..."

In the above statement, the error-band half-width is 1/30th inch, and the alpha level is 0.10. Keefer (1988) describes how these statements may be used to select a variance for a normal distribution that meets the specification. The estimate for σ is denoted by:

$$\sigma = \frac{W}{K}$$

where:

W = $\frac{1}{2}$ the error band width

K = 1-$\frac{\alpha}{2}$ quantile of the standard normal distribution

A more costly alternative procedure is to evaluate the variability of points in repeated sampling involving mapping the same area a number of times. In this case, a point coordinate represents a random variable which is observed in repeated samples and for which the standard deviation can be calculated.

In some instances, obtaining an estimate of arc variability may be straightforward. For example, in many GIS applications, polygons are formed by *proximity analysis*, in which a computer algorithm delineates those areas within a given distance of some feature. A typical case in forestry GIS is the creation of road polygons to extract acreage from the timber stands through which a road passes (thereby obtaining *net forested acres*). Another example is creation of "buffer strips" around streams or drainages. In these cases, the arcs surrounding the feature of interest (e.g., the road or stream) are established by the GIS at a fixed distance from that feature. However, on the ground, the width of the road or buffer strip is likely to vary somewhat. If we measure the variability of width via sampling, we could assign one-half of the standard deviation of width to the arcs on each side of the road or stream.

The choice of the correlation coefficent, ρ, may be more difficult. In an analysis of digitizing errors, Keefer *et al.* (1988) used time series analysis to detect serial correlation of errors. After fitting an autoregressive model to data digitized in stream mode, he encountered correlation coefficients between 0.3 and 0.9, with an average of about 0.7. It is likely that mapping processes such as digitizing technique will have a significant impact on the serial correlation of coordinate errors, but more study is obviously needed to obtain reliable estimates of the correlation coefficient under varying conditions. Even without firm knowledge of the

amount of correlation of errors in a map, the model may be quite useful: one advantage of an algorithm for calculating polygon variance is the capacity for performing sensitivity analyses in order to determine the impact of different levels of correlation on the resulting variance.

The derivations performed to obtain an expression for polygon area variance allowed a different σ and ρ to be assigned to each coordinate pair in the map. A more realistic assumption might be that all points in an arc exhibit the same variance and correlation with their neighbors. In some cases, a single σ and ρ may even suffice for all arcs in a map. Often, however, knowledge about the ability to locate various boundaries may suggest the use of different parameters for different arcs. For example, in maps derived from interpretation of color-infrared photographs, some boundaries (such as those between water and land) may be discernible with much greater precision than other boundaries (such as those between vegetation types with similar spectral signatures). In these instances, assumption of different σ's for the different arcs may be justified. In any case, it is logical to consider the values for σ and ρ to be attributes of an arc, and to be maintained as such when overlaying polygons. Then, the values may be made available to computer programs which could calculate polygon variances for the resulting overlay map. When all points in an arc are assumed to have the same parameters (σ and ρ), simplification of the variance and covariance equations is possible, and an opportunity for computational efficiencies arises. Thus, the equations for polygon area variance (4.15) and covariance (4.37) as written *are not* in the most efficient computational form. Matrix notation may also be used to express the variance and covariance formulas, but calculation of these quantities by matrix algebra may not be more efficient than use of simple sums.

One issue which is encountered in implementation of the variance and covariance expressions is the selection of an estimate of σ for nodes. If the arcs which are incident at a

node have different assumed variabilities, there may not be a single obvious choice for which variability to assign to the node in area variance calculations. It would seem reasonable to assign the variance of the points in the *least variable* incident arc to a node as the variability of that node. For example, if a node is at the intersection of an ownership boundary, a timber stand boundary, and a soils unit boundary, the location of the node is known to be on the ownership boundary, and hence is known at the highest of the three precisions. This is especially true when data sets are digitized and overlaid with a priority given to the most precise layers, which is possible when GIS software allows a user to "snap" an arc to an existing (and usually more precise) arc or node.

A final consideration in the identification of point variability is the average distance between points along an arc. The simulation programs which were used to test the equations derived above performed very poorly when σ was large relative to the distance between points in an arc. When σ approached values of $\frac{1}{4}$ to $\frac{1}{3}$ of the distance between points, simulated errors occasionally caused points on an arc to *reverse order*, creating "loops" in the arc. The lesson to be learned from this is that it may be unrealistic to assign a value of σ which does not reflect the distance between digitized points. Actually, the inconsistency lies in digitizing points at a higher resolution than the map and data call for. However, when analyzing data that has already been digitized, it makes sense to use a σ which is compatible with the density of the data.

POLYGON AREA ERRORS - VALIDATION OF THE DISTRIBUTION

The suitability of the normal distribution to model polygon area errors was examined in

a limited test involving simulations of a few contrived polygons. It was suggested that areas of polygons with few vertices would be less likely the normal distribution than areas of polygons with more vertices. To test this, the Anderson-Darling A^2 statistic was calculated for each of twenty simulations (80 iterations each) of the six polygons created from sampling a circle. These values are tabulated in Table 3. A null hypothesis of normal errors (with mean and variance specified by the expressions previously developed) at α=0.10 would be rejected for a sample of size 80 if $A^2 > 1.933$. Even for a polygon of only three vertices, the null hypothesis was rejected in only 15% of the simulations. Even data *known* to be from a normal distribution may be expected to be rejected an average of two times out of 20 when tested at $\alpha = 0.10$. Though inconclusive, there appeared to be a trend toward fewer rejections of a hypothesis of normality as the number of vertices in the polygon increased. While this was only a limited evaluation of errors in single artificial polygons, it appears that an assumption of normality for area errors seems appropriate. Simulation of different sizes and shapes of sliver polygons, at varying values for σ and ρ, would be desirable before making inferences using the normal distribution to model areas of sliver polygons.

POLYGON AREA ERRORS - EXAMPLE APPLICATION

Polygon area variances and covariances were calculated for the Webster tract using the nine sets of assumptions depicted in Table 2. These variables, and the per-acre values from Table 1, were entered into equation 3.5 to obtain nine estimates for variance of total tract value. Coefficients of variation of total tract value were calculated for each of the nine assumptions, and are given in Table 4. Figure 13 shows the coefficent of variation as a function of the values used for σ and ρ. It is readily apparent that increasing σ or ρ will increase variability in a

Table 3. Anderson-Darling A^2 statistics[1] for 20 simulations each of six polygons with varying numbers of vertices.

Number of Vertices in Polygon					
N=3	N=5	N=7	N=9	N=11	N=15
		2.31608	Reject H_0		
3.22710	4.56917	2.24194			
2.76724	2.36631	1.98364	2.88770		
2.03021	1.94432	1.95903	2.12643	2.41514	1.96851
1.73295	1.32892	1.72841	1.73555	1.72182	1.63094
1.67825	1.28616	1.39735	1.35096	1.24803	1.54893
1.37749	1.24127	1.20659	1.29623	1.17020	1.26306
1.25466	1.10529	1.18011	1.28882	1.02360	1.01745
1.05587	1.09097	0.96616	1.00782	0.87936	0.96420
0.98085	0.93748	0.94840	0.98653	0.86596	0.90264
0.92962	0.78622	0.87671	0.95864	0.79844	0.79936
0.86852	0.76338	0.84288	0.95382	0.75302	0.77579
0.75689	0.72450	0.82498	0.92506	0.65272	0.76962
0.73806	0.61611	0.71898	0.87453	0.64449	0.60908
0.71356	0.60953	0.69086	0.83841	0.55401	0.57480
0.65800	0.59430	0.63707	0.81243	0.49825	0.57464
0.65350	0.53202	0.52043	0.78192	0.47489	0.55566
0.47606	0.39913	0.50621	0.58696	0.47460	0.51931
0.46569	0.36320	0.47999	0.57171	0.45585	0.39865
0.43224	0.35265	0.16157	0.49050	0.35425	0.35004
0.36627	0.26125		0.45313	0.27800	0.34501
			0.32093	0.16953	0.28715
				0.13584	0.23950
Average = 1.10300	1.04153	1.05654	1.01181	0.74133	0.76640
# Rejected= 3	3	4	2	1	1

[1]The critical value of A^2 for a test at $\alpha = 0.10$ is 1.933.

Table 4. Variability of total timber value estimates[2] for the Webster tract under nine assumptions of σ and ρ for point location errors.

Assumption	σ used for internal arcs	ρ	Standard Deviation of Value ($)	Coefficient of Variation of Value (%)
A	25.0	0.3	3402.59	2.89
B	25.0	0.6	4791.32	4.07
C	25.0	0.9	5860.13	4.98
D	50.0	0.3	6823.42	5.79
E	50.0	0.6	9594.85	8.15
F	50.0	0.9	11732.27	9.96
G	75.0	0.3	10297.48	8.74
H	75.0	0.6	14446.12	12.27
I	75.0	0.9	17656.80	14.99

[2]Total timber value was estimated to be $117,834.

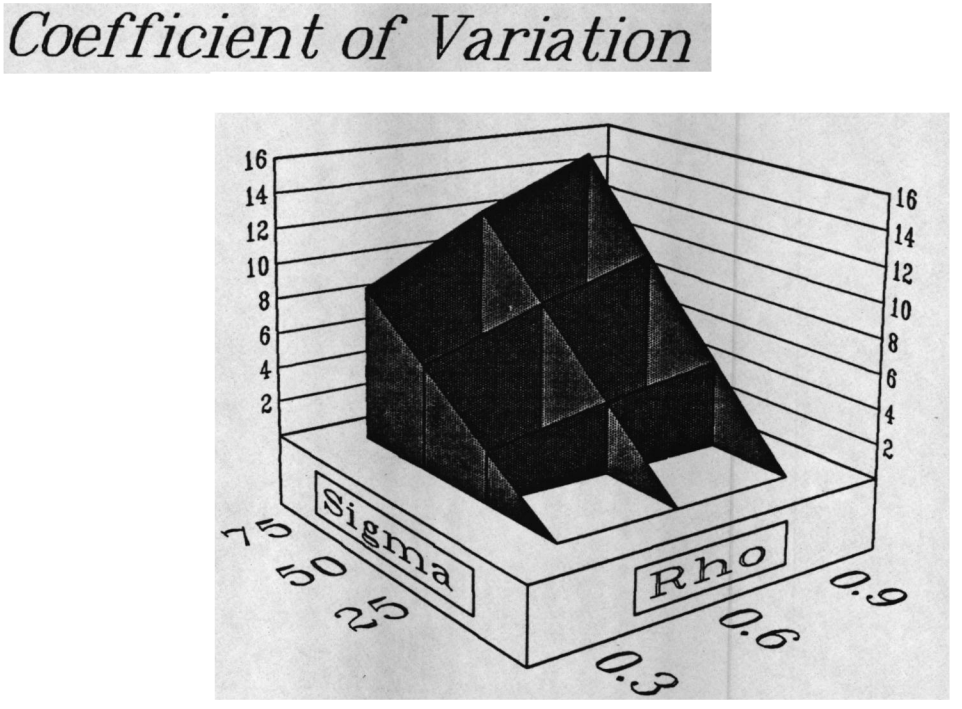


Figure 13. Coefficient of variation of value as a function of assumed values for σ and ρ.

seemingly linear fashion. The choice of σ appears to have a greater influence on the resulting variability of value than does the choice of ρ. Under the set of assumptions which seem most realistic (σ=50, ρ=0.6), the standard deviation of value was 8% of the mean tract value.

The influence of the degree of correlation on area variability may not be intuitively obvious. One might correctly point out that given a polygon composed of a single arc, point errors with the maximum correlation (ρ=1) would result in a shift in polygon location, but would produce *no errors* in area. Very highly correlated errors tend to shift an entire arc in one direction. When a polygon is composed of more than one arc, and the individual arcs are shifted independently of each other, a higher variability will result. Part of the explanation might be that *independent* errors may tend to offset each other; an error at one point may increase polygon area by the same amount that another point error decreases it. When errors are highly correlated, this offsetting effect may be reduced or eliminated.

While this analysis was only a cursory demonstration of an application of the variance formulas, it indicates that variability of summary estimates due to imprecision in spatial data can be significant. The assumptions used here were meant to be only approximations; more intensive analysis of the mapping process would produce more reliable estimates of σ and ρ. However, even such approximations as used herein can be useful for comparative purposes. For example, suppose two tracts of land are being considered for acquisition, and only one will be purchased. If the two are mapped and inventoried in similar manners, comparisons of variability can be made under a variety of assumptions. If the values of the tracts are similar, yet have widely different variances, the tract with lower variance of value would appear to present less risk, and would be the desirable choice. Precise and accurate knowledge of σ and ρ is not necessary in such circumstances.

An interesting extension of the analysis presented here would be to consider both components of variability: that due to variance of inventory estimates, and that due to spatial imprecision. As presented by Schumacher and Bull (1932), the standard error of total volume can be seen as consisting of two terms:

$$SE_{TV}^2 = SE_V^2 * A^2 + SE_A^2 * V^2$$

where: SE_{TV} = standard error of total volume estimate

SE_V = standard error of volume per acre estimate

A = area in acres

SE_A = standard error of area estimate

V = volume per acre

This equation can also be expressed in terms of relative standard errors:

$$SE_{TV\%} = SE_{V\%} + SE_{A\%}$$

where: $SE_{X\%} = \frac{SE_X}{X} = CV_X$ (coefficient of variation of X)

While this example only considered area variability, the above expressions could be used to account for both variance components. Knowledge of the relative contribution of per-acre volume variance and area variance toward total variance might suggest where effort might best be directed to increase precision. More samples might be taken to reduce per-acre volume variability, or more detailed mapping (through better equipment, larger-scale photography, or better ground control) would likely reduce area variability. Consequently, inventory resources could be directed where they would be most efficient in terms of variance reduction.

DISTANCE ERRORS - DERIVATION

Case 1: Vertex Distance

Since the assumption of normality has been adopted for this case, we can state:

$$X_s \sim N(x_s,\sigma_s^2) \qquad Y_s \sim N(y_s,\sigma_s^2)$$

$$X_v \sim N(x_v,\sigma_v^2) \qquad Y_v \sim N(y_v,\sigma_v^2)$$

which implies:

$$X_s - X_v \sim N(x_s - x_v\ ,\ \sigma_s^2 + \sigma_v^2)$$

$$Y_s - Y_v \sim N(y_s - y_v\ ,\ \sigma_s^2 + \sigma_v^2)$$

so

$$\frac{X_s - X_v}{\sqrt{\sigma_s^2 + \sigma_v^2}} \sim N\left(\frac{x_s - x_v}{\sqrt{\sigma_s^2 + \sigma_v^2}}\ ,\ 1\right)$$

and

$$\frac{Y_s - Y_v}{\sqrt{\sigma_s^2 + \sigma_v^2}} \sim N\left(\frac{y_s - y_v}{\sqrt{\sigma_s^2 + \sigma_v^2}}\ ,\ 1\right)$$

By squaring and applying the definition of a non-central chi-square random variable[3] (Johnson and Kotz, 1970) we obtain:

$$\left(\frac{X_s - X_v}{\sqrt{\sigma_s^2 + \sigma_v^2}}\right)^2 \sim \chi'^2_1\left(\lambda = \frac{(x_s - x_v)^2}{\sigma_s^2 + \sigma_v^2}\right) \tag{4.38}$$

$$\left(\frac{Y_s - Y_v}{\sqrt{\sigma_s^2 + \sigma_v^2}}\right)^2 \sim \chi'^2_1\left(\lambda = \frac{(y_s - y_v)^2}{\sigma_s^2 + \sigma_v^2}\right) \tag{4.39}$$

[3] A non-central chi-square is denoted by $\chi'^2_\nu(\lambda)$ where ν is the degrees of freedom and λ is the non-centrality parameter. By definition, $\left(N(\mu,1)\right)^2 \sim \chi'^2_1(\lambda = \mu^2)$

Adding the left sides of (4.38) and (4.39) yields:

$$\left(\frac{X_s - X_v}{\sqrt{\sigma_s^2 + \sigma_v^2}}\right)^2 + \left(\frac{Y_s - Y_v}{\sqrt{\sigma_s^2 + \sigma_v^2}}\right)^2 = \frac{(X_s - X_v)^2 + (Y_s - Y_v)^2}{\sigma_s^2 + \sigma_v^2}$$

$$= \frac{D^2}{\sigma_s^2+\sigma_v^2} \tag{4.40}$$

where D^2 is the square of the measured distance from the point to the vertex. Since non-central chi-squares are additive $\left(\chi'^2_{\nu_1}(\lambda_1) + \chi'^2_{\nu_2}(\lambda_2) = \chi'^2_{(\nu_1+\nu_2)}(\lambda_1+\lambda_2)\right)$, we can add the right sides of (4.38) and (4.39) to obtain the distribution of (4.40):

$$\frac{D^2}{\sigma_s^2 + \sigma_v^2} \sim \chi'^2_2\left(\lambda = \frac{d^2}{\sigma_s^2 + \sigma_v^2}\right)$$

where d^2 is the square of the true distance: $(x_s - x_v)^2 + (y_s - y_v)^2$. This expression can be useful in making statements about the probability of observing certain distances. For example, if we wish to know the probability that a point is at least a distance k from a vertex, when the error variance of the point and the line are known, we can state:

$$\text{Prob}\{D \geq k\} = \text{Prob}\{D^2 \geq k^2\}$$

$$= \text{Prob}\left\{\frac{D^2}{\sigma_s^2+\sigma_v^2} \geq \frac{k^2}{\sigma_s^2+\sigma_v^2}\right\}$$

$$= \text{Prob}\left\{\chi'^2_2\left(\lambda = \frac{d^2}{\sigma_s^2 + \sigma_v^2}\right) \geq \frac{k^2}{\sigma_s^2+\sigma_v^2}\right\}$$

A special case of this would be a test to determine if a point is on a line, or two points are in essentially the same location. Assume we calculate the distance between the points to be k

units. If the points were, in fact, coincident, we would have:

$$\text{Prob}\left\{D \geq k \mid d^2=0\right\} = \text{Prob}\left\{\chi'^2_2(\lambda = 0) \geq \frac{k^2}{\sigma_s^2+\sigma_v^2}\right\}$$

But $\chi'^2_2(\lambda = 0) = \chi^2_2$, so we can use commonly available tables of the percentage points of the χ^2 distribution. We would then calculate $\frac{k^2}{\sigma_s^2+\sigma_v^2}$. If the area under the tail of a χ^2_2 distribution to the right of this point were sufficiently small, we would conclude that the points were distinctly different. The above development follows along the lines of a prevalent application of distance between two bivariate normals: the probability of hitting a target when the target location and the point of impact are bivariate normal random variables (Grad and Solomon, 1955). In such targeting applications, the desired probability is the probability that the distance between the target and the point of impact of the projectile is less than the effective "kill" radius of the weapon.

It would probably be useful to know the mean and variance of D. However, deriving the distribution of:

$$\frac{D}{\sqrt{\sigma_s^2 + \sigma_v^2}} \sim \sqrt{\chi'^2_\nu(\lambda)}$$

in order to obtain them has proven to be intractable. Note that the situation we are evaluating, that of the distance between two independent bivariate normal random variables, is a common one in multivariate statistics (Tatsuoka, 1971). Yet most of the multivariate statistical tools (such as Hotelling's T^2 or Mahalanobis's D^2, as well as the targeting applications just mentioned) for detecting differences between groups utilize measures of *squared* distances. Thus, it appears that in other efforts, the distribution of distance between multivariate normal random variables has been elusive, while inferences using squared distance measures have sufficed.

Case 2: Perpendicular Distance from the Point to a Line Segment

The distribution sought in this section is the *conditional* distribution of distance given the position along the line segment ($X'_p = X'_s$). We have hypothesized (3.9):

$$D|X'_s \sim N(y'_s, \sigma_s^2 + \sigma_p^2)$$

Where $\sigma_p^2 = f(\sigma_1^2, \sigma_2^2, \rho, X'_s)$. The expression for σ_p^2 yielding the best graphical fit to variances observed in simulations was:

$$\sigma_p^2 = \sigma_1^2 p_2^2 + \sigma_2^2 p_1^2 + 2\rho\sigma_1\sigma_2 p_1 p_2 \tag{4.41}$$

where: p_1 = proportion of the line segment from point 1 to point X'_p

$$= \frac{(X'_p - X'_1)}{(X'_2 - X'_1)}$$

p_2 = proportion of the line segment from point X'_p to point 2

$$= (1 - p_1)$$

$$= \frac{(X'_2 - X'_p)}{(X'_2 - X'_1)}$$

Part of the evidence that led to this formulation was noting the location of the value of X'_p at which σ_p^2 was lowest. Simulations indicated that when ρ=0, this point was found at:

$$p_{1min} = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \qquad \text{or, equivalently:} \qquad p_{2min} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Incorporating a nonzero correlation coefficient yielded slightly different locations. It was soon discovered that the X'_p which evidenced the lowest σ_p^2 was such that:

$$p_{1min} = \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} = \frac{\sigma_1(\sigma_1 - \rho\sigma_2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{4.42}$$

or,
$$p_{2min} = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} = \frac{\sigma_2(\sigma_2 - \rho\sigma_1)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{4.43}$$

In fact, taking the derivative of (4.41) with respect to p_1 or p_2 and solving for p_1 or p_2 will yield the above equations.

An interesting result of these expressions is the value of σ^2_p at the minimum:

$$\sigma^2_{p_{min}} = \frac{\sigma_1^2\sigma_2^2(\sigma_2-\rho\sigma_1)^2 + \sigma_1^2\sigma_2^2(\sigma_1-\rho\sigma_2)^2 + 2\rho\sigma_1^2\sigma_2^2(\sigma_1-\rho\sigma_2)(\sigma_2-\rho\sigma_1)}{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)^2}$$

$$= \frac{\sigma_1^2\sigma_2^2\left(\sigma_2^2-2\rho\sigma_1\sigma_2+\rho^2\sigma_1^2+\sigma_1^2-2\rho\sigma_1\sigma_2+\rho^2\sigma_2^2+2\rho\sigma_1\sigma_2-2\rho^2\sigma_1^2-2\rho^2\sigma_2^2+2\rho^3\sigma_1\sigma_2\right)}{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)^2}$$

$$= \frac{\sigma_1^2\sigma_2^2\left(\sigma_2^2(1-\rho^2) + \sigma_1^2(1-\rho^2) - 2\rho\sigma_1\sigma_2(1-\rho^2)\right)}{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)^2}$$

$$= \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

Table 5 shows the formulas for p_{1min} and $\sigma^2_{p_{min}}$ for the general case, and for the specific cases when $\sigma_1=\sigma_2=\sigma$, and when ρ = -1, 0, or 1. If we assign similar variances to all points in an arc, then most often, σ_1 will equal σ_2, the exceptions being if point 1 or 2 is a node with a variance different from the rest of the arc. Thus, we will usually have p_{1min} = 0.5, and a simpler expression for $\sigma^2_{p_{min}}$. The expressions shown in the table have a degree of intuitive appeal. For example, when ρ=1 and $\sigma_1=\sigma_2$, the errors at each endpoint are identical. Therefore, every line segment with errors will be parallel to the nominal line segment (without errors). This would imply that since p_{1min} is undefined, there is no single point along the line at which variance is minimized; the variance is the same for all locations along the segment. Conversely, if ρ= -1, then the errors at point 2 will be *equal but opposite in direction* from the errors at point 1. This would imply that every erroneous line segment would intersect the nominal line segment at its' midpoint, and that the variance of errors at that point would be 0. If $\sigma_1 = \sigma_2$, the equations simplify considerably, and we note a linear effect of ρ on $\sigma^2_{p_{min}}$.

Table 5. Formulas for p_{1min} and $\sigma^2_{p_{min}}$.

		General Case	$\rho = -1$	$\rho = 0$	$\rho = 1$
p_{1min}	$\sigma_1 \neq \sigma_2$	$\frac{\sigma_1(\sigma_1-\rho\sigma_2)}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}$	$\frac{\sigma_1}{\sigma_1+\sigma_2}$	$\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}$	0 if $\sigma_1<\sigma_2$ 1 if $\sigma_1>\sigma_2$
	$\sigma_1=\sigma_2$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	undefined
$\sigma^2_{p_{min}}$	$\sigma_1 \neq \sigma_2$	$\frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}$	$\frac{\sigma_1^2\sigma_2^2}{(\sigma_1+\sigma_2)^2}$	$\frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$	$\min(\sigma_1^2, \sigma_2^2)$
	$\sigma_1=\sigma_2$	$\frac{\sigma^2(1+\rho)}{2}$	0	$\frac{\sigma^2}{2}$	σ^2

The behavior of the function in (4.41) was noted graphically to be very consistent with simulations, over a wide range of the parameters σ_1, σ_2, ρ, and X'_s. Other functions tested included a parabola and a cosine curve which were fit to duplicate the noted results at the endpoints and the point (p_{1min}) at which variance of Y' errors were smallest. Neither of these curves performed well at other points, or were consistent across ranges of the parameters.

Equation (4.41) therefore provides a parameter which allows us to specify completely the hypothesized normal distribution of perpendicular distance from a point to a line (3.9). The translation of (X,Y) to (X',Y') was performed to simplify notation, and has no effect upon the results. Given any point (X_s,Y_s) and its variability σ_s^2, and any line segment from X_i,Y_i to X_{i+1},Y_{i+1} and its error structure (σ_i^2, σ_{i+1}^2, ρ_i), we can first calculate the point of intersection of the perpendicular from the subject point to the line segment, and then calculate the parameters of the hypothesized distribution.

DISTANCE ERRORS - VALIDATION OF THE DISTRIBUTION

The normal distribution was hypothesized for perpendicular distance from a point to a line. To evaluate the validity of this assumption in an admittedly restricted test, twenty simulations of 80 iterations each were performed using 54 sets of assumptions. The Anderson-Darling statistic was used to compare the empirical distributions resulting from the simulations to the hypothesized distribution. At α=0.10, and n=80, the rejection region for this statistic is A^2>1.933. At this Type I error rate, data from a normal distribution might be expected to result in a rejection about two times out of 20. Table 6 shows the number of rejections (out of 20 simulations) for each of the 54 sets of assumptions.

Table 6. Number of rejections of a null hypothesis[4] of normal distance errors in 54 sets of simulations.

		$\rho = 0.3$			$\rho = 0.7$		
		$X'_s = -25$	$X'_s = 0$	$X'_s = 25$	$X'_s = -25$	$X'_s = 0$	$X'_s = 25$
$\sigma_2^2 = 1$	Y'_s=32	2	3	2	4	2	1
	Y'_s= 8	2	2	2	2	1	2
	Y'_s= 2	1	3	1	3	1	5
$\sigma_2^2 = 3$	Y'_s=32	2	3	3	3	2	2
	Y'_s= 8	2	1	1	2	2	3
	Y'_s= 2	1	4	2	1	4	2
$\sigma_2^2 = 5$	Y'_s=32	0	2	1	2	2	1
	Y'_s= 8	3	3	1	1	1	1
	Y'_s= 2	3	1	2	2	0	1

[4]Testing at α=10%, the null hypothesis is rejected if $A^2 > 1.933$.

The average number of rejections for the 54 sets of simulations was 1.96. While not *confirming* the normal distribution as the best model, this data certainly provides no evidence for rejecting it. Inspection of the relationship between frequency of rejections and the parameters shows no discernible pattern; none of the parameters used seems to lead to an undue number of rejections or acceptances.

Therefore, under restricted circumstances, the normal distribution seems adequate for modeling errors in perpendicular distance from a point to a line. The mean of these errors will be zero, and the variance will be dependent upon the variances and correlations of the points involved, as well as the position of the intersection of the perpendicular.

DISTANCE ERRORS - EXAMPLE APPLICATION

The 4-chain by 5-chain grid of plots laid over the Webster tract resulted in 107 plots being located within the tract (Figure 14). The distances between the plots and the nearest stand boundary line were calculated in a Fortran program. Ten of the 107 plots were nearest to a vertex, while the distance from the remaining 97 plots to the nearest stand line was a perpendicular distance. For each plot, under each of the six sets of assumptions, a p-value was recorded which indicated the probability of observing the measured distance (d) if the plot were on the stand boundary (d=0). Recall that large p-values are associated with ambiguous plots; for plots that are not on the stand boundary, one wishes to obtain a small p-value, thereby leading to the correct rejection of the hypothesis: d=0. The relative frequency of these p-values is shown in Figure 15.

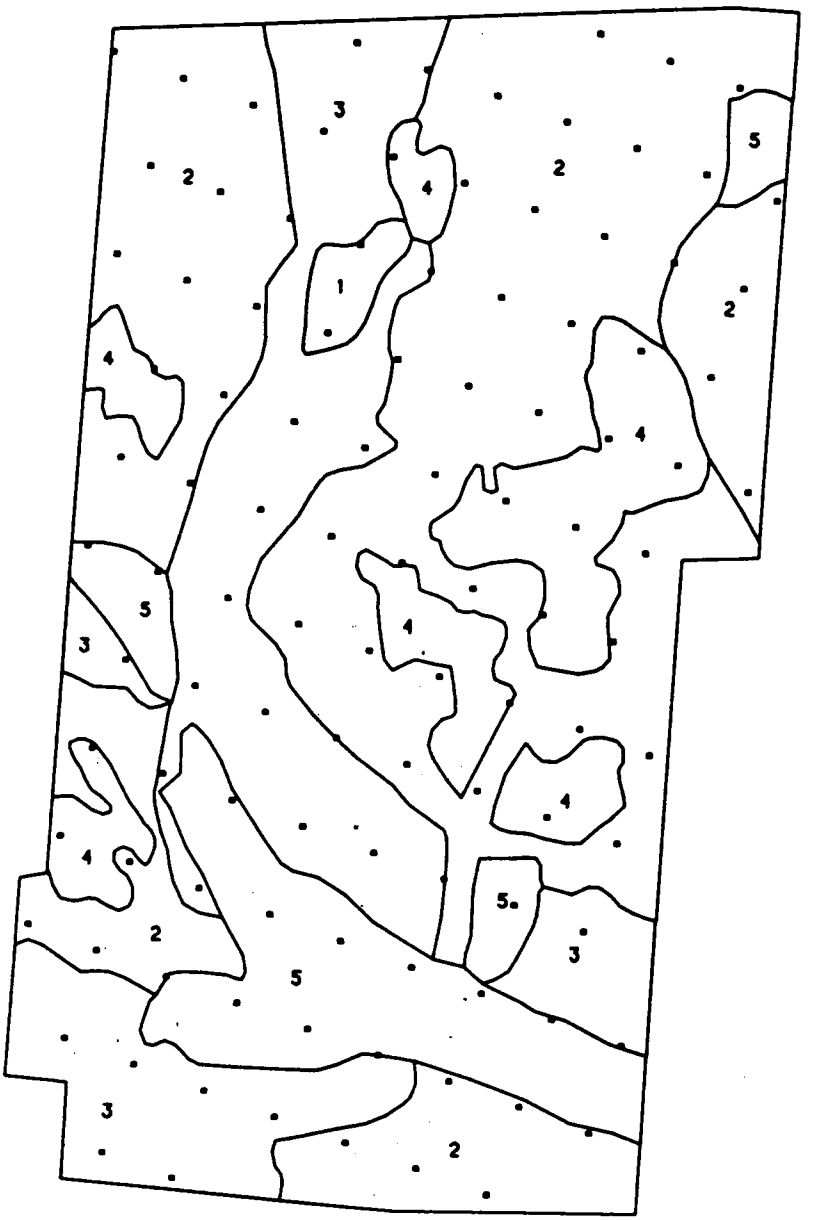
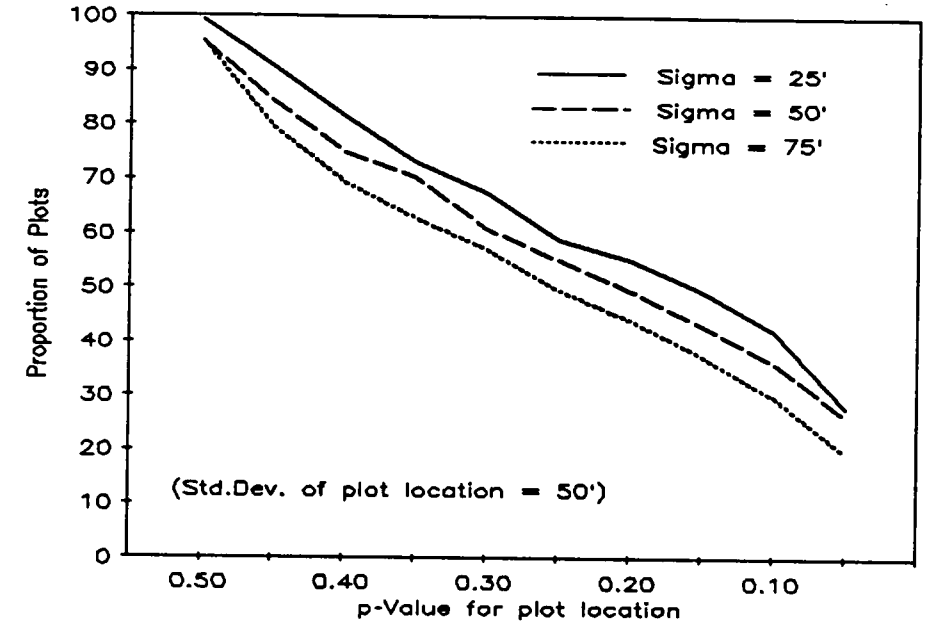


Figure 14. Timber cruise plot locations for the Webster tract.



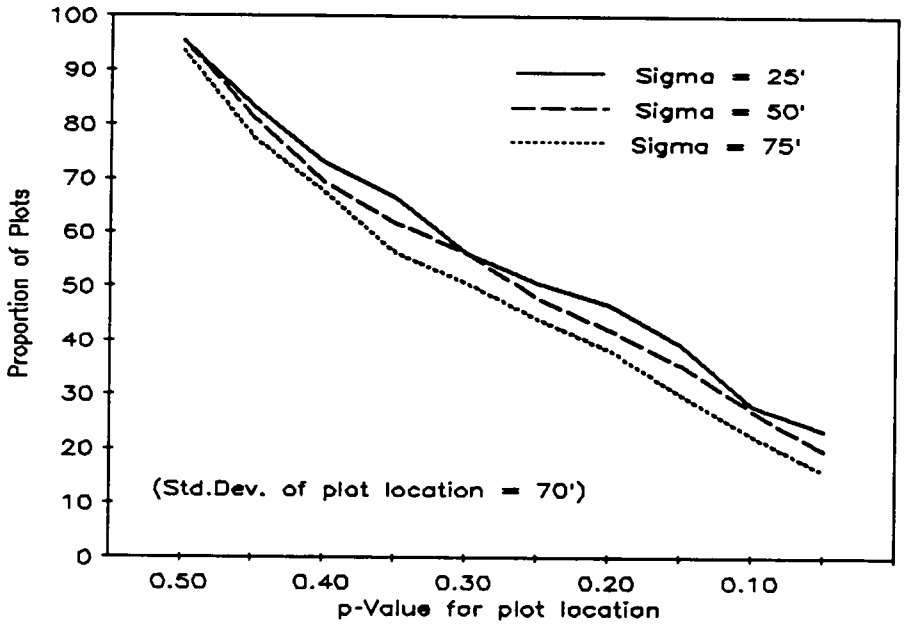


Figure 15. Percentage of plots by p-value for six sets of assumptions.

For the most generous of the six sets of assumptions (σ_s=50.0, σ_v=25.0, ρ=0.6), only 42% of the plots were indicated as being separate from the stand lines with probability 90%. Slightly more than half of the plots could be located within a stand with 80% probability. At the most extreme of assumptions (σ_s=70.0, σ_v=75.0, ρ=0.6), 22% were confirmed to be within a given stand at 90% probability, and only 38% at 80% probability.

These figures at first appear to be alarming. It does not seem credible that half of the inventory plots are ambiguously located; indeed, that is not the case. Several factors are unaccounted for in this brief example. When locating plots in the field, a forester typically uses a compass and pacing to determine location. However, there are often landmarks (such as stand boundaries themselves) which supplement simple bearings and distances to aid in location. When a plot is indicated on the map to be within a pine stand but adjacent to a lake, the plot will almost certainly be taken in the pine stand, even if strict adherence to bearings and distances might lead to a location in the lake. The point is that *discernible* boundaries will influence the ability to locate nearby points. However, if the boundary were more indeterminate (or even invisible, such as a county line or unmarked property boundary), these results are more believable. As a comparison, Blakemore (1984) reported that only half of the 780 sites he studied were uniquely assignable to one administrative polygon in his epsilon band analysis.

Figure 16 indicates which plots could and could not be judged to be significantly distant from the stand lines with 80% probability (using the most generous of the assumptions). It was noted that some plots identified as ambiguous ($p > 0.20$) were farther from the boundaries than others which were not deemed ambiguous. For example, plot 24 was 55 feet from a stand line and was not considered ambiguous, while plot 4 was 88 feet from a line and *was* considered ambiguous. The difference is that plot 4 was closest to a vertex, while the distance from plot

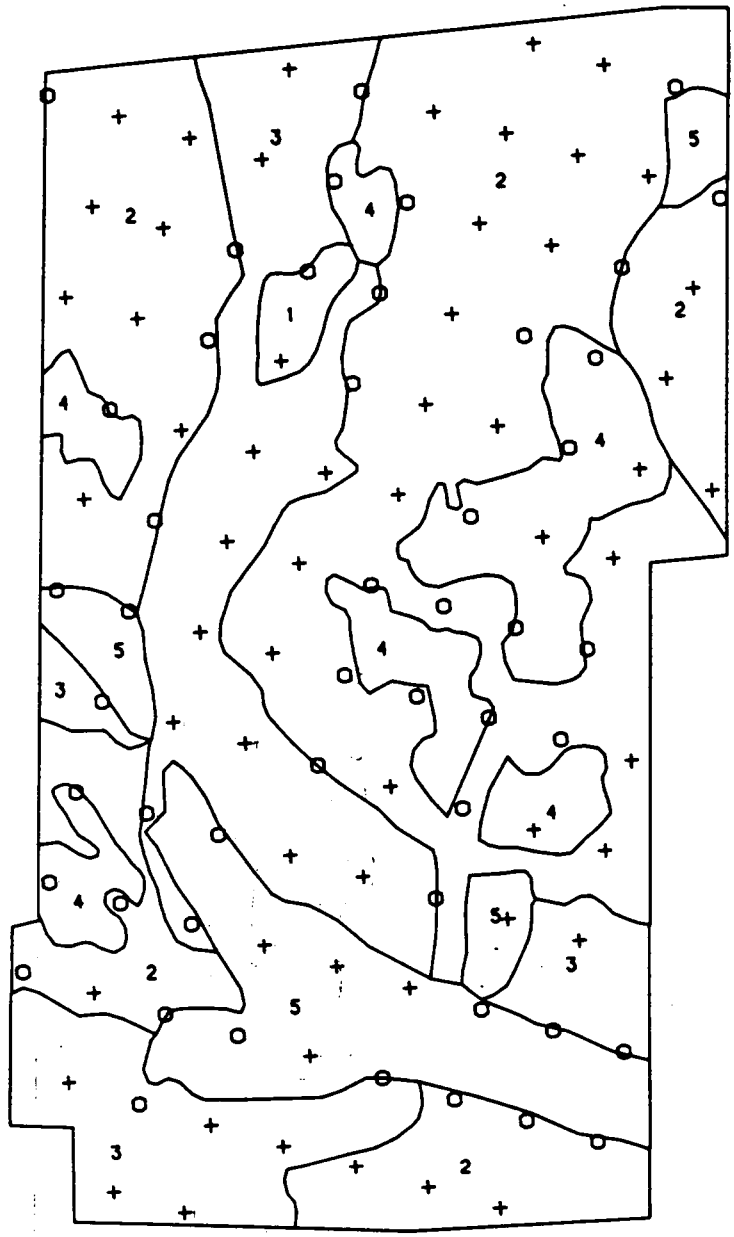


Figure 16. Locations of ambiguously defined cruise plots. Circles indicate plots which could not be located in a specific stand with at least 80% probability. (+)'s indicate plots which could be located with at least 80% probability in a stand.

24 to the stand line was a perpendicular distance. The concavity of the isodensity lines about the line segment create a narrower "zone of uncertainty", allowing closer points to remain unambiguously defined. This observation leads to a paradox which will be discussed in the following section.

The findings of this admittedly superficial example reinforce warnings made by several authors that "use of a GIS . . . could lead to false perceptions about the quality of the results" (Bailey, 1988). The precision with which lines and points are drawn by a digital plotter lends undue credence to the precision of the data being mapped. The plots and stands in this example exhibit far more variability in location than is evident in Figure 14.

DISCUSSION OF MODEL STRENGTHS AND WEAKNESSES

The models developed here offer an opportunity to examine the impacts of spatial location errors on measurements such as area and distance calculated in a GIS. The models appear to behave well in simulations, and are based upon fairly weak assumptions regarding point errors. However, there are some drawbacks which merit discussion.

First, it was noted that the definition of the centroid location influenced the variance of polygon area. This may have been due to the approximation method used; covariances between non-adjacent triangles were omitted in the derivation of the variance of polygon area. It is hypothesized that complete specification of polygon area variance, including covariances of non-adjacent polygons would result in a variance formula independent of centroid location.

However, this dependence upon the centroid does not appear to be a serious shortcoming since the simulations performed to test the normality of errors did not indicate that the variance estimate was in error. Further simulations of polygon area errors could support the validity of this model, or indicate the need for a centroid-insensitive formulation.

Second, the results obtained depend on some arguable assumptions. Notably, the polygon variance expressions require assumptions that X and Y errors are not correlated, and that serial correlation is of order 1 (based upon results from Keefer, 1988). The distance distributions involved assumptions of normality, which, while reasonable, cannot be proven to be valid. Viewed as an initial effort, these results might be improved upon in future work by relaxing the assumptions. However, by including such terms as the correlation between X and Y errors in the model, additional variables must be specified in order to use the model. So little is currently known about correlation of errors in GIS systems that further parameterization may not be called for at this point.

Third, the models presented here considered only one aspect of location errors: errors of *commission*. That is, we considered what errors occur at points which have been digitized into a spatial database. Errors of *omission* have been ignored. Since mapping by definition involves generalization, there may be significant features (such as meander loops in streams or convexities in boundary lines) which are lost in the mapping process. Such errors are difficult to model in any situation, since there is little basis for knowing when and where generalization has resulted in omissions. Some automated mapping and drafting systems include the ability to "un-generalize" by *adding* detail to map features. An example is "smoothing" lines by fitting spline curves to digitized points. Such endeavors are generally practiced by cartographers seeking to improve the appearance of map products, who are not overly concerned with the

integrity of the coordinate data therein.

The paradox referred to in the previous section is that according to the models developed here, the precision of line location (σ_p^2) *improves with increased distance from a sampled point.* The lower variances towards the middle of line segments suggests that the fewer points are used to represent a line, the better. Following such a line of thinking to an extreme would lead to tremendous omission errors, which are not being modeled. If we were to assume that some degree of omission (generalization) errors are inevitable between the points sampled on a line, then reverting back from the concave isodensity region in Figure 11b to the parallel isodensity region in Figure 11a might be advised. It might also be argued that the complexity involved in obtaining the estimate of σ_p^2 based on p_1 (proportional distance from a vertex) represents an over-quantification. The epsilon band model, wherein the band of uncertainty about a line has a constant width, may be more suitable for the applications considered here. Certainly, such a model would obviate the paradoxical results obtained, and provide more consistency in application than the two-case model for distances developed here.

Other situations may indicate that the concave isodensity region is appropriate. For example, land ownership in a majority of the United States follows rectangular patterns established by the Public Land Survey system of townships and square-mile sections. Many polygons (representing ownership, land cover, road networks, etc.) in these areas are square or rectangular, with long distances between vertices, and errors at corners are more likely to be independent. In such cases, ommission errors are less likely, and boundaries may truly be more precisely located at a distance from polygon corners.

The modeling approach followed herein has several advantages over current applications of models such as the epsilon-band. First, a constant error-band width is not assumed for all arcs in a map. This feature provides for more realistic modeling of errors accumulated in an overlay map, wherein arcs can be traced back to different source maps, and may have different error structures. Second, variances and covariances are statistically defined and approximate distributions are suggested; previous models avoided probabilistic statements. Finally, errors in both points and line segments were considered. This represents an extension of the qualitative point-in-polygon analysis cited in the literature review.

The models for area and distance errors should have a variety of beneficial applications. By obtaining variance estimates for a variety of polygons, more could be learned about the influence of such factors as polygon size, shape, and complexity on the variance of area. The effect of varying assumptions (or map accuracy standards) could be tested on a case-by-case basis. As mentioned in the polygon area application example, the relative importance of spatial and attribute errors in a cartographic modeling situation can be estimated. If nothing else, applications of models such as are developed here might lead to a wider recognition of the indeterminacy of spatial phenomena in GIS systems.

DISCUSSION OF A MODEL FOR LINE LENGTH

When intially proposing the work described here, the derivation of mean and variance of line lengths was considered also. The derivation was soon found to be very problematic, and further work in this area was abandoned. A brief discussion of the problems involved might provide an interesting contrast to the models developed for distance and area.

Obviously, line length derives from distances between points. In a GIS, length along a line or arc is calculated as the sum of distances between successive points in a digitized line:

$$L_N = \sum_{i=1}^{n-1} \sqrt{(X_i - X_{i+1})^2 + (Y_i - Y_{i+1})^2} \tag{4.44}$$

where: L_N = length of a line composed of n segments

X_i, Y_i = coordinates of point i along the line

The first difficulty encountered is attempting to evaluate $E(L_N)$. In order to do this, one must obtain the expectation of a function involving a square root for which the joint distribution of X_i and Y_i must be known. If we assume the normal distribution (as seems most realistic), we are faced with the same situation as in considering the distance between two bivariate normal random variables. The squared distance is known to be distributed as χ'^2_2, but the distribution of distance is unknown and intractable.

Next, evidence from several authors (Baugh and Boreham, 1976; Keefer, 1988) suggests that contrary to the situation encountered with area, errors in line length result in a bias when length is measured using the conventional expression (4.44). A similar bias is noted when using D^2 to estimate d^2:

$$E(D^2) = E\left(\frac{D^2}{\sigma_v^2 + \sigma_s^2}\right)(\sigma_v^2 + \sigma_s^2)$$

$$= E\left(\chi'^2_{\nu=2}\left(\lambda = \frac{d^2}{\sigma_v^2 + \sigma_s^2}\right)\right)(\sigma_v^2 + \sigma_s^2)$$

$$= (\lambda + \nu)(\sigma_v^2 + \sigma_s^2)$$

$$= d^2 + 2(\sigma_v^2 + \sigma_s^2)$$

If D were unbiased for d, then by applying the above equation, and using the fact that $\mathrm{Var}(D) = \mathrm{E}(D^2) - [\mathrm{E}(D)]^2$, we would obtain $\mathrm{Var}(D) = 2(\sigma_v^2 + \sigma_s^2)$.

Finally, to further complicate matters, obtaining the distribution of distance would not be sufficient. It is readily noted that the distances between successive pairs of points on a line are correlated, as was the case with adjacent triangles in the derivation of polygon area. Thus, pairwise covariances would be required to obtain the variance of line length:

$$\mathrm{Var}(L_N) = \sum_{i=1}^{n-1} \mathrm{Var}(D_i) + 2\sum_{i=1}^{n-2} \mathrm{Cov}(D_i, D_{i+1})$$

where: D_i = distance between points i and $i+1$: $\sqrt{(X_i - X_{i+1})^2 + (Y_i - Y_{i+1})^2}$

Once again, these pairwise dependencies interfere with the strict application of the Central Limit Theorem in obtaining the distribution of L_N. All these complications suggest that modeling errors in line length will require a different approach than those followed in this work.

Chapter 5 - SUMMARY

Geographic Information Systems are becoming commonplace in forest resource management organizations. They have progressed from being little more than automated drafting machines to being requisite tools for managing and manipulating large spatial databases. While the technological capabilites of these systems have rapidly improved, the reliability of results has often come into question. The need for estimates of accuracy and precision of derived variables such as area, length, and distance has been stated repeatedly in the literature.

The objective of the work described here was to develop a procedure for incorporating information or assumptions about the locational variability of points in arc-node databases into the analyses common in forest management applications of GIS. First, assumptions regarding the variability of points were presented in statistical expressions. Then, using the algebra of expectations of functions of random variables, the mean and variance of polygon area were derived. The derivation was based upon triangles formed by line segments in the polygon boundary and a centroid location. Note that the derivation of the variance expression was an approximation, which may have resulted in a dependency of area variance on centroid location. Therefore, in order to obtain consistent estimates of variance, a minimum-variance centroid was defined. Next, the covariance of area of adjacent polygons was derived. It was thought that the distribution of polygon areas was approximately normal; some simulations of polygon area errors revealed no reason to believe otherwise. The centroid definition and expressions of polygon area variance and covariance provided the necessary tools to evaluate the variability of estimates obtained by multiplying per-unit-area figures by area estimates.

Next, by extending the assumptions about point location errors to include normality, it was possible to obtain the distribution of distance between two points. An approximate distribution of perpendicular distance from a point to a line segment was also described. The hypothesized distributions withstood testing of simulated distance errors using a variety of parameters. An application of the use of distributional information about distances was demonstrated: a point-in-polygon analysis revealed that many points which are located deterministically within polygons cannot be shown to be distinguishable from the polygon boundaries at high confidence levels.

The models of error assumed and derived herein have several drawbacks. First, there is currently very little known about some of the parameters (notably the correlation between adjacent point errors) which are required for application of the variance expressions. Second, the variance and covariance of polygon area depended upon the location of the centroids of the polygons. A model which is not dependent on either centroid locations or coordinate axis scale or orientation might be preferable. Finally, the models considered only errors at digitized points in a spatial database, neglecting potential generalization errors which occur between such points. However, it would be difficult at this point in time to account for such omission errors, as they tend to be even more elusive than the errors committed at recorded coordinates.

The advantages of the models developed here include the allowance for error structures which differ among arcs, the probabilistic statements which can be made with distributional assumptions, and the expansion of previous methods to incorporate variability in both points and arcs. Obviously, more research is needed in several areas. Sensitivity analysis would indicate the effect of varying parameter values on the estimates of variability that are obtained. Regressing polygon area variability against measures of polygon shape, complexity, and size may

provide insights into the effects of these variables on area precision. More efforts are also needed to improve ways of obtaining estimates of the parameters used in the models: variability and correlation of point location errors. Finally, no satisfactory expression for the accuracy or precision of line length estimates has yet been developed.

The ultimate goal of studies of error in forestry GIS systems is to provide resource managers with some indication of the reliability of results of GIS analyses. As GIS systems, users, and databases become more prolific, more caution must be exercised in the interpretation of the products derived from GIS. Only with some understanding of the reliability of these products can resource managers prudently apply the information developed, and reap the benefits that GIS advocates have promised.

Chapter 6 - BIBLIOGRAPHY

Aronoff, S. 1982a. Classification accuracy: a user approach. Photogrammetric Engineering and Remote Sensing 48(8):1299-1307.

Aronoff, S. 1982b. The map accuracy report: a user's view. Photogrammetric Engineering and Remote Sensing 48(8):1309-1312.

Averack, R. and M.F. Goodchild. 1984. Methods and algorithms for boundary definition, pp. 238-250 in: *Proceedings of IGU International Symposium on Spatial Data Handling*, Zurich, Switz., Aug. 1984.

Avery, Thomas Eugene and Harold E. Burkhart. 1983. *Forest Measurements, Third Edition*, McGraw-Hill Book Company. 331 pp.

Bailey, Robert G. 1988. Problems with using overlay mapping for planning and their implications for geographic information systems. Environmental Management 12(1):11-17.

Baugh, Ian D.H. and Jeremy R. Boreham. 1976. Measuring the coastline from maps: a study of the Scottish mainland. Cartographic Journal 13:167-171.

Bennett, H.C. 1977. The cartographic data base- reliability of chaos? Proceedings of American Congress on Surveying and Mapping. pp. 675-680.

Berry, J.K. 1987. Computer-assisted map analysis: potential and pitfalls. Photogrammetric Engineering and Remote Sensing 53(10):1405-1410.

Blakemore, M. 1984. Generalisation and error in spatial data bases. Cartographica 21:131-139.

Bondesson, Lennart. 1986. *Estimation of standard errors of area estimates of forest compartments obtained by traversing.* Swedish University of Agricultural Sciences, Section of Forest Biometry, S-901 83 Umea, Sweden. Report 24. 49 pp.

Bouille, F. 1982. Actual tools for cartography today. Cartographica 19(2):27-32.

Breed, Charles B. 1971. *Surveying, Third Edition.* John Wiley & Sons, N.Y. 495 pp.

Buckles, B.P. and F.E. Petry. 1982. A fuzzy representation of data for relational databases. Fuzzy Sets and Systems 7:213-226.

Burrough, P. A. 1986. *Principles of Geographical Information Systems for Land Resources Assessment.* Oxford University Press, New York. 193 pp.

Campbell, James B. 1987. *Introduction to Remote Sensing.* The Guilford Press, New York. 551 pp.

Carter, James R. 1988. Digital representation of topographic surfaces. Photogrammetric Engineering and Remote Sensing 54(11):1557-1580.

Chrisman, Nicholas R. 1980. Assessing LANDSAT accuracy: a geographic application of misclassification analysis. Second Colloquium on Quantitative and Theoretical Geography, Trinity Hall, Cambridge, England.

Chrisman, Nicholas R. 1982a. Beyond accuracy assessment: correction of misclassification. pp. 123-132 in: *ISPRS Commission IV Symposium*, American Society for Photogrammetry and Remote Sensing, and American Congress on Surveying and Mapping, Falls Church, Va. 571 pp.

Chrisman, Nicholas R. 1982b. A theory of cartographic error and its measurement in digital data bases. Auto Carto 5, pp. 159-168.

Chrisman, Nicholas R. 1982c. Methods of spatial analysis based on error in categorical maps. Ph.D. dissertation, University of Bristol.

Chrisman, Nicholas R. 1984a. The role of quality information in the long-term functioning of a geographic information system. Cartographica 21:79-87.

Chrisman, Nicholas R. 1984b. On storage of coordinates in GIS. Geoprocessing 2:259-270.

Chrisman, Nicholas R. 1987a. The accuracy of map overlays: a reassessment. Landscape and Urban Planning 14:427-439.

Chrisman, Nicholas R. 1987b. Efficient digitizing through the combination of appropriate hardware and software for error detection and editing. International Journal of GIS 1(3):265-277.

Chrisman, Nicholas R., and Brian S. Yandell. 1988. Effects of point error on area calculations: a statistical model. Surveying and Mapping 48(4): 241-246.

Congalton, R.G., R.G. Oderwald, and R.A. Mead. 1983. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogrammetric Engineering and Remote Sensing 49(12):1671-1687.

Covington, W.W., D.B. Wood, D.L. Young, D.P. Dykstra, and L.D. Garrett. 1988. TEAMS: A decision support system for multiresource management. Journal of Forestry 86(8):25-33.

Cowen, David J. 1988. GIS versus CAD versus DBMS: what are the differences? Photogrammetric Engineering and Remote Sensing 54(11):1551-1555.

Franklin, W. R. 1984. Cartographic errors symptomatic of underlying algebra problems, pp. 190-208 in: *Proceedings of an International Symposium on Spatial Data Handling*, Zurich, Switz.

Frolov, Y.S. and D.H. Maling. 1969. The accuracy of area measurements by point counting techniques. Cartographic Journal 6:21-35.

Goodchild, Michael. 1978. Statistical aspects of the polygon overlay problem, in: *Harvard Papers on Geographic Information Systems* (ed. G. Dutton), Vol. 6, Laboratory for Computer Graphics and Spatial Analysis, Graduate School of Design, Harvard University.

Goodchild, Michael. 1980a. The effects of generalization in geographic data encoding, pp. 191-205 in: *Map Data Processing* (ed. H. Freeman and G.G. Pieroni), Academic Press, New York.

Goodchild, Michael. 1980b. Fractals and the accuracy of geographical measures. Mathematical Geology 12:85-98.

Goodchild, Michael. 1982. Accuracy and spatial resolution: critical dimensions for geoprocessing, pp. 87-90 in: *Computer Assisted Cartography and Geographic Information Processing: Hope and Realism.*

Goodchild, M.F. and Odette Dubuc. 1987. A model of error for choroplethic maps, with applications to geographic information systems. Auto Carto 8, pp. 165-174.

Goodchild, Michael. 1988. The national center for geographic information analysis: an update. Unpublished presentation at GIS/LIS '88, San Antonio, Texas, December 1, 1988.

Grad, Arthur, and Herbert Solomon. 1955. Distribution of quadratic forms and some applications. Annals of Mathematical Statistics 26:464-477.

Greenland, A. and R. Socher. 1985. Statistical evaluation of accuracy for digital cartographic databases. Auto Carto 7, pp. 212-218.

Hanson, Robert J. 1988. Conversion of municipality records for a geographic information system: considerations and techniques. pp. 111-121 in: *GIS/LIS '88 Proceedings*, American Society for Photogrammetry and Remote Sensing, and American Congress on Surveying and Mapping, Falls Church, Va. 980 pp.

Hord, R.M. and W. Brooner. 1976. Land use map accuracy criteria. Photogrammetric Engineering and Remote Sensing 42:671-677.

IMSL, 1987. STAT/LIBRARY: Fortran Subroutines for Statistical Analysis. IMSL, Inc. Houston, Texas. 1231 pp.

Jenks, G.F. and F.C. Caspall. 1971. Error on choroplethic maps: definition, measurement, reduction. Annals of the Association of American Geographers 61(2):217-244.

Jenks, G.F. 1981. Lines, computers, and human frailties. Annals of the Association of American Geographers 71(1):1-10.

Johnson, Norman L., and Samuel Kotz. 1970. *Distributions in Statistics: Continuous Univariate Distributions, Vol. II.* John Wiley & Sons, New York. pp. 130-148.

Johnston, Kevin. 1987. Natural resource modeling in the geographic information system environment. Photogrammetric Engineering and Remote Sensing, 53(10):1411-1415.

Keefer, Brenton J. 1988. Effect of manual digitizing error on the accuracy and precision of polygon area and line length. MS thesis, Department of Forestry, Virginia Polytechnic Institute and State University, Blacksburg, VA. 180 pp.

Keefer, Brenton J., James L. Smith, and Timothy G. Gregoire. 1988. Simulating manual digitizing error with statistical models. pp. 475-483 in: *GIS/LIS '88 Proceedings*, American Society for Photogrammetry and Remote Sensing, and American Congress on Surveying and Mapping, Falls Church, Va. 980 pp.

Lang, L. 1988. The movers and shakers of GIS. Professional Surveying 8(4):4-9.

Leung, Yee. 1987. On the imprecision of boundaries. Geographical Analysis 19(2):125-151.

Loehle, Craig. 1983. The fractal dimension and ecology. Speculations in Science and Technology 6(2):131-142.

MacDougall, E. B. 1975. The accuracy of map overlays. Landscape Planning 2:23-30.

MacEachren, Alan M. 1982. Choropleth map accuracy: characteristics of the data. Auto Carto 5, pp. 499-507.

Maffini, Giulio. 1987. Raster versus vector data encoding and handling: a commentary. Photogrammetric Engineering and Remote Sensing 53(10):1397-1398.

Maling, D. H. 1989. *Measurement from Maps: Principles and Methods of Cartometry.* Pergamon Press, Oxford. 577 pp.

Mandelbrot, B.B. 1977. *Fractals: Form, Chance, and Dimension.* Freeman, San Francisco. 365 p.

McAlpine, J.R. and B.G. Cook. 1971. Data reliability from map overlay, in: *Proceedings of the 43rd Congress of the Australian and New Zealand Association for the Advancement of Science,* Brisbane, Australia.

McHarg, Ian L. 1971. *Design with Nature.* Doubleday & Co., Doubleday/Natural History Press, Garden City, N.Y.

Mead, D. A. 1982. Assessing data quality in geographic information systems, pp. 51-62 in: *Remote Sensing for Resource Management,* Soil Conservation Society of America, Johannsen and Sanders, ed.

Meyer, Walter H. 1963. Some comments on the error of the total volume estimate. Journal of Forestry 61(7):503-507.

Morrison, Joel, editor. 1988. The Proposed Standard for Digital Cartographic Data, The American Cartographer 15(1).

Monmonier, Mark S. 1982. *Computer-assisted Cartography- Principles and Prospects.* Prentice-Hall, Inc., Englewood Cliffs, N.J. 214 pp.

Morrison, J.L. 1980. Computer technology and cartographic change, *in*: The computer in contemporary cartography, Taylor, ed., pp. 5-23, Chichester, Wiley.

Muller, J.C. 1977. Map gridding and cartographic errors- a recurrent argument. Canadian Cartographer 14(2):152-167.

Newcomer, J. A. and J. Szajgin. 1984. Accumulation of thematic map errors in digital overlay analysis. The American Cartographer 11(1):58-62.

Nicholson, W.L. 1958. On the distribution of 2 x 2 random determinants. Annals of Mathematical Statistics 29:575-580.

Otawa, Toru. 1987. Accuracy of digitizing: overlooked factor in GIS operations., pp. 295-299 in: *GIS* '87, American Society for Photogrammetry and Remote Sensing, and American Congress on Surveying and Mapping, Falls Church, Va. 756 pp.

Parker, H. Dennison. 1988. The unique qualities of a geographic information system: a commentary. Photogrammetric Engineering and Remote Sensing 54(11):1547-1549.

Perkal, J. 1966. On the length of empirical curves. Discussion Paper 10, Ann Arbor Michigan. Michigan Inter-University Community of Mathematical Geographers.

Petersohn, C. and A.P. Vanderohe. 1982. Site-specific accuracy of digitized property maps. Auto-Carto 5, pp. 607-620.

Peuker, T. and N. Chrisman. 1975. Cartographic data structures. The American Cartographer 2(1):55-69.

Peuker, T.K. 1976. A theory of the cartographic line. Internat'l Yearbook of Cartography Vol 16., pp. 139-143.

Peuquet, Donna. 1984. A conceptual framework and comparison of spatial data models. Cartographica 21(4):66-113.

Prisley, S.P. and J.L. Smith. 1987. Using classification error matrices to improve the accuracy of weighted land cover models. Photogrammetric Engineering and Remote Sensing 53(9):1259-1263.

Robinson, V.B. and A.H. Strahler. 1984. Issues in designing GIS under conditions of inexactness. Proceedings of the 10th International Symposium on Remotely Sensed Data. pp. 198-204.

Robinson, V. and A. Frank. 1985. About different kinds of uncertainty in collections of spatial data. Auto Carto 7, pp. 440-449.

Rosenfield, A. 1980. Tree structures for region representation, in: *Map Data Processing*, (Freeman and Pieroni, eds.), pp. 137-150. Academic Press, N.Y.

Schumacher, F.X., and H. Bull. 1932. Determination of the errors of estimate of a forest survey, with special reference to the bottomland hardwood forest region. Journal of Agricultural Research 45:741-756.

Shelberg, M.C. and H. Moellering. 1983. IFAS: a program to measure the fractal dimensions of curves and surfaces, pp. 483-492 in: *Technical Papers of the 43rd Annual Meeting of ACSM*, Washington, D.C.

Shumway, Clyde. 1986. Summary of Forest Service GIS activities. pp. 49-52 in: *Proceedings of Geographic Information Systems Workshop.* American Society for Photogrammetry and Remote Sensing, and American Congress on Surveying and Mapping, Falls Church, Va. 426 pp.

Sieg, Gregory E. 1988. Integrating geographic information systems and decision support systems. pp. 901-910 in: *GIS/LIS '88 Proceedings*, American Society for Photogrammetry and Remote Sensing, and American Congress on Surveying and Mapping, Falls Church, Va. 980 pp.

Slama, Chester C., ed. 1980. *Manual of Photogrammetry, Fourth Edition*, American Society of Photogrammetry, Falls Church, Virginia. 1056 pp.

Smith, James L. 1987. Evaluation of the effect of photo-measurement errors on predictions of stand volume from aerial photographs. Photogrammetric Engineering and Remote Sensing 52(3):401-410.

Snyder, John P. 1982. *Map projections used by the U.S. Geological Survey.* Geological Survey Bulletin 1532. U.S. Government Printing Office, Washington, D.C. 313 pp.

Solomon, H. 1978. *Geometric Probability.* Philadelphia Society for Industrial and Applied Mathematics, 174 pp.

Sonnenburg, Duane. 1988. The role of the legal land parcel in land information systems/land information management. pp. 152-158 in: *GIS/LIS '88 Proceedings*, American Society for Photogrammetry and Remote Sensing, and American Congress on Surveying and Mapping, Falls Church, Va. 980 pp.

Stephens, M.A. 1974. EDF statistics for goodnes of fit and some comparisons. Journal of the American Statistical Association. 69(347):730-737.

Story, Michael and R.G. Congalton. 1986. Accuracy assessment: a user's perspective. Photogrammetric Engineering and Remote Sensing 52(3):397-399.

Switzer, P. 1975. Estimation of the accuracy of qualitative maps, pp. 1-13 in: *Display and Analysis of Spatial Data*, Davis & McCullagh, ed., Wiley, New York.

Switzer, P., and A. Venetoulias. 1987. Spatial classification error rates related to pixel size, pp. 221-239 in: *Contributions to the Theory and Application of Statistics*, Alan E. Gelfand, ed. Academic Press.

Tatsuoka, M. M. 1971. *Multivariate Analysis- Techniques for Educational and Psychological Research.* John Wiley & Sons, New York. 310 pp.

Thompson, L.G.S. 1981. Digitizing and automated output mapping errors. Photogrammetric Engineering and Remote Sensing 47(10): 1455-1457.

Thompson, Morris M. 1979. *Maps for America- Cartographic Products of the U.S. Geological Survey and Others*, U.S. Government Printing Office, Washington, D.C., 265 pp.

Timber-Mart South. 1989. Vol 14, No. 1, First Quarter, 1989. Highlands, N.C.

Tomlinson Associates. 1985. *Advanced Geographic Information Systems Workloads Analysis- George Washington National Forest Individual Forest Report.* Tomlinson Associates,

Traylor, C. 1979. The evaluation of a method to measure manual digitizing errors in cartographic data bases. PhD dissertation, University of Kansas, Lawrence, KS. 117 pp.

VanRoessel, Jan W. 1988. Conversion of cartesian coordinates from and to generalized balanced ternary addresses. Photogrammetric Engineering and Remote Sensing 54(11):1565-1570.

Vitek, J.D. and D.G. Richards. 1978. Incorporating inherent map error into flood-hazard analysis. Professional Geographer 30(2):168-173.

Vitek, J.D., S.J. Walsh, and M.S. Gregory. 1984. Accuracy in GIS: an assessment of inherent and operational errors. Proceedings of Pecora IX Symposium. pp. 296-302.

Walsh, Stephen J., Dale R. Lightfoot, and David R. Butler. 1987. Recognition and assessment of error in geographic information systems. Photogrammetric Engineering and Remote Sensing 53(10):1423-1430.

Weibel, Robert and Barbara P. Buttenfield. 1988. Map design for geographic information systems. pp. 350-359 in: *GIS/LIS '88 Proceedings*, American Society for Photogrammetry and Remote Sensing, and American Congress on Surveying and Mapping, Falls Church, Va. 980 pp.

Wenger, Karl E., ed. 1984. *Forestry Handbook, Second Edition.* John Wiley & Sons, New York. 1335 pp.

Wiles, S. 1988. Evaluation of photographic properties for area estimation. MS thesis, Department of Forestry, Virginia Polytechnic Institute and State University, Blacksburg, VA. 100 pp.

Wolf, Paul R. 1974. *Elements of Photogrammetry.* McGraw-Hill, Inc. New York. 562 pp.

Yoeli, P. 1984. Error-bands of topographical contours with computer and plotter. Geoprocessing 2:287-297.

The vita has been removed from the scanned document