South Dakota State University

Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange

Electronic Theses and Dissertations

1978

# Application Analysis of a Cellular Geographic Information System

Michael Edward Wehde

Follow this and additional works at: https://openprairie.sdstate.edu/etd

 Part of the Electrical and Computer Engineering Commons

## Recommended Citation

Wehde, Michael Edward, "Application Analysis of a Cellular Geographic Information System" (1978). *Electronic Theses and Dissertations*. 5634.
https://openprairie.sdstate.edu/etd/5634

This Thesis - Open Access is brought to you for free and open access by Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. For more information, please contact michael.biondo@sdstate.edu.

# APPLICATION ANALYSIS OF A CELLULAR GEOGRAPHIC INFORMATION SYSTEM

BY

MICHAEL EDWARD WEHDE

A Thesis submitted
in partial fulfillment of the requirements for the
degree Master of Science, Major in
Engineering, South Dakota
State University

1978

SOUTH DAKOTA STATE UNIVERSITY LIBRARY

APPLICATION ANALYSIS

OF A CELLULAR

GEOGRAPHIC INFORMATION SYSTEM

This thesis is approved as a creditable and independent investigation by a candidate for the degree, Master of Science, and is acceptable for meeting the thesis requirements for this degree. Acceptance of this thesis does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

APPLICATION ANALYSIS

OF A CELLULAR

GEOGRAPHIC INFORMATION SYSTEM

Abstract

MICHAEL E. WEHDE

Under the supervision of Professor Robert Finch

AREAS, the Area REsource Analysis System, was developed by the author as an analytical tool for resource data manipulation in a spatial context. The data base and processing software design factors relevant to economical application are reviewed.

The investigation seeks to define a procedure for selection and justification of the grid cell size used to quantify the spatial domain. The performance of a cellular information system is measured in terms of mapping and inventory accuracy. These measures of performance are evaluated for various cellularizations of a single map. The interboundary distance distribution is justified as the only unique characterization of the map and a mathematical model for converting this distribution into an estimate of mapping accuracy is evaluated.

## ACKNOWLEDGEMENTS

The author wishes to express appreciation to Mr. Victor I. Myers, Director of the Remote Sensing Institute and the institute staff for enabling the research to be performed. The work was supported by the NASA research grant number NGL-42-003-007. The comments of Dr. Robert Finch of the Electrical Engineering Department were most helpful.

Consultations with Mr. Delmar Johnson, the Systems Programmer, at the SDSU computer facility were invaluable to the successful implementation of AREAS. The author is grateful to the following people: Lynette Nelson, for patience in typing repeated adjustments to the text; Mary Buckmiller and Kathy Kitzmiller, for graphic arts assistance in figure production; and Jan Griesenbrock, for timely production of photographic products during the course of the investigation and during the production of figures. The author also recognizes the cooperation of his family and the encouragement of his wife, Pat, as most contributory to the completion of this work.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

LIST OF FIGURES

LIST OF FIGURES (CONTINUED)

LIST OF TABLES

# CHAPTER 1. INTRODUCTION AND LITERATURE REVIEW

## 1.1 Introduction

Digital hardware evolution and software innovation have contributed to increased storage capacity, high speed processing and improved cost-performance ratio in data processing applications. Continuing improvements in storage, speed and costs of digital systems broadens the spectrum of applications. Present technological capabilities have well surpassed the minimum requirements for useful processing of geographically oriented, spatially distributed, data sets.

Spatial data sets carry an added information dimension, location. Whether explicitly or implicitly contained in the data set, locational information allows interaction and display in a map reference framework. It is primarily this feature which distinguishes geographic data sets. Methods for storage, retrieval and manipulation of such data sets, while preserving the location information, vary in complexity, accuracy and cost.

The inability of man to mentally cope with the complexity of processing or physically cope with the masses of data involved motivated investigations of computer processing. Other motivations have been noted. Nichols [1] pointed out that computerization of resource maps allow automated reduction of many levels of map information down to a few classes, suitabilities or potentials according to the specific purpose of the analysis and, thereby, restores human visualization to the problem. McDonald [2] cites an

ever growing requirement for increased rates of delivery. Either situation alone would be cause for pursuit of computer capability.

Computer systems do have the capability. Over the last two decades software packages have proliferated in response to the combination of available hardware and specific problems to be handled. Radlinski [3] reports that a United States Geological Survey cooperative, 2-year study with the International Geographical Union Commission on Geographical Data Sensing and Processing found fifty-four land or resource related data bases or systems existing within the United States Geological Survey alone. At that time over 500,000 megabits of information were involved and the data were expected to triple by the year 1980. On the national and international scene software variations abound, each with an identifying acronymn as claim of uniqueness. Examples which are identified later are LUNR, LUIS, MIADS, ORRMIS, GRIDS, CMS, NARIS, DIME, MLIS, GRDSR, MIDAS, GIST, FRIS, CGIS, MAP/MODEL, PILS, and STORET [4]. Tomlinson [5] cites AUTOMAP, SYMAP, CAN-HYDRO, SYMVUU, GIMMS, OEM, NCC, ARDS, CENSUS, OBLIX, and WWW. Bryant [6] offers IBIS as yet another alternate approach.

## 1.2 Geographical Information Systems

Although there is disagreement between geographic information system users and the specialists who develop the technology as to whether or not the data base should be considered a part of the "system", the data base does influence decisions about the hardware and software system development. As a significant factor in the

design and operation of the information system, the data base must receive careful attention.

### 1.2.1 Data Base Characteristics

All computer data sets are collections of data. The entries may be present measurements of a variable or identifications of class membership or theme level. There is in addition at least an implicit reference to the time or date of observation. The relevance of this time datum is dependent on the dynamics of the theme or variable.

Spatial or geographical information processing, however, deals in data sets, often termed data bases, which contain three items of information. Together with the theme or variate value and the time reference, spatial location is included [7]. Steiner [7] lists the factors which determine the spatial characteristics of the data base as (A) the area span (geographical extent or coverage of the data base), (B) the spatial resolution (minimum size cell or spatial unit represented), and (C) the spatial frequency (the occurrence of the resolution elements throughout the coverage, from a sampling system to a complete representation). He further points out that similar characteristics may be defined for the time or date datum, i.e. temporal span, temporal resolution, and temporal frequency. These characteristics of course only apply to data bases containing a theme or variable repeated at multiple time references as would be the case in studies of time dynamics.

Clearly the content variation in data bases discussed to this point is limited to multiple theme or variable data at a point in time, multiple time samples of a theme or variable, or both. The intended analyses guide formulation of the composition of the data base and also dictates to some degree the development of the information processing system. This is the first diverging influence which propogates new processing systems.

Determining data base content in light of purpose does not necessarily determine the measurement scale for the datum points. Steiner [7] describes several scales to choose from. First, a nominal scale is typically binary and simply indicates presence or absence of a theme. Second, an ordinal scale places observations in a size sequence, ranking or ordering of the theme or variable value. Third, a continuous scale is usually associated with a continuous variable and may be either interval or ratio type. Interval continuous scaling has an arbitrary zero and maintains data value differences, e.g. temperature measurements. Ratio continuous scaling has a physically meaningful absolute zero which then allows analysis by ratios of data values as well as differences among values, e.g. rainfall measurement. In addition the conception of a thematic scale is an extension of nominal scaling to multiple levels not of an ordinal or continuous nature.

The application or intended analysis which guides selection of data base content variables or themes will also typically guide selection of the measurement scales. The variety of data scales,

however, has greater impact on the processing system than the data base content discussed earlier. This is the second diverging influence which motivates creation of new processing systems.

In geographic information systems, data base design decisions of content and scale are subordinate to consideration of the alternatives for representation of the locational information. In fact geographic information systems are classified into types based primarily on the location reference element of the data base. Five of the most commonly recognized types according to approach to location reference are (1) the uniform grid which divides the space into an x-y cellular network, (2) the parcel in which spatial subdivisions arise in either natural or artificial (political) boundary context and are irregularly shaped, non-uniformly sized or both, (3) the area boundary in which addresses of the enclosing border are stored in conjunction with the data for the enclosed region, (4) the network in which lines connect nodes together in a spatial mesh or net, and (5) the point where spot spatial addressing places the data record into the overall spatial domain but the data values have no single uniform spatial coverage around the specific site [4]. Bryant [6] suggests a sixth type, image format which is a sequential raster of pixels and greatly resembles the uniform grid in spatial philosophy but not necessarily in processing methodology.

Tomlinson [8] proposes four types of geographical data bases according to the handling of the locational identifiers. He suggests (1) external index, (2) coordinate reference, (3) arbitrary grid,

and (4) explicit boundary. The first is identifiable by data sets containing no direct reference to location nor physical arrangement of the data which might imply spatial location. A separate index file contains the geographic locations and address pointers to identify the position of the data in the data base. The second system carries geographic coordinates together with the data entries. The third system divides the spatial domain into a regularly spaced, cellular network and the data are stored in such a manner as to preserve the relative location of these cells. The fourth system stores specific locational coordinates for the boundaries which enclose a homogeneous region. Although all of these types of data base accomplish the task of maintaining geographic or spatial reference, it is particularly important to note that only the latter two actually store and preserve boundary information. Hence only these two systems lend themselves to spatial displays of boundaries [8].

This preservation of boundary information for purposes of spatial display is precisely in line with a prime motivation of geographical information processing, i.e. mapping of information. For this reason only two types of information systems are commonly recognized. These are the grid encoding and line encoding approaches [9].

It is quite evident that the approach to handling the locational identifiers in a geographical data base, has the greatest impact on the processing system of any of the factors discussed. This is the third and most strongly divergent influence on development of data base systems.

With data base contents, analysis purposes, measurement scales, and geographical reference approaches varying from application to application and, furthermore, the idiosyncrasies of specific hardware configurations confounding the problem, the reason for the proliferation of data processing systems becomes apparent.

### 1.2.2 Information System Examples

Examples of systems employing the uniform grid approach to geographic reference are as follows: LUNR - Land Use and Natural Resources by the Office of Planning and Coordination for the State of New York and Cornell Center for Aerial Photographic Studies; LUIS - Land Use Information System by the University of Massachusettes; MIADS - Map Information Assembly and Display System by the U.S. Forest Service within the U.S. Department of Agriculture; ORRMIS - Oak Ridge Regional Modeling Information System by the Oak Ridge National Laboratory; GRIDS - Grid Related Information Display System by the Southern California Regional Information Study; and CMS - Composite Mapping System by the Economic Development Administration within the U.S. Department of Commerce [4].

Examples of geographic information systems employing the parcel approach to location are as follows: NARIS - Natural Resources Information System developed at Center for Advanced Computation at the University of Illinois for the Northeast Illinois Natural Resource Service Center; DIME - Dual Independent Map Encoding by the Bureau of Census within the U.S. Department of Commerce; MLIS - Minnesota Land Management Information System Study

by the University of Minnesota; GRDSR - Geographically Referred Data Storage and Retrieval System by the Dominion Bureau of Statistics in Canada; GIST - Geographic Information System for the Office of the Mayor of New York City and FRIS by the Swedish Central Board for Real Estate Data [4].

Examples of geographic information systems employing the area boundary approach to spatial location are as follows: CGIS - Canadian Geographic Information System by IBM corporation for the Canadian Department of Regional Economic Expansion; MAP/MODEL by the University of Oregon for the Bureau of Governmental Research and Service; and PIOS (no further information available). The lone example of network approach is STORET (no further information provided) [4].

IBIS - Image Based Information System is a newcomer in processing approach although the storage method is actually a fine cell or uniform grid basis [6].

The multitude of systems is clear evidence of the usefulness of computer technology in the area of geographic information processing. Furthermore, the many systems exist as a result of the influences discussed earlier which tend to generate new systems to meet new situations.

### 1.2.3 System Studies and Recommendations

From a data processing or technological viewpoint a system is usually considered to be the combination and interaction of hardware and software. The typical consumer of geographic information would

like to approach the system with only questions and analyses in mind, hoping to find both the processing capability and the requisite data sets available. Thus, from a user viewpoint in geographical information systems, there would appear to be merit in including the data base as part of the information system.

Steiner [7] cites data base development as a five phase activity; data specification, data acquisition, data storage/retrieval/manipulation, data dissemination, and data applications. These activities, if included as a part of the information system, would promote continued development of new systems for new problems as the users and their data attempt to interact with the storage and processing capabilities of their hardware/software.

There are many opinions as to the content and developmental needs in generating useful geographic information systems.

Bryand and Zobrist [10] offer four criteria which they believe must be satisfied to make a geographic information system "useful". These criteria are (1) that point and area locations are provided, (2) that variable aggregation or sub-setting be possible, (3) that there be representation of spatial arrangement of the data, and (4) that data interface with mathematical and statistical analysis programs. These criteria are indicative of recognized needs for spatial manipulation and analytical capabilities in a useful system.

In a feasibility study for the Illinois Resources Information System, six areas were deemed important to the creation of a viable information system [4]. These were as follows: (1) that the system

involve a large user community, (2) that point, network, and area data exist in one system, (3) that multiple data bases and resolutions be employed within the system, (4) that multiple problem-oriented user interfaces or terminal languages be provided, (5) that multiple data entry facilities spanning a variety of forms and formats of data be provided, and (6) that advanced graphics output capability be included.

The United States Geological Survey in cooperation with the International Geographical Union's Commission on Geographical Data Sensing and Processing conducted a two-year study of spatial data systems. W.A. Radlinski, the Associate Director of USGS, in his opening address to the 1977 ASP-ACSM Annual Convention reported on specific findings of that study as follows [3]: (1) many seemingly simple tasks require development, (2) storage schemes need to consider large files up to terrabits ($10^{12}$), (3) the digitization and edit function must be made more economical, (4) one system to handle point, line, area, network and image data is needed, (5) topological structure in data needs to be understood, and (6) storage/retrieval methods far exceed analytical/interpretive methods. His summary of cooperative developmental areas called for (1) establishing standards on scale, resolution, accuracy etc. by means of user survey, (2) sorting the potential or candidate data to be maintained in a system into priority by usefulness, (3) coordination of the identification of theoretical research needs, (4) cooperative hardware development, (5) technology forecasting to influence system

development planning, and (6) establishing a mechanism for institutional communication and coordination. His review and recommendations represent the broadest and most progressive noted in the literature.

The Illinois Resource Information System feasibility study, being somewhat older, was less extensive in its data base design recommendations [4]. It called for optimization of high-speed retrieval and analysis functions to support large quantities of data rather than for development of update/insertion activities, and also recommended multiple resolutions in the data sets with processing capability to change among them as needed.

Phillips [9], although biased toward value of graphics, did make the significant observation that computer graphics must serve the role of reducing "impedance of the interface" between the data and the users.

Common threads of meaning in the recommendations cited appear to be (1) concern for the user and his interests and (2) concern for controlled, progressive, positive developmental effort in geographic information systems.

## 1.3 System Comparisons and Evaluations

Tomlinson [8] emphasized in his review of geographical data bases, that only cellular geographic reference and boundary coordinate types of data bases contain the boundary information necessary for spatial display in mapping form. These information systems are referred to as cell and polygon systems. Image based information

processing could also be included as a cellular type. These three primary systems are employed for their mapping capability.

### 1.3.1 Data Organization

The processing systems are markedly influenced by the data organization and can be characterized by the approach used to handle the geographic or spatial reference.

With the image raster approach the spatial or x-y position is implicitly recognized by the scan position within the image [10]. The processing approach is described as image manipulation [6]. A polygon data organization encloses each homogeneous spatial region with a polygon. The x-y coordinate pairs for the curved line segments which enclose the polygon are recorded.

The grid referenced or cellular approach divides the spatial domain into uniformly spaced and sized cells (a grid). The relative position of the cells in an array preserves spatial relationships without allocating storage to the location information [11]. Steiner [7] characterizes the grid based coding schemes as (1) complete - x and y coordinates and datum recorded for each geographic point, (2) sequential - only datum values recorded in sequence, and (3) compact sequential where string length or change point methods compress the sequence. String length coding involves a datum and a repeat factor while change point coding involves an initiator coordinate and a datum which applies until the next initiator coordinate. All three grid based schemes accomplish the same geographic referencing but processing economics differ.

The cellular and polygon approaches will be compared further. The image approach is the newest and least documented in comparative performance and costs. By virtue of its similarity to a full-matrix cellular approach, the image form will be dropped from discussion - realizing that continuing development in this area may warrant a separate review.

### 1.3.2 Cellular Versus Polygon System

Smith, Van Gorkom, Dyer [12] outline the evaluation of geographic information systems. They suggest evaluation of five attributes as follows:

(1) map characteristics

(2) analysis characteristics

(3) cost efficiency

(4) user convenience

(5) applicability to a state-wide system

Furthermore, they recommend twelve factors upon which to base judgement of effectiveness. These are as follows:

(1) the distortion of basic data

(2) output comprehensibility

(3) multiple map analysis capability by adjoining map segments

(4) multiple map overlay

(5) complex compositing with weighting factors

(6) user costs or costs per acre

(7) applicability to extensive analyses

(8) applicability to intensive analyses

338881 SOUTH DAKOTA STATE UNIVERSITY LIBRARY

(9) analysis flexibility

(10) ease of update

(11) ease of user access

(12) extendability to state-wide system.

These criteria are clearly established from a user viewpoint. Tomlinson [5] poses brief criteria of (1) data volume capability, (2) geo-reference method, and (3) manipulation or analyses capability. These appear to be evaluative criteria from the viewpoint of system development or implementation. Summary criteria might be the storage and analyses capabilities, corresponding costs and the product quality or accuracy.

Steiner [7] cites grid system simplicity in data handling and output phases as advantageous. Bryant and Zobrist [10] compare the geographic encoding approaches in general terms as follows. The grid cell method is manually operated with poor spatial resolution and difficult update capability. The polygon method is expensive for large data sets and inherently prohibits certain operations. Continuing to summarize the drawbacks of both systems, they cite three primary disadvantages of each as follows:

Grid cell - (1) spatial resolution tied to a cell

(2) data are nominal or ordinal but not both

(3) manual coding difficult/costly as is update process

Polygon - (1) editing is computationally expensive

(2) topological extraction of sub-areas is complex

(3) computer system for overlay of polygons is expensive.

These summaries are indicative of a possible single tradeoff decision between cost and accuracy when choosing between these approaches.

To acquire a meaningful set of operating cost data for comparison of the systems is difficult because only one system is typically utilized and reported. Smith [12] overcame this problem by utilizing a benchmark project to be completed by contracting with two agencies - each representing one approach. Soil, slope and geology maps at 1:24000 scale were provided for conversion to geographic data base and analyses. The grid system was operated on a 2.5 acre cell. To convert to data base, the costs were $342 for the grid system and $2825 for the polygon. Similarly the analyses specified cost $430 via the grid processor and $4520 via the polygon system. The cell or grid system is much less expensive to operate.

The costs of data conversion versus analysis by the grid system in Smith's study were on the same order of magnitude. Tomlinson [5] reports that data preparation costs for cellular systems usually run four to five times the manipulation and display costs. Factors accounting for this apparent discrepancy are the quantity of analysis done once the data are prepared and a high variation in analysis costs according to the specific storage approach, complete, sequential or compact sequential, within cellular

systems. The author has experienced cost ratios of 10-1 to as much as 20-1 in comparison of complete coding to compact sequential coding data base processing.

Operating cost factors appear to favor the grid system. There are criticisms of the grid system which warrant consideration. Smith [12] reports that grid type maps are criticized for quality of final product - especially where a line printer is used, and they are criticized for data precision in a grid framework. Both criticisms are answerable - the first by selection of another output media from among the ever increasing list of alternatives and both by utilization of small cells.

Phillips [9] and Sinton [13] provide a comparative summary. The cell system has grid coarseness as a limit on data resolution and the finer the grid the greater the storage requirements. It does work best and fastest on large or complex data sets because of data accessability and adaptability to many cell-oriented devices, line printers, film recorders, computer memory etc. The polygon system has excellent spatial definition and therefore a better map accuracy and product quality. The storage and retrieval phases are relatively more complex and difficult and the processing of data is more costly with the polygon system.

Sinton [13] notes that polygon users are typically cartographically oriented which requires the more precise ground locations while grid-cell users are resource analysts desiring manipulative ability and operating economy.

Smith [12] draws a conclusion in his study which well summarizes the comparative discussion here. The polygon system provides higher quality but is far too expensive. The cell system is operable and affordable.

### 1.3.3 Performance Evaluations

The matter of geographical information system performance may be approached from two viewpoints. On the one hand system analysts are concerned with storage and processing efficiencies in accomplishing the required analyses. On the other hand the consumer of information is concerned with the accuracy of the two primary analysis products, spatial inventories (tabulations) and spatial displays (maps).

On the surface these may seem very similar. Both viewpoints take similar avenues to evaluation, that is, they both monitor the system in terms of successful performance of analyses, accuracy of inventory, accuracy of mapping and corresponding costs. The former viewpoint, however, is evaluation for purposes of system design and implementation while the latter is evaluating applicability to a problem. Specific examples will reinforce this distinction. The grid system will be evaluated.

Switzer [14] outlined a mathematical approach to evaluation of the grid system. His evaluation was based on map accuracy. His primary purpose was system design as an analyst would view the system, although a later development in his work applies well to the consumer viewpoint also.

Switzer's evaluation of map precision was based on the degrees of discrepancy between the "estimated" map in hand and the "true" map as actual spatial relationships. In the set theoretic sense, the intersection operation defining the area of agreement would measure map precision in the spatial domain.

The coding method employed was the cell center datum. Whatever the cell size, the cell is considered homogeneous and of whatever data category or value occurred at the cell center. This differs from the cell dominant approach where spatial plurality within the cell determines assigned homogeneity. The cell size was recommended to be smaller than twice the smallest map region to be retained. This constraint arises obviously due to extreme error situations when many strata appear within one cell and central datum alone defines the result. When cell sizes are small so that map boundary lines on a cell basis look like linear segmentors, then there will be almost complete agreement of cell central datum coding and cell dominance coding techniques.

Switzer's mathematical formulation of the coding process as it relates to mapping error is diagrammed in Figure 1.1. A single cell is shown. Datum assignment would be to stratum j based on membership of central point B. In a mapping sense, an instant error arises due to omission of stratum i points from representation even though they are within the cell and a distance d away from B. The boundary line is drawn linear as would approximately be the case with small cells. In this situation cell dominance would have to agree with central datum in assignment of stratum j.

Geometric probability can be directly employed to evaluate $P_{ij}(d)$, the probability of stratum i points within a cell assigned stratum j and separated from the central datum by distance d.

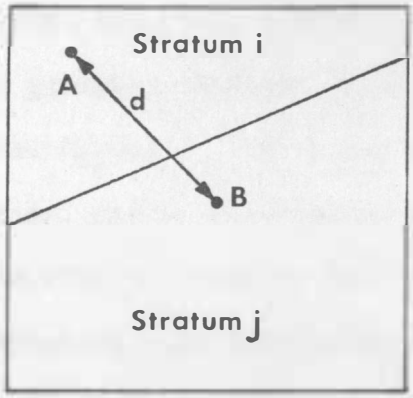



Figure 1.1. The central-datum cell coding concept as related to mapping error. Cell assigned to stratum j because of membership of central point B even though there is stratum i present in the cell at points such as A a distance d away from B.

Clearly the area mis-match between the coded cell - entirely stratum j and the original "true" cell is exactly the proposed estimator of mapping precision. Thus, map precision depends on $P_{ij}(d)$.

Switzer's first concern was evaluation of cell geometries. For a cell of given size, he reasoned, $P_{ij}(d)$ depends on the map and the cell geometry. Therefore, it is only necessary to hold the map constant as well as the cell size and vary the cell geometry. Resultant changes in map precision would provide the evaluation. Performing precisely this investigation he concluded that square cells are superior to rectangular and are only slightly inferior to a hexagonal network where all adjacent cells are equally spaced. Also, he noted that map precision or accuracy increases at the same rate as the maximum cell dimension decreases (increasing number of cells).

This investigation is an example of evaluation aimed at system design. A cell geometry was selected for implementation. Costs were not considered.

Switzer continued his study with a user oriented evaluation. Holding the sampling geometry constant, i.e. square cell of given size, he reasoned that $P_{ij}(d)$ is then a map property. The estimated map (after digitization) can be analyzed to predict the actual property $P'_{ij}(d=o)$ and that property in turn used to predict the map precision. This predictive sequence claims the evaluative distinction of measuring map precision from the estimated map alone without having the "true" map as reference. This is a commendable effort in mapping evaluation and the only assumption made is that map data are fairly dense.

The drawback to this evaluation is that the estimated map, i.e. the digitized representation, must exist before the map precision can be predicted. If the precision turns out unacceptable to the user, he is forced to redigitize to achieve a new estimated map. In operating characteristics noted earlier, cellular systems involve as much or several times more investment in the data entry than in analysis. Hence an iterative digitization and evaluation is not an economical approach to performance oriented digitization. What is needed is a simple map analytical method for performance prediction which can be usefully employed before digitization of maps begins.

Users have attempted crude evaluations of the grid system based on spatial inventory performance. Via this approach the "true"

map is tabulated by areas of category or theme and compared to the tabulation achieved with the geographic information system. Nichols [1] utilized soil maps of low, medium and high cartographic detail and cellularizations of 8.64 ha (21.33 acres), 16.20 ha (40 acres), and 64.80 ha (160 acres) to evaluate percentage tabulation performance versus the true, input map areas. The comparison was based on cell counts from the grid information system versus a published survey generated by dot sample counting on the maps. He reported performance versus cartographic detail as 71.6%, 64.4% and 41.3% for low, medium and high detail. Looking further at the medium detail map, he reported 70.5%, 64.4% and 48.8% for cells of 8.64 ha, 16.20 ha, and 64.80 ha. He felt that the accuracy was not acceptable for his purposes.

It should be noted that in Nichols' study, the most critical performance factor, cell size, was arbitrarily chosen. Again no pre-digitization guideline was available for selection of the grid cell size.

Recent articles have suggested probablistic models for the evaluation of match between a computer output map and reality. Recall this is exactly Switzer's approach to evaluation of map precision. Therefore,these models may also prove useful in evaluation of geographic information systems.

Van Genderen's view of land use map accuracy parallels Switzer's map precision. Cell by cell match of the computer map to actual ground conditions measure accuracy. He noted, however, that

total comparison is often physically and economically impossible. A sampling technique is proposed. He suggests using stratified random sampling by category and a binomial probability model for the two situations, match and mismatch. His concern was how many samples, N, should be taken if under probability of error, p, one does not wish to risk a chance result of no errors in the sample. The objective of the sampling in the first place is to estimate p, but he is concerned with overconfidence due to misleading "perfect" scores.

Van Genderen's discussion is actually a specialized application of Hord's [16] earlier work along the same line. Hord proposed the binomial approach and expounded on use of the normal distribution to calculate confidence intervals on the estimate of p. In a broader context he also applied this model to not only the cell by cell matching, but to boundary line comparisons and to control point location comparisons of the map in hand to the actual spatial relationships. Since class accuracy, boundary accuracy and control point accuracy need not numerically agree, he also proposed RMS average as a single figure for map accuracy.

It should be noted that the Hord-Van Genderen probablistic mis-match model, like Switzer's output map analysis, requires the product map be generated before evaluation can be made. To reiterate, this is not useful for judgements prior to digitization. The missing link is analysis of the true map prior to digitization to enable predictions of mapping and inventory accuracy, with varying cell size.

## 1.4 The Remote Sensing Institute System

Earlier comparisons of grid cell and polygon systems came to the conclusion that the grid cell is operable and affordable. The Remote Sensing Institute staff, primarily resource oriented, sought manipulative and analytical capability and operating economy.

Another factor which contributed to grid-cell selection was recognized by Smedes [18]. He noted among the multitude of data sources for a geographic information system - social, economic, remote sensing etc. - that some data forms are directly cell oriented. For example, multispectral scanners aboard aircraft and spacecraft and digital image matrices are sequential cell oriented.

The first geographic information system utilized became a grid cell system. The author, who developed the present system, foresaw an implementation strategy which provided improved operating economy and storage efficiency [17].

Complete, sequential and compact sequential cell coding approaches were reviewed earlier. In addition to these alternatives for data organization, there is a choice of record storage approach which impacts storage and operating efficiencies. The choices are (1) low efficiency where fixed length records are used to store non-compressed data, (2) medium efficiency where (a) variable length records store non-compressed data or (b) fixed length records store compressed data and (3) high efficiency where variable length records store compressed data [4]. The author developed a system utilizing variable length blocked records to store compact sequential data in

sub-word formats [19]. A more detailed system description will be provided in Chapter 2.

## 1.5 Grid Cell Size Selection Bases

To access the analytical, manipulative and display capabilities of a geographic information system, the user must provide for conversion of his map or spatial data into the format of the system data base. Users of a grid cell information system face the obvious and extremely significant question of appropriate cell size. Crude guidelines do exist in the literature. Suggestions are as follows:

(1) use the resolution of the source data [11,20]

(2) consider processing capability and cost and use finest grid affordable [7]

(3) use grid appropriate to the particular data application [20,21]

(4) grid-cell size selected according to data detail such as urban at 10 meter, urbanizing at 250 meter, other at 25 kilometer [7]

(5) grid-cell selected according to output device i.e. if line printer, use rectangular to offset aspect ratio [7,2]

(6) grid cell size small enough to force smallest spatial unit on the map to be more than 50% of the cell [9,14]

(7) consider the volume of data generated for processing [11].

These suggested guidelines for cell size selection do not appear to be based on knowledge of accuracy or performance capability of a

grid system. There is a hint that a cell size-accuracy relationship is appreciated, particularly in suggestion number 2.

With data conversion (digitization) under a grid oriented system being a significant part of the operating cost, the user might well hesitate to select a very fine cell spacing. Meyers, Durfee and Tucker [20] make the very valid point, however, that use of a smaller grid may generate more cells in the encoding process but may still save time by making dominant theme assignment trivial compared to large cells covering several themes. It is nevertheless recognized that a significant factor in cell size selection is the time and cost of the coding method used [11].

## 1.6 Research Objectives

All users of geographical information systems have a vested interest in learning more about appropriate application of the system. The effectiveness of a system can be judged in terms of application costs and product accuracies. With a grid cell system both of these measures depend on the cell size selected. The literature reviewed represents an effort to define a cell size selection and an effort to evaluate mapping and inventory products based on system/map characteristics.

The objective of the research herein reported is to overcome two serious shortcomings in the work to date - namely (1) that mapping and inventory accuracies have only been analyzed/modelled after the fact not prior to digitization of the data base and (2) that cell size selection criteria have not been based on the cost-accuracy

tradeoff which appears to be present. These shortcomings are directly interrelated. The missing factor is performance predictability given only the proposed input map. The desired outcome of the research is a procedure for analyzing candidate map data for parameters which will assist the user in deciding on cell size.

There are two levels of completion of the objective. Cell size selection criteria may be generated for application of the Remote Sensing Institute system and/or a cell size selection assistance procedure can be defined for all users of grid-cell information systems. The difference between them is the system operating costs which will vary from system to system. The performance prediction procedure should hold for all cell-dominance coded, cellular geographic information systems regardless of the analyses they perform. The research effort in performance prediction is applicable to both levels of completion of the objective.

The route of investigation is to answer the following questions:

(1) Do mapping/inventory product accuracies relate to cell size given a data set?

(2) Do characteristics of the input map vary predictably with cell size?

(3) Which characteristics are useful in accuracy prediction?

(4) What procedure or technique might be employed to predict accuracy for a given map?

## 1.7 Preview of Contents

In the remainder of this paper the following topics will be discussed.

Chapter two will review the particular geographic information system developed by the author. The design and implementation were guided by convenience and economical operation criteria. The processing capabilities are also reviewed.

Chapter three discusses measures of performance of the system, a detailed performance experiment, and results. Consideration is given to the effect of cell size on the map in general as opposed to individual mapping units.

Chapter four explores interboundary distances, the distribution of relative frequencies of these distances, the distribution behavior with changing cell size, horizontal-vertical directional bias, and the distribution mean as an estimator of mapping error.

Chapter five analyzes the effect of alternate positions of the grid network with respect to an interboundary distance. A positional average error model is derived and tested.

Chapter six reports the undertaking of a model correction due to inherent adjacency of interboundary distances in any map transect. A correction procedure is derived and implemented for comparison to the results of the positional average model in Chapter 5.

Chapter seven summarizes the contents of this paper, the results obtained, the conclusions drawn and elaborates on additional courses of further investigation yet open for pursuit.

# CHAPTER 2. THE GEOGRAPHIC INFORMATION SYSTEM

The cellular geographic information system developed for the Remote Sensing Institute by the author is identified by the acronym AREAS which stands for Area REsource Analysis System [17,19]. This chapter is devoted to an overview of that system.

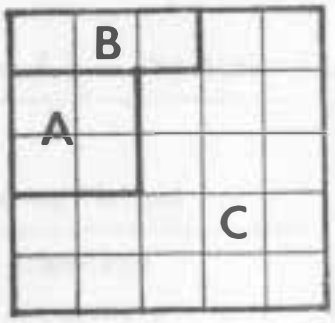
System design criteria were storage and processing efficiencies and operating convenience. Achievement of these objectives would promote economy in the applications environment.

## 2.1. The Data Base

The decision which dramatically impacts storage and processing efficiency of the system is selection among the alternatives for data base organization. Three factors are important, namely, the coding scheme, the storage assignment and the record organization within the file.

The introductory discussion of grid cell coding approaches in Chapter 1 only briefly reviewed the complete, the sequential and the compact sequential coding schemes. A more explicit comparison of these approaches is made in Fig. 2.1.

The impact on storage requirements should be evident. As an example consider a tabulation of the characters (digits and letters) required by each method for the data shown in Figure 2.1 and for the same data coded at twice the spatial resolution. See Table 2.1. The advantage of method 3 lies in compact representation of homogeneous




(a) gridded map segment

1,1,B,1,2,B,1,3,B,1,4,C,1,5,C,2,1,A,2,2,A,2,3,C,2,4,C,2,5,C,3,1,A,
3,2,A,3,3,C,3,4,C,3,5,C,4,1,C,4,2,C,4,3,C,4,4,C,4,5,C,5,1,C,5,2,C,
5,3,C,5,4,C,5,5,C

(b) complete coding

B,B,B,C,C
A,A,C,C,C
A,A,C,C,C

(c) sequential coding

C,C,C,C,C
C,C,C,C,C

1,1,B,4,C
2,1,A,3,C
3,1,A,3,C

(d) compact sequential coding

4,1,C
5,1,C

Figure 2.1. Cell-oriented coding schemes. (a) The spatial data gridded for input, (b) the complete coding approach, (c) the sequential approach, and (d) the compact sequential approach by the change-point technique.

Table 2.1. Character storage comparison of coding methods.

| Coding Method | RESOLUTION | |
|---|---|---|
| | cell in Fig. 2.1 | 0.5 cell in Fig. 2.1 |
| 1. complete | 75 | 310* |
| 2. sequential | 25 | 100 |
| 3. compact sequential | 21 | 42 |

(* ten double digits involved)

spans of data. This method also allows the spatial resolution to be changed (doubled for example) with only a linear (doubling) impact on storage while the other methods increase as the square of the factor (quadrupled in this case). Storage saved is storage available for finer resolution gridding of the data set.

The comparison tabulated above is somewhat artificial. The data set was not large enough in spatial expanse to represent multiple coordinate digits, only single character themes were present and the boundary locations were extremely sparse and totally hypothetical. The bias of these factors, however, was conservative.

Amidon [22] carried out a more rigid investigation of data compression by a span oriented, compact sequential approach and concluded that (1) compression of resource data may vary from 3 to 95%, (2) compression depends on the detail in the data and (3) direction of the span coding has negligible bearing on the compression achieved.

Under the design constraint of storage efficiency, the selection of a compact sequential coding approach was almost

inevitable. The next matter of importance to storage utilization was the assignment of storage for the data.

McDonald [2] documents storage savings of 85% in a sequentially-coded, 9 level, 100-cell-square matrix by not allocating full 60 bit words (a CDC 6600 computer) to each cell. This savings was totally accomplished by utilizing sub-words for each cell. It is only fair to note however that the savings is dependent on the word length of the machine being utilized. The example does clearly encourage consideration of sub-words in designing storage allocation.

Sub-word storage was adopted in the data base design for AREAS. Capacity limits were arbitrarily defined as needed throughout the design. Introducing a 256 level theme limit allows 8 bits or one computer byte to represent the datum for each homogeneous span. For a typical 32-bit, 4-byte word length computer, this represents 75% storage savings for the datum values. Limiting row and column identifications to half-words of 16 bits would accommodate spatial grids of 65,536 elements square. This would save 50% of the storage required for geographical identifiers.

The third factor in storage design of the data base is the organization of the records within the data file. The change-point coding scheme naturally gives rise to variable length records. To be consistant with the storage and processing efficiency objectives, this must be the choice of record format. Blocking of the records must also be considered in establishing the record format. While the record is the logical quantity of information transferred by a program

operation such as read or write, the block is the physical quantity transferred by the operating system over the data channel to or from the peripheral storage device. When block sizes are larger than the record length, computer buffer memory and the operating system work out the housekeeping. The beneficial result of buffering via blocking is fewer physical input-output, IO, operations required to transfer all the records of the data set.

Blocked records were chosen for AREAS design to promote processing efficiency. The blocks were defined as large as possible to permit fewest physical IO operations in processing AREAS data sets.

Characteristics and limits of the physical storage media must also be dealt with in specifying the record format. For reasons of capacity, convenience of access and operating system assistance in security, integrity and data management, disk storage was chosen over tape storage. Physical units of disk storage are the track and cylinder. A track is the path of one read/write head position over one recording surface during rotation of the disk. With several surfaces in parallel and simultaneously positioned read/write heads, the parallel tracks form a cylinder. The matter of record format specification should consider the capacity of a track in comparison to the block size, i.e. the largest quantity of data to be physically transferred in one operation. Having specified a maximum block size will likely involve physical transfers of data blocks larger than the track size. Specification of track overflow is necessary to cause track filling and appendage of tracks as needed to accommodate the data.

To summarize the data base design from storage and processing efficiency standpoints, the data is to be change-point, compact-sequential coding in sub-word format, stored in variable length, blocked, track overflow records on disk. The data base design is still not complete from the processing efficiency standpoint, however. Yet to be considered are the matters of minimum operating system handling of the data bits and the inclusion of identifiers and keys in the data base to automatically assist processing.

In the matter of data bit handling by the operating system, minimization was the objective. The minimum handling for input-output is direct copy without formatting of any kind. This direct verbatim transfer of the data bits is accomplished by unformatted read-write operations and unnecessary conversions to and from zoned decimal are avoided. The information, once it is converted to binary form, is maintained in that form.

In the matter of identifiers and keys within the data base, provision for identification of the data and automated interface to processing programs were the objectives. As leading information within each data set there is included a data set type identifier which traces the immediate history and nature of the data, a data set label which with proper utilization can provide subsequent recognition of the data set identity and/or intent, the physical extents which indicate to processor programs how many rows, columns, and codes (theme levels) are present, and the code table which lists all theme levels or codes. The position within the code table

somewhere from first to the two hundred fifty sixth position is the datum entry within the data records. This reduces four character alphanumeric codes to a number representable by 8 bits or a single byte. Although header entries were initially intended to automate processing by letting programs pick up control values directly from the data set, the purpose of these data base entries became access keys. Processing programs require user input of the values which then must match those of the data base for processing to continue.

## 2.2 Creating a Data Set

Two types of input data may be converted to the defined format. Manually coded card decks or pixel based image matrices may be converted.

For the manual process a map is available and a resolution cell size has been selected. A grid-generating, plotter program, RSIPLL, produces regularly spaced, rectangular or square grids at any scale up to the physical limitations of the plotter. By mounting a transparent drafting material on the plotter, an interpretation and coding overlay grid can be obtained. This grid when placed over the map provides assistance in assignment of the row and column geographic identifiers which accompany each datum.

The first processing program converts the coded data cards into the defined data base scheme. An assumption was made that row, column and code entries would be from one to four characters in length. The implied limit of 9999 rows or columns has not yet been anywhere near approached in single map entry. To eliminate constrained

format coding, a free-form input format was defined where row designations, column designations and data are separated by commas and an asterisk indicates end of row. The form is continuous, free of blanks within the row and, therefore, conservative of coding materials (forms, cards etc.). This format also eliminates alignment errors in coding or keypunch as there are no requirements on character positioning.

To facilitate implementation of the input program, CARDSIN, with the free-form input data, ASSEMBLER language coding was used. Character by character the data is deciphered as either data or control and the one byte and half-word data entries inserted into the data base record. Extensive error checking is built in with reporting of error locations and problem identifications throughout the data set. Situations such as missing rows, repeated rows, repeated columns, unidentified code entries, etc. are pinpointed to assist the user error correction process. By iterative correction and reprocessing eventually an error free run will be achieved which assures that a grammatically correct data set has been created. Interpretation and coding errors which simply misplace boundaries can only be eliminated by mapping the data set at a scale to directly overlay the original map and again iteratively correcting discrepancies.

The alternate input data format is image based. One word pixel matrices are assumed to correspond pixel to cell on one-to-one basis from the image to the desired data base. The program, CELLSDB, was

written to convert Landsat satellite computer compatible tape information in a thematic, computer classified format to the data base format.

Both programs generate a one page report of data set characteristics which can be directly inserted in loose-leaf form into a project or data set record book. This form of direct record keeping is intended to minimize operator housekeeping efforts. Either approach results in a type one data set as outlined in Figure 2.2.

Each square is one byte of computer storage. The first three records are the housekeeping keys and identification entries. The remaining record pairs define rows of the grid network. The byte and half-word (2 byte) compactions within the data records are evident.

## 2.3 Processing Capabilities

Although the design criteria and approach to implementation of AREAS heavily emphasized system and technical considerations impacting efficiency, the developmental sequence for processing capabilities was driven by user demand. Three basic analytical capabilities were deemed essential. These are discussed briefly in developmental sequence in the following paragraphs.

The interpretation process is a mapping of n levels or codes in an existing data set into n or fewer levels in a new specialized data set. Examples would be the mapping of soil association data

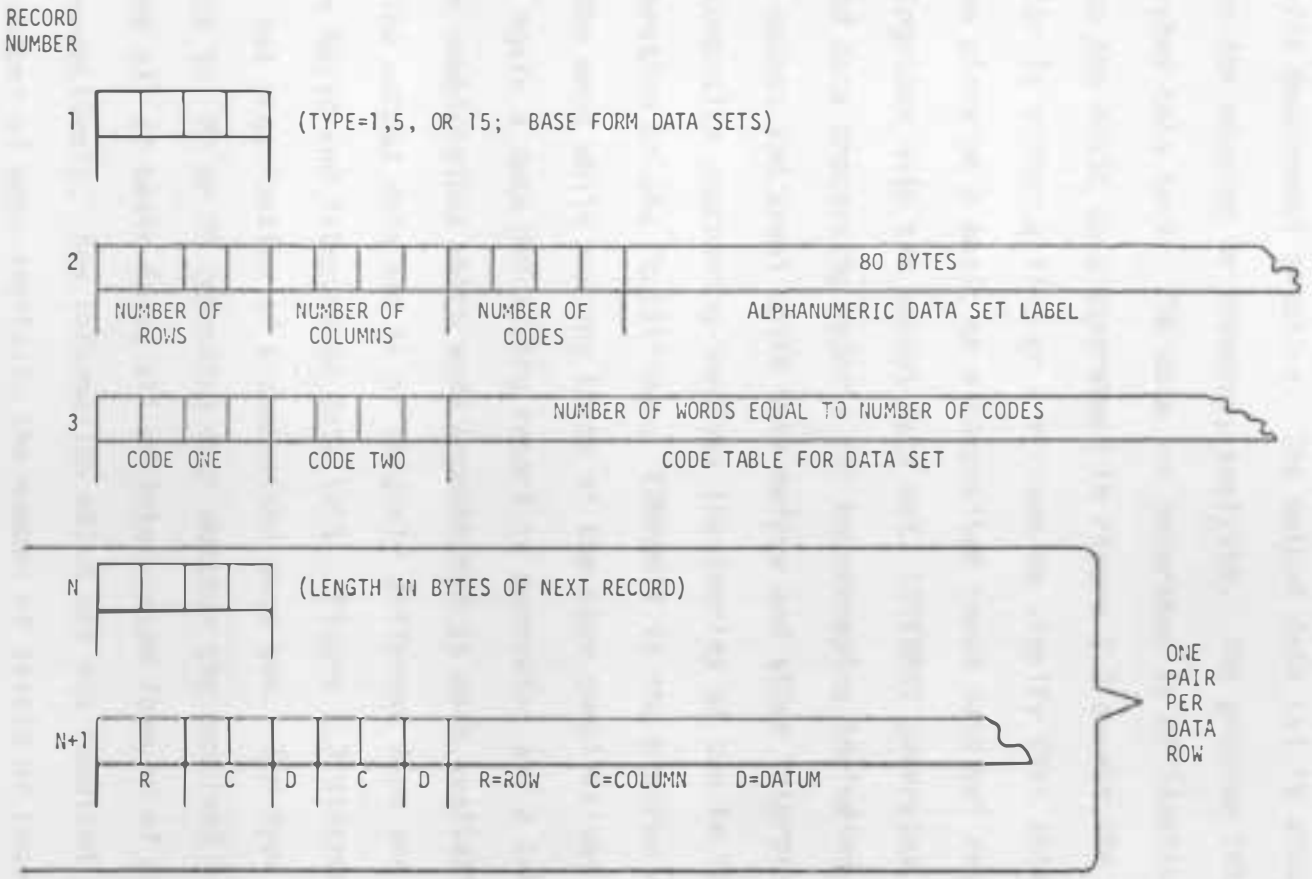



Figure 2.2. Organization of the records in the basic data set of AREAS.

codes into levels of irrigation suitability, construction limitation or specific management practice. The output data set is unique and available for mapping or inventory analyses. The program INTERPRT accomplishes this task. The data set generated is of identical format to the basic data diagrammed in Figure 2.2 except the type identifier is either a five or a fifteen to signify that interpretation has taken place on a basic or a composited input data set respectively.

Together with the output data set, INTERPRT generates page formatted data processing reports for housekeeping including tables of cell counts and areal units both before and after interpretation.

Composite processing involves the overlay of two to four maps and generation of one "total" map. COMPOSIT is the program which merges the maps while keeping track of the code combinations as they occur. Again a data processing report is generated and a table of the code combinations which were encountered is made available.

The output data set is of slightly different form and content than the basic and interpreted data sets. Figure 2.3 diagrams the content and organization of a composited data set. The type number is either 10, 20 or 30 depending upon whether the combined input maps were all of basic form, all of interpreted form or of mixed forms respectively. New information which was not required previously is the number of maps overlaid, the number of levels or codes in each input map and the altered form of code table at the end of the data set.

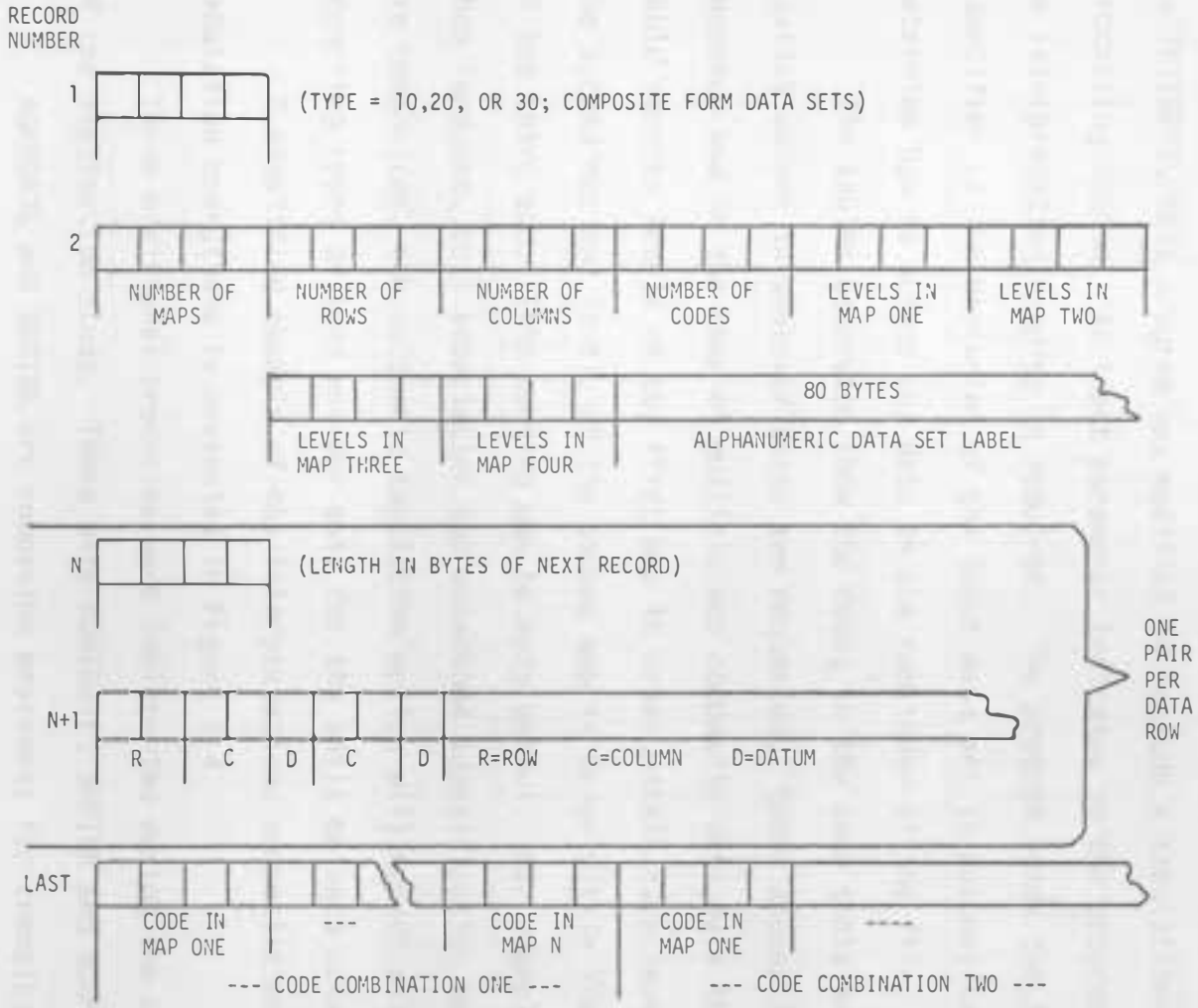



Figure 2.3. Organization of the records in a composited data set of AREAS.

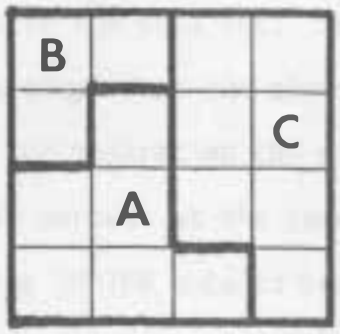
The third requested process was inventory tabulation in spatial units. Since most of the required logic was already available in INTERPRT, this program was modified to include a tabulation-only processing option. An input parameter indicates to the program that no interpretation mapping is required. The program reads the type identifier at the beginning of the input data set to automatically determine how to access the data in the remainder of the file.

The tables generated show the codes in the code table and the spatial extent in whatever units are requested. Codes appear in sorted sequence and in the case of multiple map composite data sets the table reports levels of the first map in order within each level of the second map and levels of the second map in order within levels of the third etc. This sorting can be very useful. For example when land use, soil association and watershed identification maps are composited, the automatic tabulation option will report all occurring types of land use per soil for the soils on each watershed.

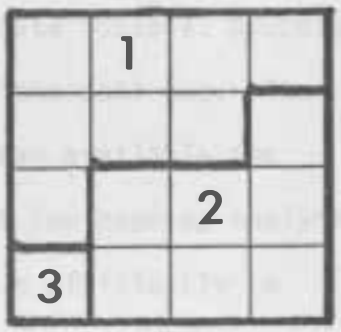
A simplified example of the interpretation, composite and tabulation operations is presented in Figure 2.4.

Three additional processes were implemented during the course of the application study. These were AGREGATE, SHRINK and BOUNDARY.

AGREGATE and SHRINK are companion processes for changing the effective grid cell size. Any integer factor may be applied and the grid cell size will effectively increase by that factor in both dimensions. AGREGATE moves the boundary locations around to effect

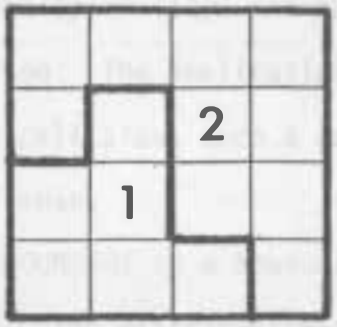



Soils A,B,C

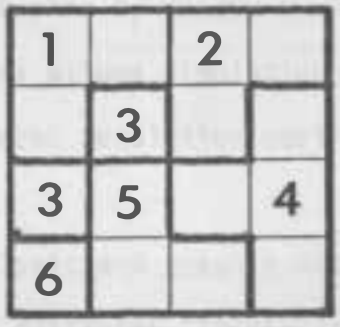



Watersheds 1,2,3

(a) 2 example map segments, 1 acre cells




(b) interpretation of soils into 2 categories of workability [A=1, B=2, C=2]




(c) composite of the 2 maps in (a) 1=B1, 2=C1, 3=A1, 4=C2, 5=A2, 6=A3

| CATEGORY | ACRES |
|---|---|
| A | 6 |
| B | 3 |
| C | 7 |
| TOTAL | 16 |

| CATEGORY | ACRES |
|---|---|
| A 1 | 2 |
| B 1 | 3 |
| C 1 | 3 |
| A 2 | 3 |
| C 2 | 4 |
| A 3 | 1 |
| TOTAL | 16 |

(d) tabulations of the data sets in prints (a) and (c)

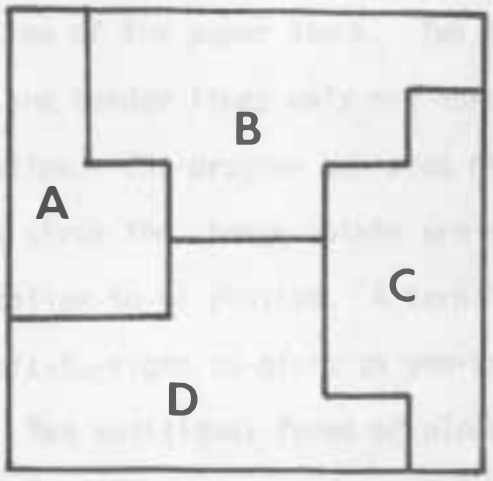
Figure 2.4. Example interpretation, composite and tabulation processes.

this cell size change without reducing the number of rows and columns in the data set. SHRINK is the separate optional process which accomplishes the physical reduction of the data set. The reason for separating the processes was to have available the AGREGATE outputs at the same physical extents for overlay analyses while the SHRINK outputs could be handled more efficiently in plotter processes. AGREGATE and SHRINK processes are diagrammed in Figure 2.5. Code assignment for the larger cells is made by area dominance (plurality) and multiway ties are broken by random number generation. The application of these programs allows simulation of various cell sizes once a data set of the finest resolution desired is available.

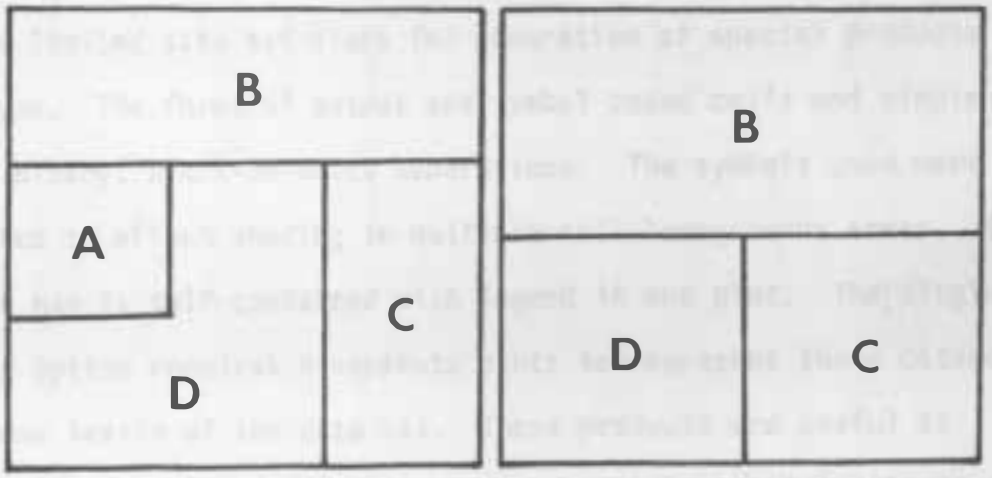
BOUNDARY is a boundary separation analysis and graphic display program. The distribution of inter-boundary distances (spans) in the horizontal dimension, vertical dimension and total data set are developed and displayed. Also distributions of distances from a theme or class to the adjacent different classes are similarly developed via horizontal, vertical and total approaches. The Kolmogorov-Smirnov test checks for horizontal-vertical dissimilarity and means of all distributions are reported. Details will be provided in the application discussion.

## 2.4 Mapping and Displays

Map products are generated on a drum plotter using Calcomp plotting subroutines and the interfacing logic of program PLOTTER. This program automatically segments and pages large plots so data set

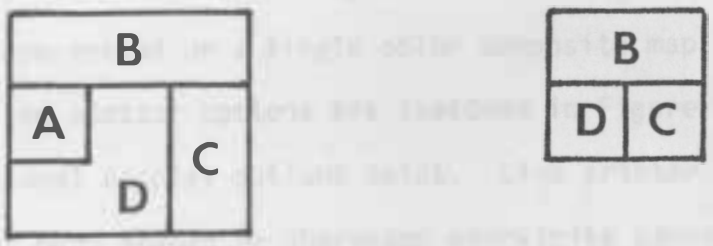



(a) 6 x 6 Data Set




Factor = 2 Factor = 3

(b) AGREGATE Process



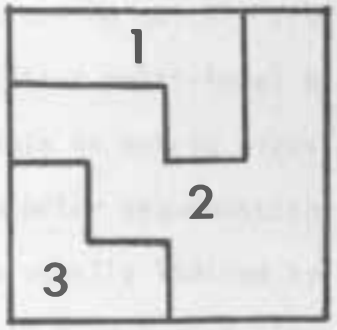

(c) SHRINK Process

Figure 2.5. Example AGREGATE and SHRINK Processes.

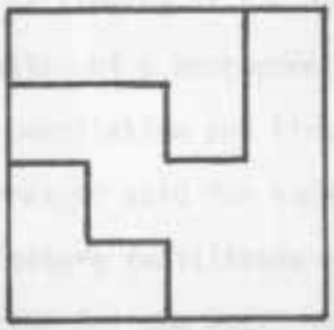
size is limited only by consideration of the scale and length dimension of the paper stock. Two forms of plot are producible. These are border lines only and border lines with code number annotation. The program operates directly on the defined data base format since the change points are exactly the required within-row information to be plotted. Alternate lines are plotted right-to-left and left-to-right to minimize pen-up travel time.

Two additional forms of plotter maps are available. These are more decorative and more expensive. Thus special programs which handle limited data set sizes for generation of special products are employed. The forms of output are symbol coded cells and single-code, binary, black-on-white separations. The symbols used were selected to effect shading in multiple cell homogeneous areas. The output map is self-contained with legend in one plot. The single code, binary option requires n separate plots to represent the n categories or theme levels of the data set. These products are useful as sequential overlays to a base map or as high-contrast inputs to a dry color process such as Chromalin by DuPont. Through this process color overlay transparencies or a single color composite map may be produced. Examples of the plotter options are sketched in Figure 2.6.

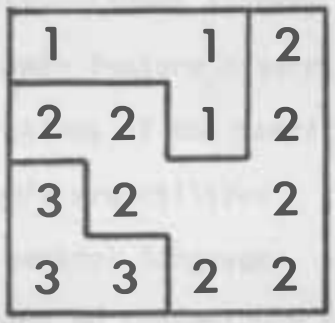
Additional display options exist. Line printer mapping of boundaries or maps shaded by character overstrike cannot compete with plotter products because of fixed scales and the non-unity, x-y aspect ratio. They can be useful for quick look and error check/verification purposes as they are much faster production items than pen plots.



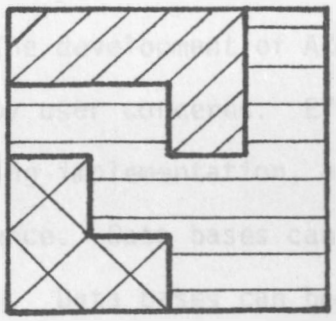

(a)



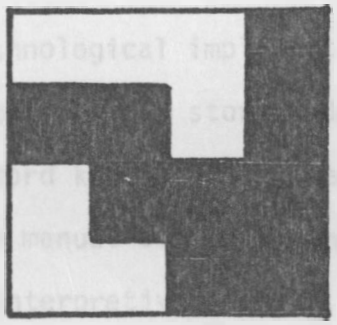
(b)




(c)



(d)



(e)

Figure 2.6. Plotter mapping options. (a) example data segment, (b) border lines, (c) border lines with code annotations, (d) symbol, and (e) binary, single code separation for code 2.

By converting the change point format data base to a one-byte-per-cell matrix, data sets can be quickly prepared to interface with film recording devices. Either multi-level black and white or full color products are achievable in matrix sizes dependent upon the hardware limitations or on prior segmentation of the data set.

Product options are usually limited by the devices available to the user, not by the data base organization.

Additional features of AREAS which promote economical operations are the cataloging of all programs into a disk-resident library, and the creation of a procedure library. The former feature assures that program compilation and linkage editing operations of the operating system are not paid for each time the program/s are utilized. The latter feature facilitates very minimal job control language preparation for the AREAS operator, thus promoting convenience and throughput.

## 2.5 Summary of Features

The development of AREAS stressed technological implementation guided by user concerns. Efficiency was emphasized in storage design, processing implementation, automation of record keeping, and operator convenience. Data bases can be created from manual coding or image matrices. Data bases can be processed via interpretive mapping, multiple map compositing, cell resolution adjustment or boundary characterization study. Outputs include inventory tabulations, variable scale, and/or form plotter maps, color or black and white

film recording images, line printer boundary maps or shade maps, and data set summaries and listings for record keeping. Developmental efforts were distributed among input, management, processing and display phases of the information system to assure a foundation of useful capability during the course of continued AREAS development.

# CHAPTER 3. SYSTEM PERFORMANCE EXPERIMENTS

Research objective one was to derive a performance prediction method based on map characteristics and useable prior to digitization of the map. The route outlined for achievement of this objective included investigation of performance of the geoinformation system as measured by (tabulation) inventory accuracy and by mapping (spatial) accuracy. The system is considered to include the data base together with the processing programs of AREAS. The performance is dependent on the grid-cell nature of the data and not on the specific programs which implement the analyses. This chapter reviews experiments on inventory and mapping accuracy as related to cell size variation.

## 3.1 The Data Base

A sample map segment of soil association data was selected for intensive study. An area of moderate boundary density and a mixture of region sizes and shapes was selected. The intent was to avoid bias that might enter artificially created data or that might arise from data with regularly shaped and/or sized patterns. The data map segment is shown in Figure 3.1.

The utilization of a very fine grid network was realized as particularly important to the study of cell size influence. A first-cut rule of cells smaller than the smallest separation of adjacent boundary lines was considered. In addition a constraint was imposed that the cell size be an even integer divisor of the

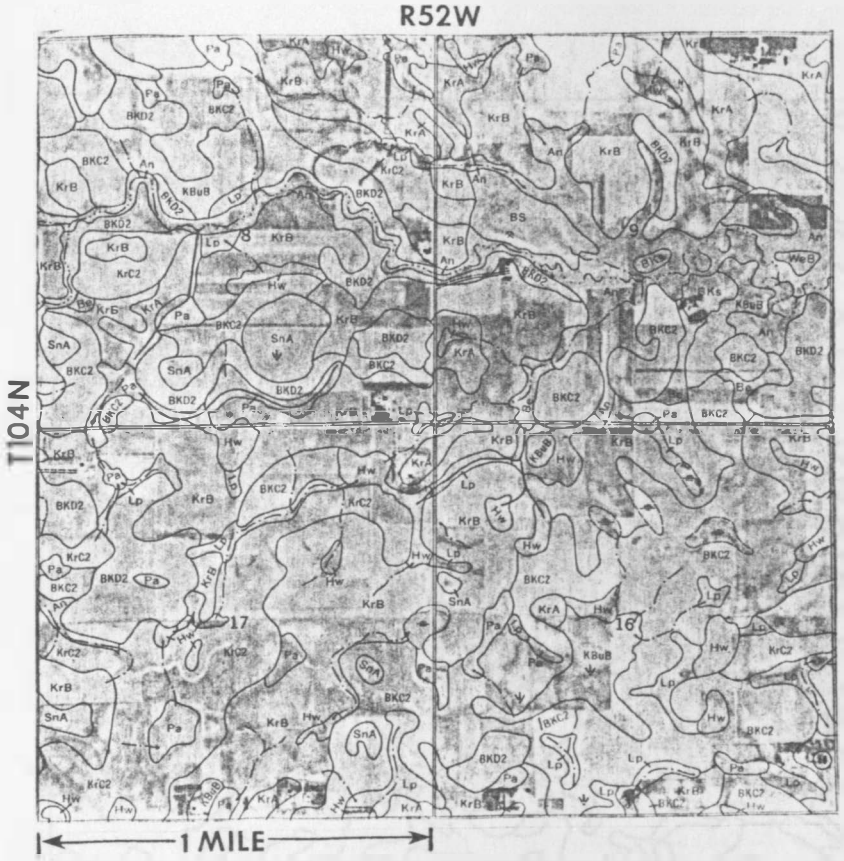



Figure 3.1. The original map of the intensive-study data set. The four sections of soil association data are located in Minnehaha County, South Dakota.

commonly used 1.008, 4.032, and 16.128 ha (2.5, 10, 40 acre) cells. Under this constraint, the study of resolutions would pass through cell sizes which users have employed and for which users have developed a conceptual feeling.

The final selection of a 0.007 ha (0.0174 acre) cell resulted in a grid of 384 elements square. The data set created was considered the "true" map as reference for all analyses. This map is shown in Figure 3.2. The data set was generated by enlarging the map,

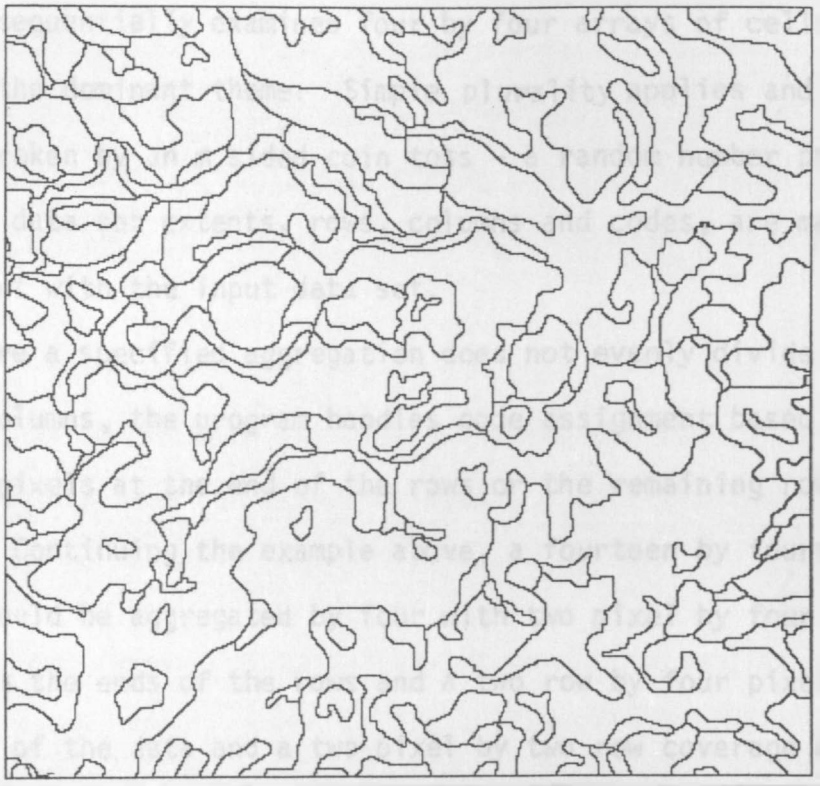


Figure 3.2. The computer map of the "true" data set at 0.007 ha (0.0174 acre) cellular grid.

preparing a computer drawn grid and manually encoding the cell contents.

For the analysis envisioned additional data sets of the same map at larger and larger cell sizes would be required. Rather than attempt many manual digitizations, the cheaper, faster and more dependable route of computer aggregation was chosen.

Program AGREGATE was created to accomplish aggregation to larger and larger cell sizes. An integer number is specified, for example four, and the change points of the data set are altered to

align on column and row boundaries divisible by four. The algorithm sequentially examines four by four arrays of cells to determine the dominant theme. Simple plurality applies and m-way ties are broken by an m sided coin toss - a random number process. The output data set extents, rows, columns and codes, are maintained in agreement with the input data set.

Where a specified aggregation does not evenly divide the input rows and columns, the program handles code assignment based on the remaining pixels at the end of the rows or the remaining rows of the data set. Continuing the example above, a fourteen by fourteen input data set would be aggregated by four with two pixel by four row coverage at the ends of the rows and a two row by four pixel coverage at the end of the data and a two pixel by two row coverage at the ends of the last two rows of data. Thus, any integer aggregation can be processed up to the program limit of sixty four.

For uneven aggregation division of the row-column dimensions of the input data set, the pixels processed at the right and bottom edge would introduce artificial structure into any analysis of boundary spacing. Therefore aggregation was only applied for even divisors of the 384 by 384 base data set. The aggregations were 2,3,4,6,8,12, 16,24,32,48, and 64. Corresponding cell sizes were 0.028, 0.063, 0.112, 0.252, 0.448, 1.008, 1.792, 4.032, 7.168, 16.128, and 28.672 ha (0.069, 0.156, 0.278, 0.625, 1.111, 2.500, 4.444, 10.000, 17.778, 40.000, and 71.111 acres).

The eleven aggregations were also processed with SHRINK which generated data sets with reduced row-column dimensions. Rows are omitted and column designations divided by the aggregation factor. These alternate forms of aggregations were useful for more rapid and economical plotting as well as some forms of boundary structure analysis.

The original data set, the eleven aggregated data sets and the eleven aggregated-shrunken data sets comprise the twenty three data sets of the data base for intensive study.

## 3.2 Performance Experiments

Performance was analyzed from the product viewpoint as mapping accuracy and inventory accuracy. The influencing variable under study was cell size. The source data set with the finest resolution cell was the "true" map reference for mapping accuracy via Switzer's approach. An inventory tabulation of the spatial quantities of each map unit in the original data set became the "true" inventory reference for the study of inventory accuracy. The processing applied to each of the eleven data sets with larger cell size is diagrammed in Figure 3.3. The inventory and mapping performance was observed by map units and summarized for the entire map. This allowed observation of size-shape influences as well as evaluation of the entire data set for each of the eleven different cell sizes.

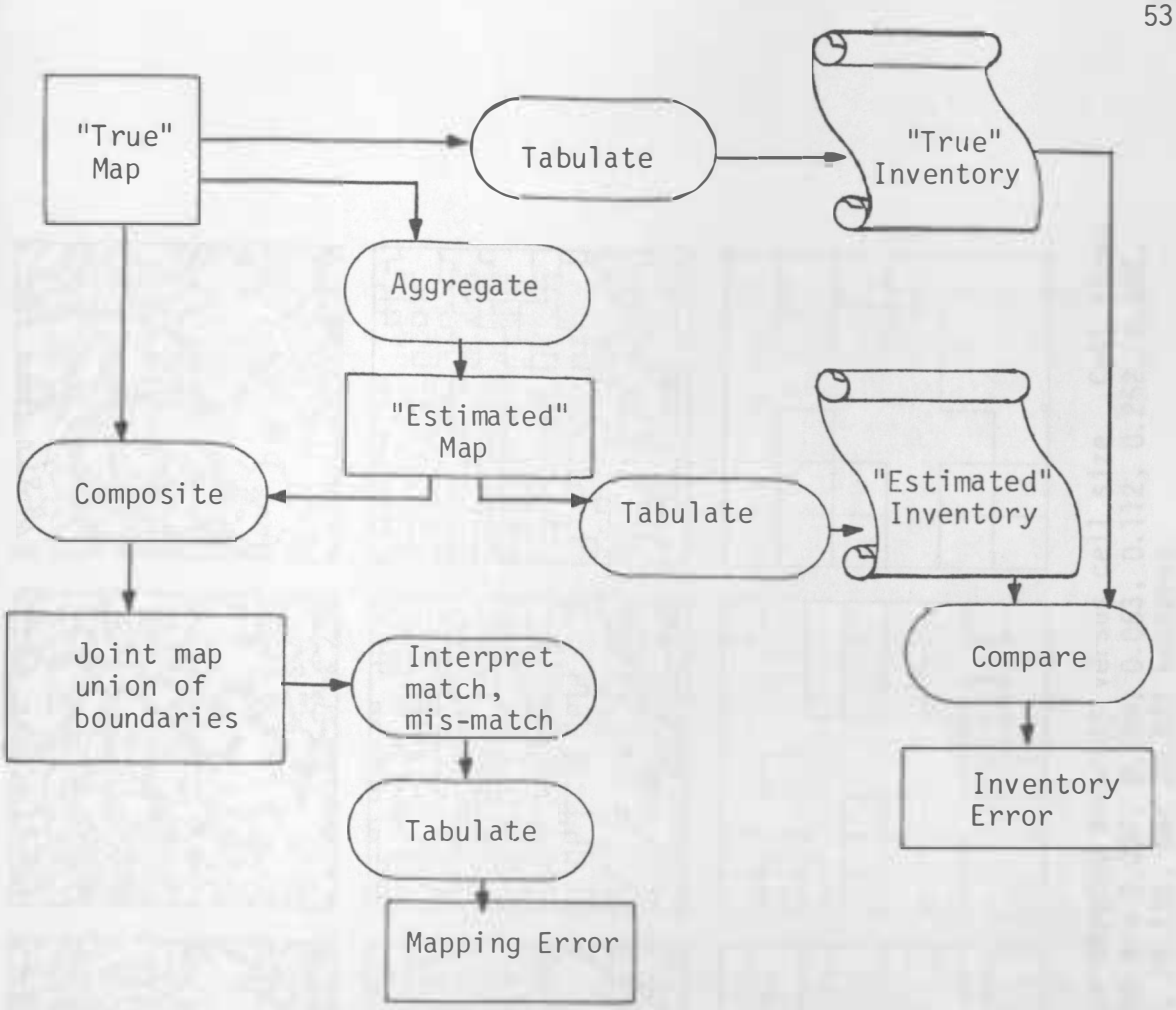



Figure 3.3 Performance evaluation processing diagram for an increased cell size.

## 3.3 Experimental Results

The appearance of the maps corresponding to the twelve resolutions in the data base can be compared in Figure 3.4. Corresponding mapping errors, the mismatch of areas when comparing maps of larger cell sizes to the smallest cell size available, are displayed spatially in Figure 3.5. Note that the basic data set (upper left in Figure 3.4) and the aggregation by 2 to the immediate right are of such a fine cell size for the output scale being used

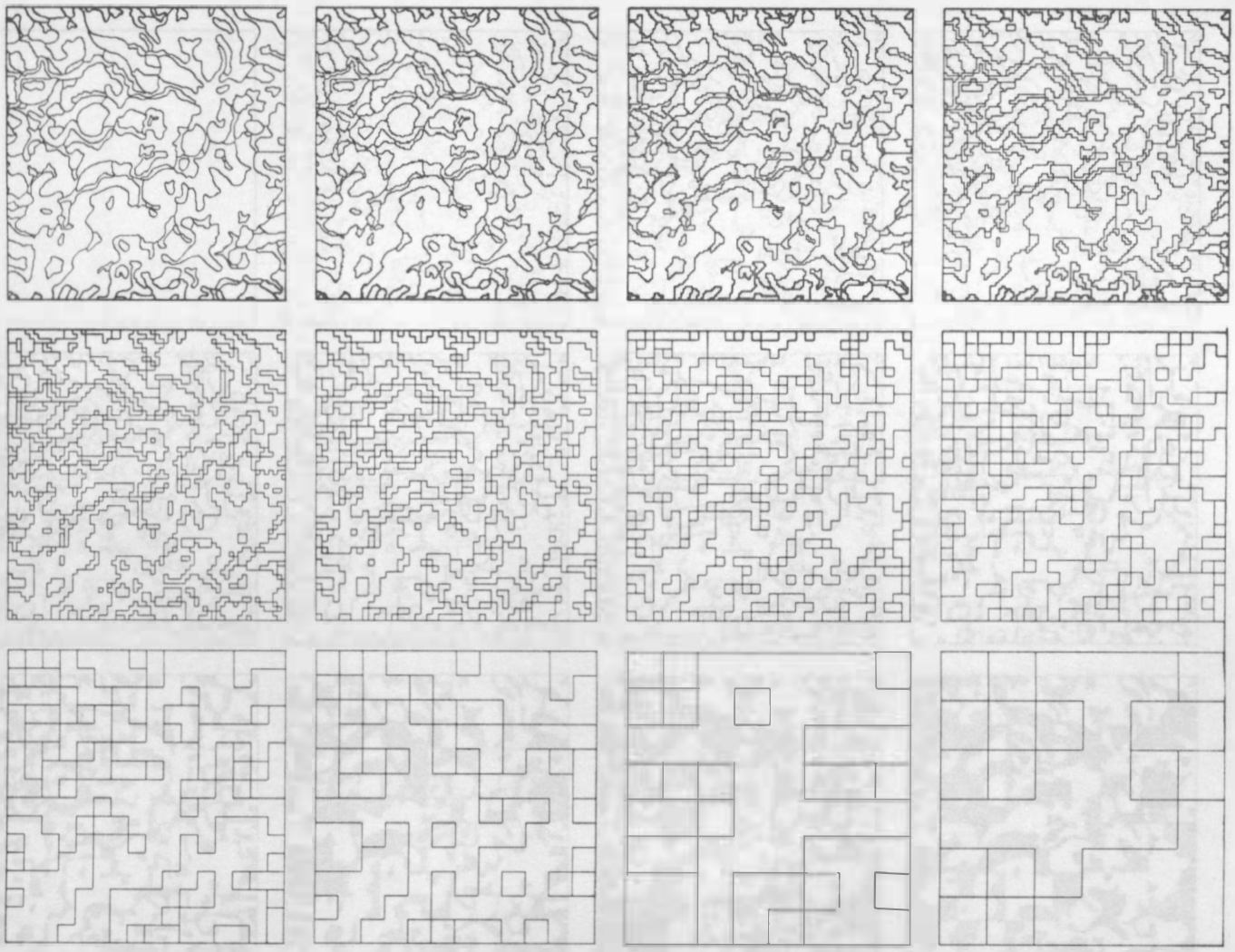


Figure 3.4. The twelve maps compared for mapping accuracy versus cell size. Cell sizes from top left to lower right are 0.007, 0.028, 0.063, 0.112, 0.252, 0.448, 1.008, 1.792, 4.032, 7.168, 16.128, and 28.672 hectares.

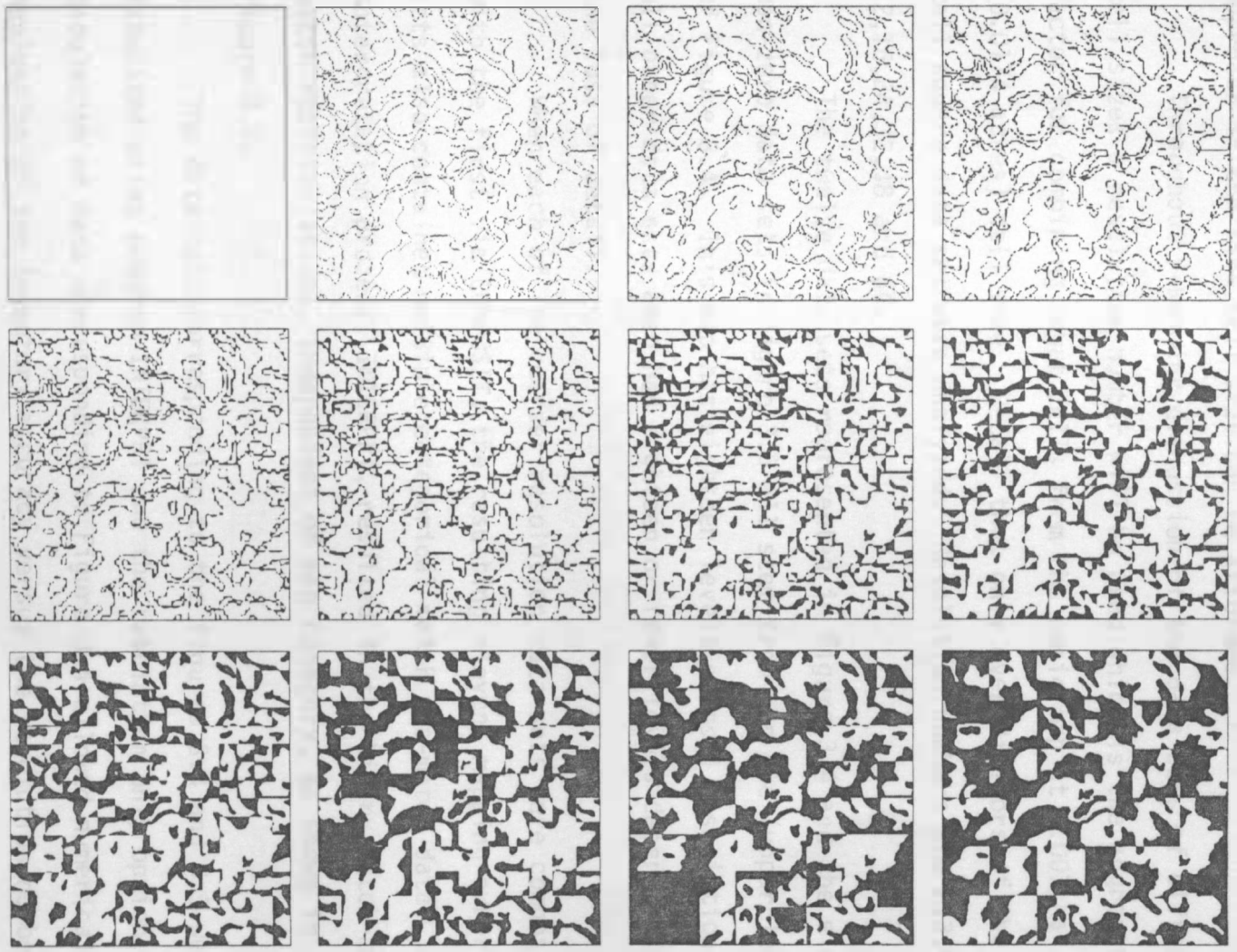


Figure 3.5. Representations of mapping error. Maps correspond to Figure 3.4.

that the cellular nature is not even apparent. Cellular representations can be as cosmetically pleasing as the polygon approach if the fine cell size can be afforded.

Throughout remaining discussion of the data sets of different cell sizes a resolution number will be used which is the aggregation factor for combining input cells from the basic map at 0.007 ha (0.0174 acres) cell size. Recall that only even divisors of the 384 cell square data set were analyzed. Resolution numbers are 2,3,4,6,8, 12,16,24,32,48 and 64.

The twelve data sets represented in Figure 3.4 and the eleven, non-zero data sets in Figure 3.5 all stem from map one (upper left) of Figure 3.4. This map has eighteen levels of soil association and all other data sets also had eighteen or fewer levels depending on the loss of detail.

When each of the coarser resolution data sets were composited with the first via COMPOSIT, the resulting eleven, mapping-error data sets also contained multiple categorical match mis-match data. An interpretation process, INTERPRT, resulted in the binary match-mis-match representations, independent of map category, as shown in Figure 3.5.

The data sets corresponding to the figures 3.4 and 3.5 were tabulated using program INTERPRT and the tabulation only option. The tabulation of data corresponding to Figure 3.4 allowed numerical evaluation of the inventory accuracy versus the resolution number (cell size). The tabulation of data corresponding to Figure 3.5

allowed numerical evaluation of the mapping accuracy versus the resolution number (see Figure 3.3). The results obtained are plotted in Figure 3.6 versus the resolution number. The percentage

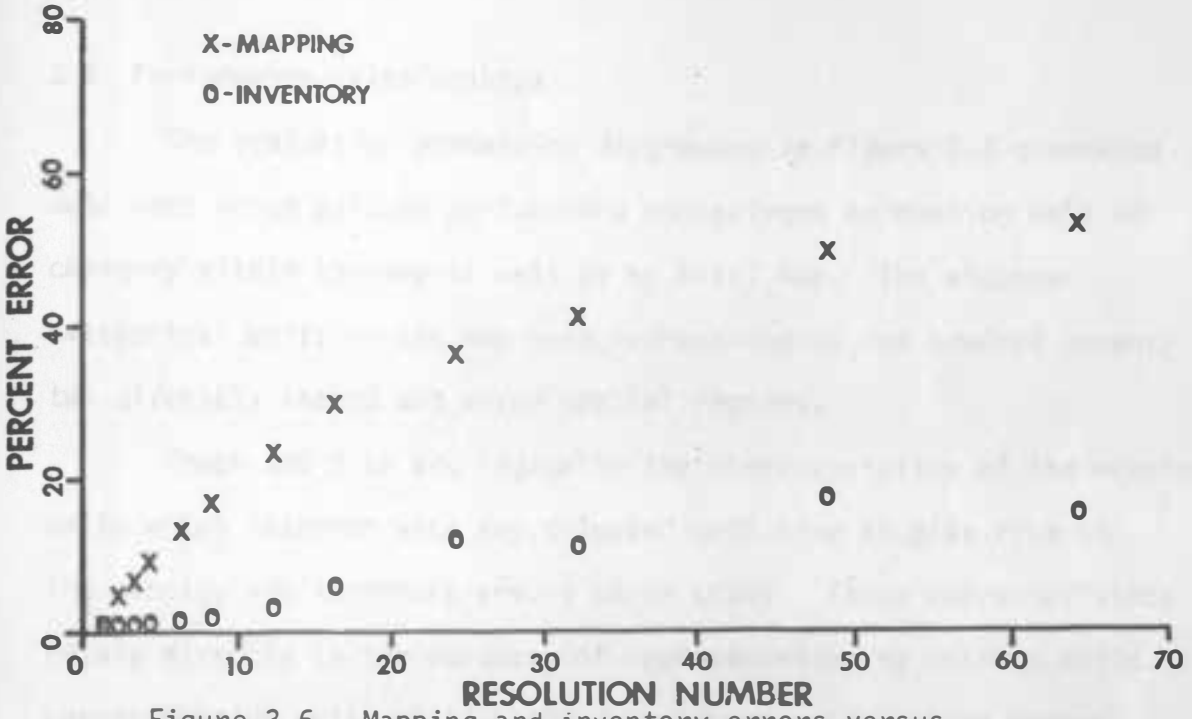



Figure 3.6. Mapping and inventory errors versus resolution number.

mapping error is the number of incorrectly mapped cells divided by the total 147,456 cells in the 384 cell square. Tabulation error will be lower than mapping error since omission-commission errors per category can offset each other. Since all cells are tabulated as belonging to some category, the average error across categories, with some categories over tabulated (+ error) and some under tabulated (- error), will always be zero. RMS, root-mean-square, error was considered to overcome cancellation but the value obtained would be an average value for a map category - not the entire map. A root sum square, RSS, was employed where the squares of category tabulation

errors are added and the square root taken. The result is an accuracy figure for the entire map. These are the percentage inventory errors graphed in Figure 3.6.

## 3.4 Performance relationships

The evaluation processing diagrammed in Figure 3.3 generated data sets which allowed performance comparisons by mapping unit or category within the map as well as by total map. The eighteen categorical units in the map were represented by one hundred seventy two diversely shaped and sized spatial regions.

Shape and size are logically the characteristics of the mapping units which interact with any selected cell size to give rise to the mapping and inventory errors under study. These characteristics relate directly to the adequacy of representation by uniform cells. Larger mapping units will accommodate larger cells. Also larger mapping units will tend to have fewer perimeter cells in relation to total cells. Shapes may be broadly classified into a complexity spectrum from simple to complex. The criterion is border versus area or perimeter cells versus total cells. Simple shapes will require less perimeter to enclose an area and, therefore, tend to be less subject to the errors which arise in cellular representation of the borders. Figures 3.7 and 3.8 show the behavior of mapping and inventory errors versus resolution number for several selected mapping units of differing size and shape. These mapping units are separated and spatially displayed in Figure 3.9. Table 3.1 summarizes

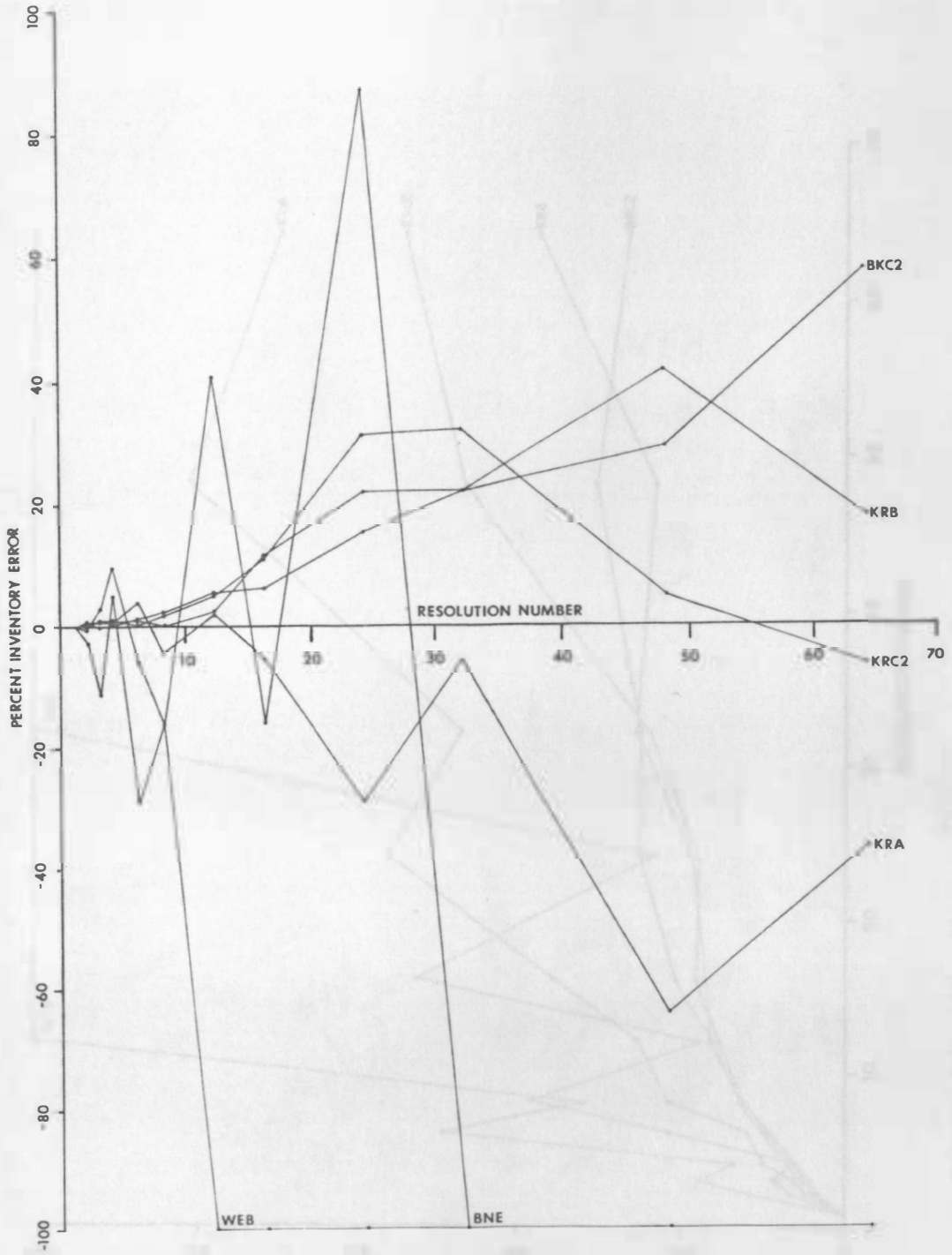



Figure 3.7. Inventory errors versus resolution number for selected mapping units.

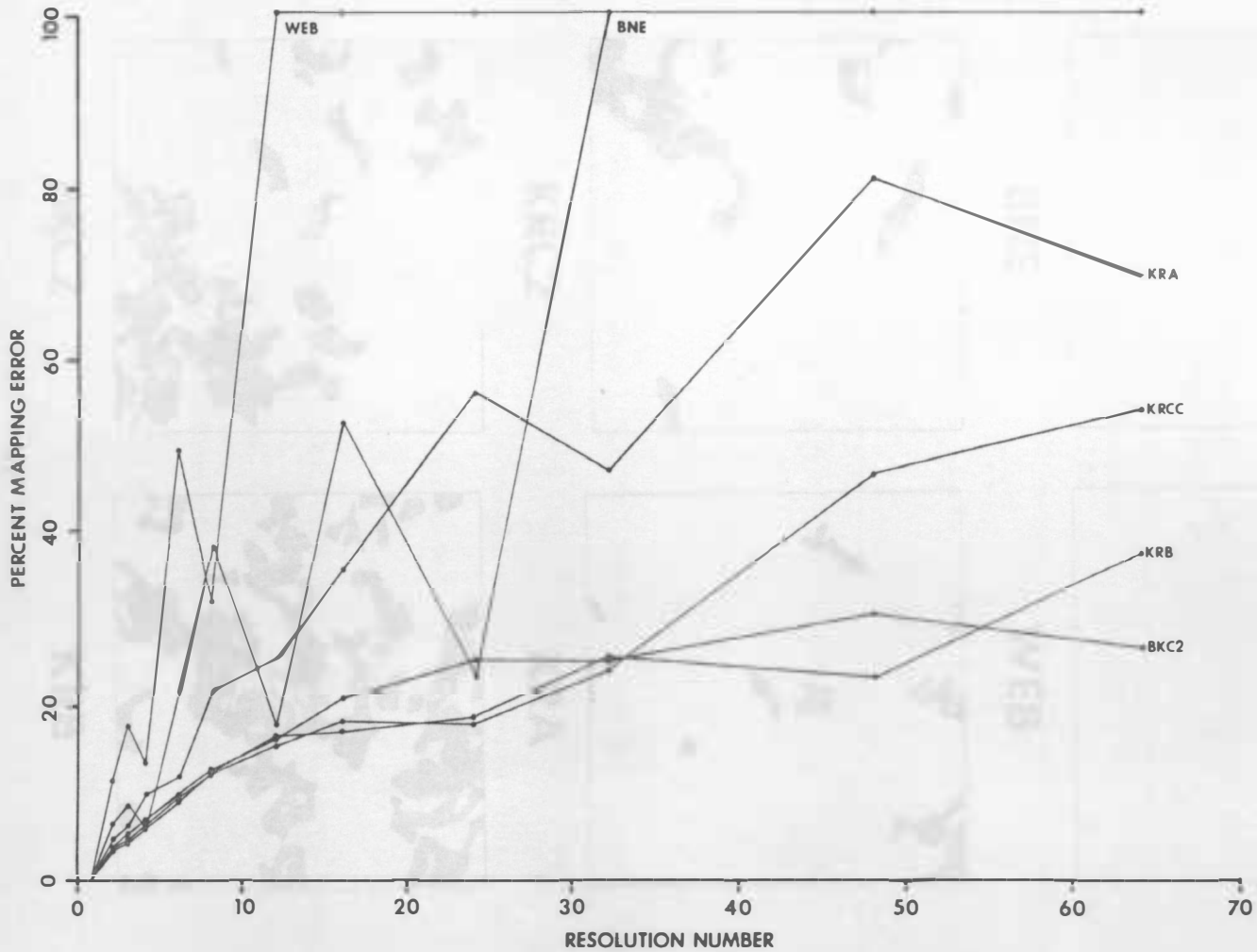



Figure 3.8. Mapping errors versus resolution number for selected mapping units.

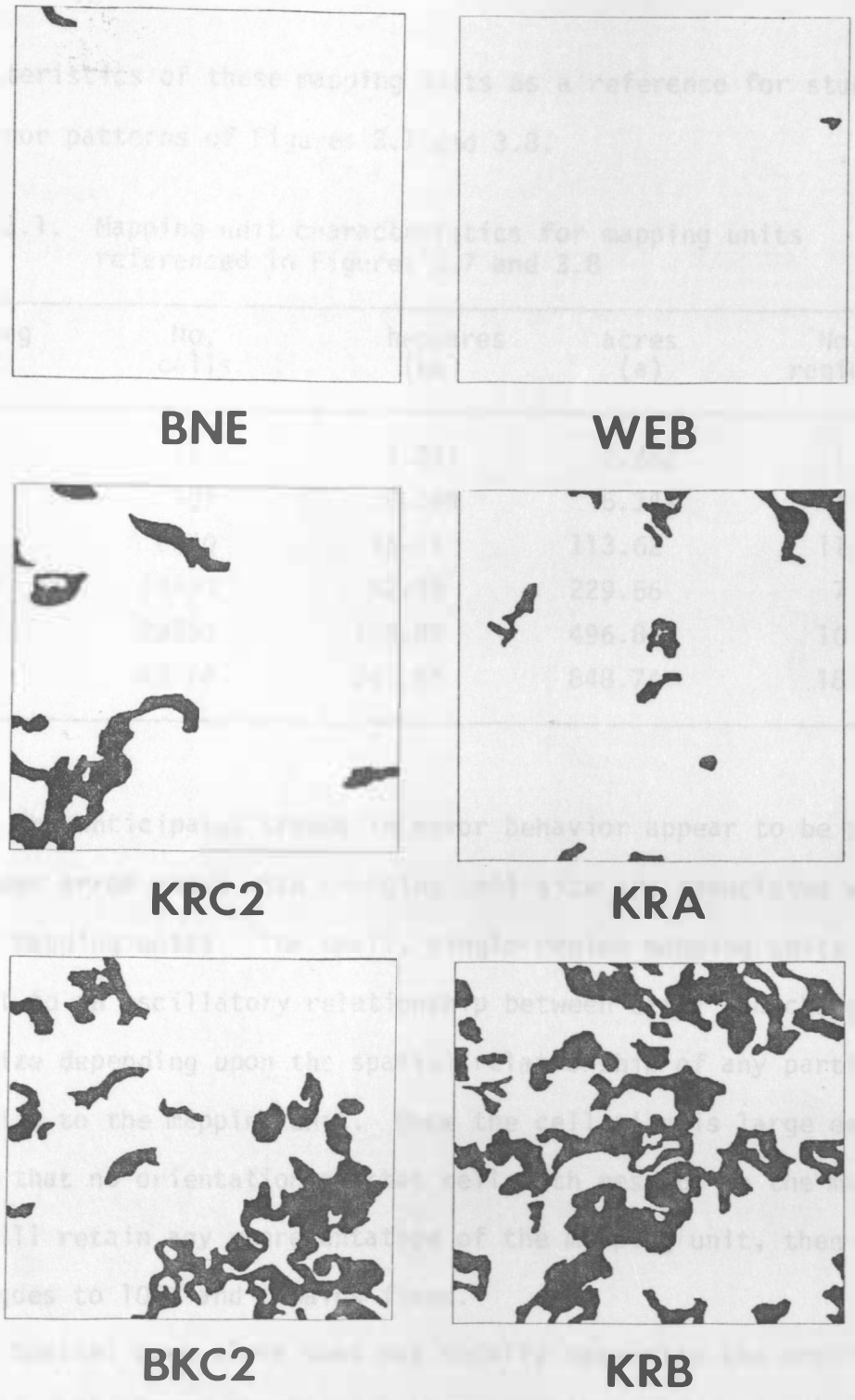



Figure 3.9. Spatial representations of the mapping units referred to in Figure 3.7.

characteristics of these mapping units as a reference for studying the error patterns of Figures 3.7 and 3.8.

Table 3.1. Mapping unit characteristics for mapping units referenced in Figures 3.7 and 3.8

| Mapping Unit | No. cells | hectares (ha) | acres (a) | No. regions |
|---|---|---|---|---|
| WEB | 153 | 1.071 | 2.662 | 1 |
| BNE | 307 | 2.149 | 5.342 | 1 |
| KRA | 6530 | 45.71 | 113.62 | 11 |
| KRC2 | 13199 | 92.39 | 229.66 | 7 |
| BKC2 | 28555 | 199.89 | 496.86 | 10 |
| KRB | 48778 | 341.45 | 848.74 | 18 |

The anticipated trends in error behavior appear to be present. The lower error rates with changing cell size are associated with the larger mapping units. The small, single-region mapping units are subject to an oscillatory relationship between error and changing cell size depending upon the spatial relationship of any particular cell size to the mapping unit. Once the cell size is large enough to assure that no orientation of that cell with respect to the mapping unit will retain any representation of the mapping unit, then the error goes to 100% and remains fixed.

Spatial area alone does not totally determine the error versus changing cell size. The largest three mapping units are not strictly in the same performance sequence, particularly when comparing mapping and inventory sequences. The additional influence of perimeter is

evident as partially determined by the number of regions of the mapping unit. Shapes of the mapping units also contribute to the extensiveness of perimeter.

Even though several factors appear to influence the errors on a mapping-unit basis, the relationships of Figure 3.6, effectively representing averages over the eighteen map units, are remarkably smooth trends which could be modelled by least squares curve fitting.

# CHAPTER 4. DATA CHARACTERIZATION ANALYSIS

Chapter three reviewed the behavior of mapping and inventory accuracies as cell size changed. This chapter will look at a map characteristic as it relates to changing cell size and in turn to the product accuracies.

## 4.1 The Distribution of Spans

The change-point version of the compact, sequential coding scheme is utilized by AREAS. Geographic reference is implicit to the cellularization and compaction is achieved by the representation of spans of identical cells. The distance equivalent of n uniform cells is a discrete representation of the continuous measure of inter-boundary distance in the input map.

From an intuitive standpoint, the distances which separate map boundaries are precisely the "character" of the input map which determines the influence of cell size on accuracy of mapping and accuracy of tabulation. If all inter-boundary distances were much larger than the selected cell size, many cells would occur between boundaries and remain accurately encoded even though the boundary cells are assigned by dominance considerations. In other words, the area in error within any given boundary cell would be a small portion of the area represented by the remaining interior cells of the same map unit. As cell size grows larger relative to a given map fewer cells occur within a region and the border cells of the region include larger area errors. This intuitive concept is diagrammed in Fig. 4.1.

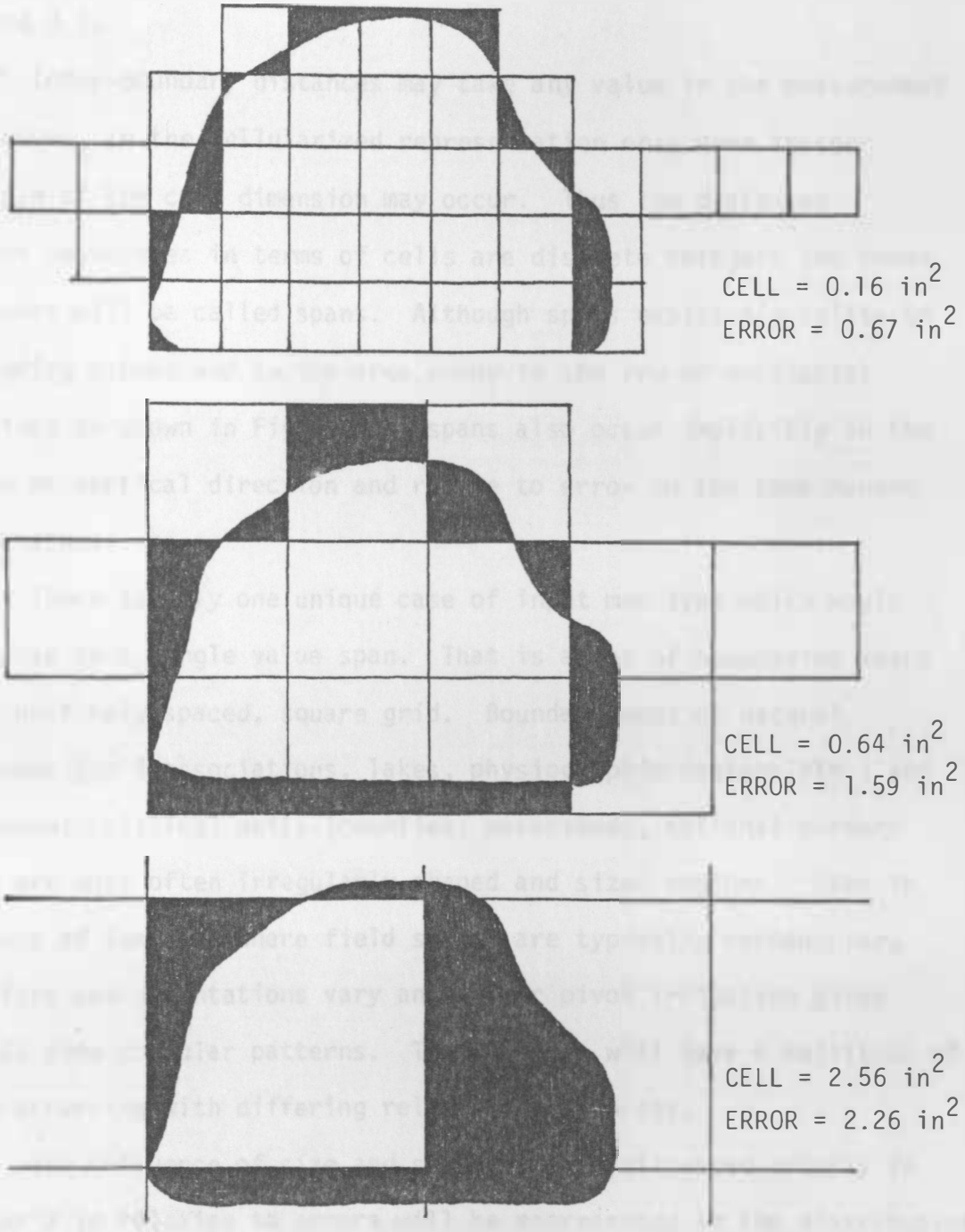



Figure 4.1 Growth of mapping error as cell size increases for fixed interboundary map distances. Area in error by cell-dominant coding process is shaded.

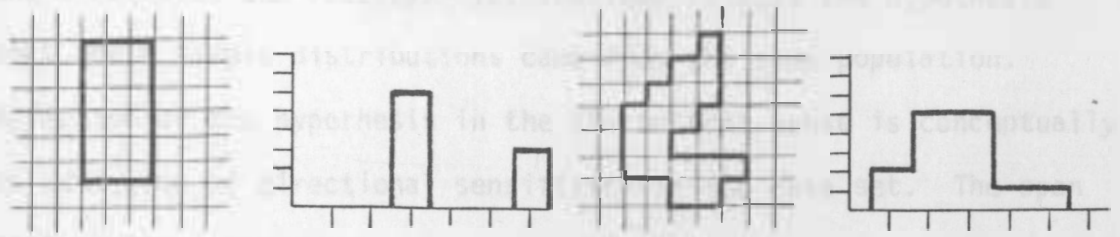
The actual occurrence of this effect is quite evident in Figures 3.4 and 3.5.

Inter-boundary distances may take any value in the measurement continuum. In the cellularized representation only some integer multiple of the cell dimension may occur. Thus the distances between boundaries in terms of cells are discrete integers and these distances will be called spans. Although spans explicitly relate to the coding scheme and to the area error in the row or horizontal direction as shown in Figure 4.1, spans also occur implicitly in the column or vertical direction and relate to error in the same manner as horizontal spans.

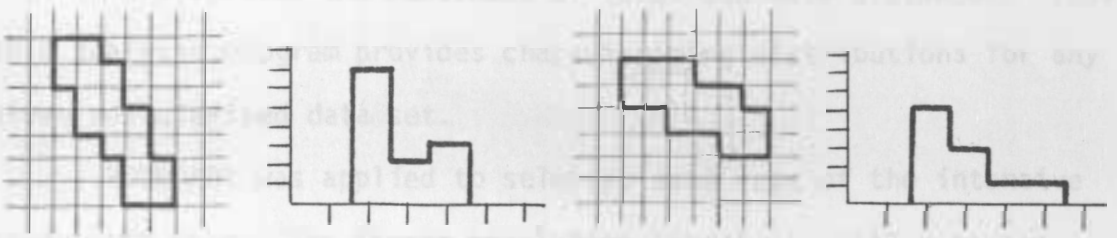
There is only one unique case of input map type which would give rise to a single value span. That is a map of boundaries which are a uniformly spaced, square grid. Boundary maps of natural phenomena (soil associations, lakes, physiographic regions etc.) and management/political units (counties, watersheds, national borders etc.) are most often irregularly shaped and sized regions. Even in the case of land use where field shapes are typically rectangular, the sizes and orientations vary and center pivot irrigation gives rise to some circular patterns. Typical maps will have a multitude of spans occurring with differing relative frequencies.

The influence of size and shape factors discussed briefly in Chapter 3 in relation to errors will be represented in the distribution of spans. The effect of various orientations of an object would also be present in a distribution of spans. Figure 4.2 shows the effects of

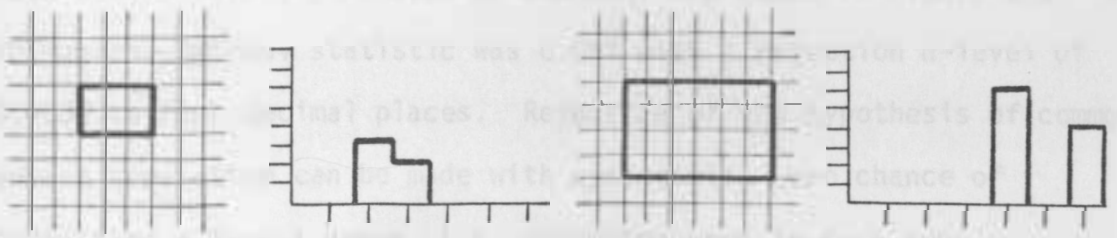
shape, orientation and size of a region on the span distribution. Note that the span distribution is shown as discrete in terms of the number of unit cells in the span.



(a) regular vs irregular shapes



(b) orientation factor



(c) size factor

Figure 4.2. The span distribution reflects region size, orientation and regularity of shape.

To experimentally evaluate the data sets created for this study, program BOUNDARY was prepared. This algorithm calculates tables and distribution means, and draws graphs of the relative frequency distributions for spans in the horizontal (row), vertical (column) and total (map) context. The Kolmogorov-Smirnov test is applied to the horizontal and vertical distributions to test the hypothesis that these sample distributions came from the same population. Rejection of the hypothesis in the statistical sense is conceptually an admission of directional sensitivity in the data set. The span distributions along and across rows of the data set are aligned with the cellularization in exactly the manner that the cellularization intuitively impacts the continuum of inter-boundary distances. Thus this analysis program provides characterizing distributions for any given cellularized data set.

BOUNDARY was applied to selected data sets of the intensive study data base. The finest resolution (smallest cell) data set, referred to as resolution one, was the "true" reference map. The span distributions generated by BOUNDARY are shown in Figure 4.3. The Kolmogorov-Smirnov statistic was 0.052 with a rejection $\alpha$-level of 0.0000 to four decimal places. Rejection of the hypothesis of common parent population can be made with essentially zero chance of committing a Type I error, i.e. rejecting when in fact true. Although the distributions appear very similar, the statistical conclusion is that they are not.

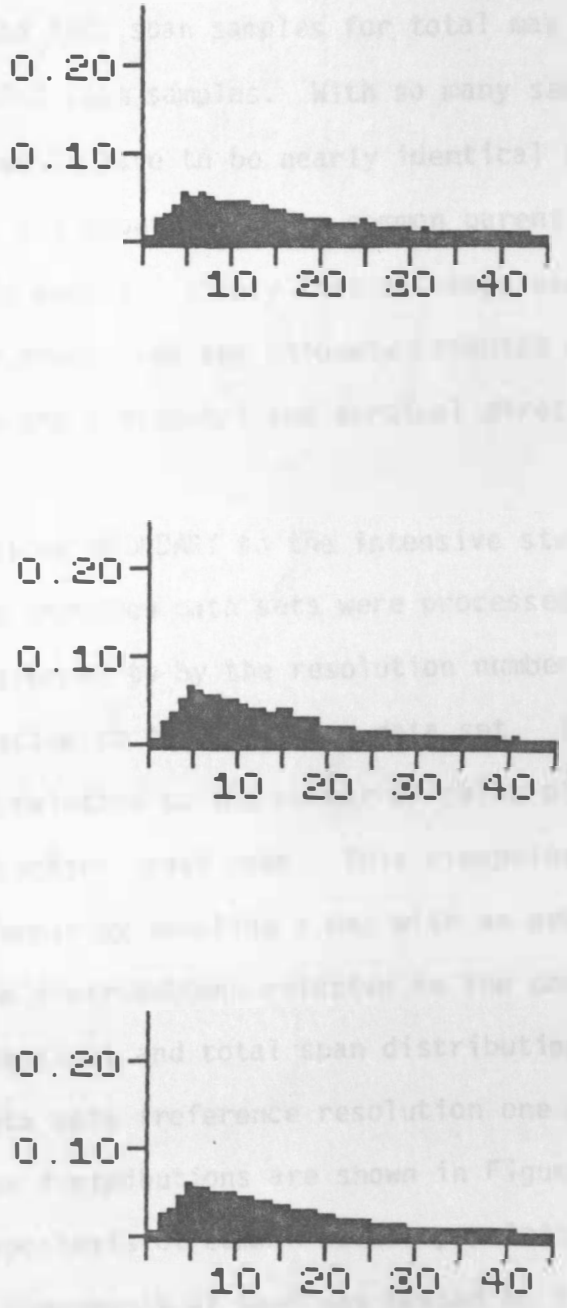



Figure 4.3. Span distributions for the reference data set of resolution number one. (a) horizontal scan, (b) vertical scan and (c) total map.

This seemingly paradoxical situation is resolvable by considering the degrees of freedom for the test. The horizontal distribution had 7981 span samples for total map coverage and the vertical had 8748 span samples. With so many samples available the distributions would have to be nearly identical to statistically fail to reject the hypothesis of a common parent population.

It would appear likely that all maps except a rare class of very-highly structured and uniquely oriented maps would be spanned differently in the horizontal and vertical directions under small-cell analysis.

In applying BOUNDARY to the intensive study data base, the aggregated and shrunken data sets were processed. Each of these data sets can be referred to by the resolution number which identifies the cell size relative to the reference data set. The span distributions generated are relative to the number of cells of that particular size which constitute each span. This viewpoint is consistent with results obtainable by sampling a map with an arbitrary cell size and obtaining span distributions relative to the chosen cell size. Horizontal, vertical and total span distributions were obtained for the twelve data sets (reference resolution one and eleven aggregations). The total span distributions are shown in Figure 4.4.

The hypothesis of common parent population for the horizontal and vertical components of span was tested by the Kolmogorov-Smirnov statistic for the twelve resolutions with the results as tabulated in Table 4.1. Note particularly that statistical dissimilarity disappears at the resolution of six which is the mode of the reference

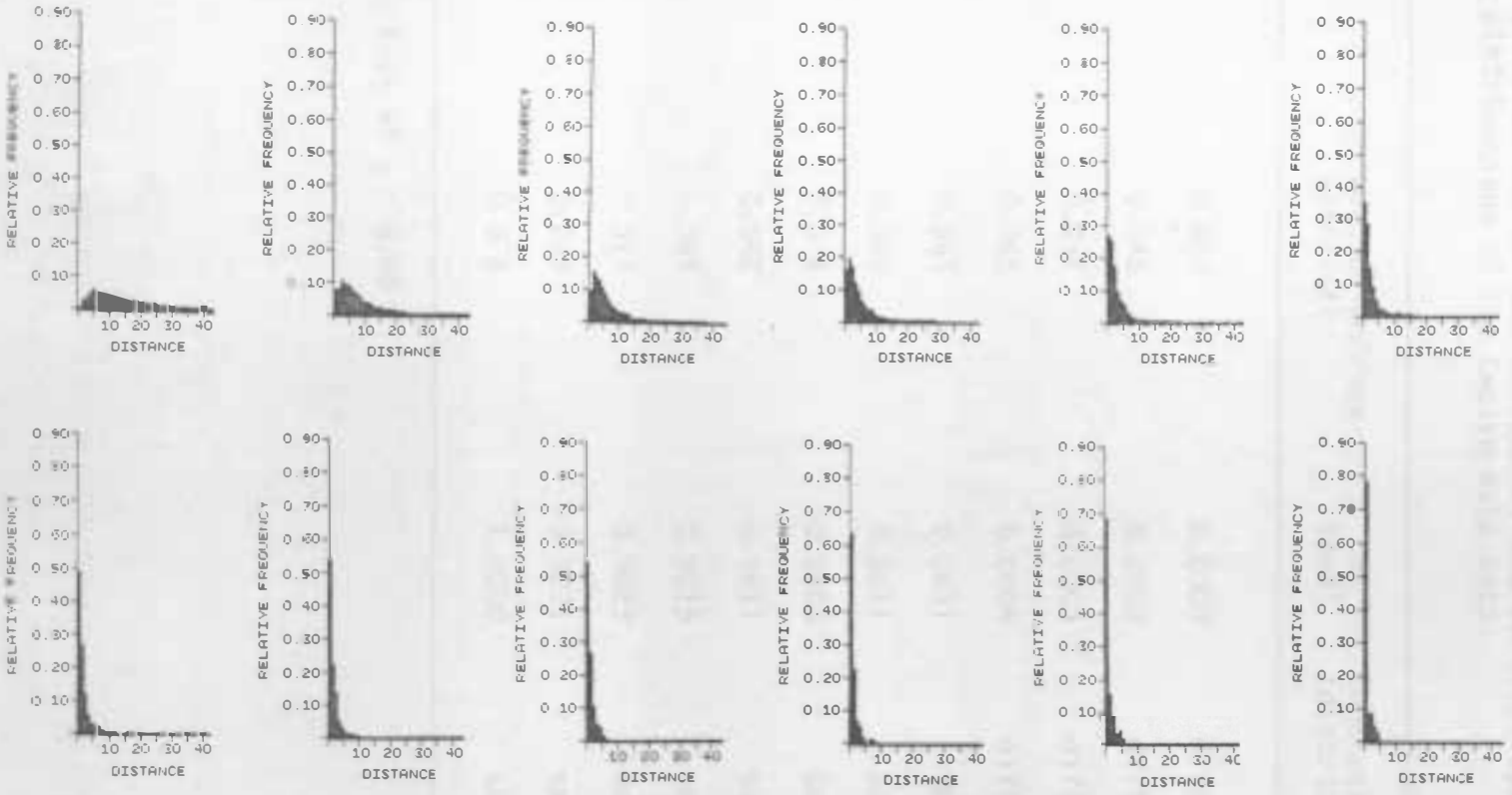



Figure 4.4. Span distributions for combined horizontal and vertical scans of the twelve different cell-sized data sets. Resolution numbers are (a) 1, (b) 2, (c) 3, (d) 4, (e) 6, (f) 8, (g) 12, (h) 16, (i) 24, (j) 32, (k) 48, and (l) 64.

Table 4.1. The Kolmogorov-Smirnov test of common parent span distribution for the horizontal and vertical span distributions of the twelve data sets.

| Resolution Number | Kolmogorov-Smirnov statistic | $\alpha$ level | Conclusion on span distributions * |
|---|---|---|---|
| 1 | 0.052 | 0.0000 | different |
| 2 | 0.046 | 0.0003 | different |
| 3 | 0.057 | 0.0003 | different |
| 4 | 0.065 | 0.0004 | different |
| 6 | 0.053 | 0.0631 | same |
| 8 | 0.042 | 0.9611 | same |
| 12 | 0.044 | 0.9408 | same |
| 16 | 0.042 | 0.9611 | same |
| 24 | 0.066 | 0.9215 | same |
| 32 | 0.071 | 0.9823 | same |
| 48 | 0.219 | 0.3221 | same |
| 64 | 0.076 | 1.0000 | same |

* Hypothesis test at $\alpha = 0.05$

data set. Apparently cellularization itself is beginning to cause the span structure in the horizontal and vertical directions to become statistically similar. This is suggestive of a test for too coarse a sampling cell size. The sampling of a map and generation of statistically similar horizontal and vertical span distributions is suggestive of a too-large sampling cell. The natural data structure should not be dominated by the cellularization. This is not to say that errors arising from cell sizes larger than the natural mode of the data are unacceptable. This remains an application judgement considering results such as in Figure 3.6 where tabulation errors never exceeded 20% even to the extreme aggregation to resolution number 64.

The key to success in the characterization analyses is whether the distribution selected (particularly a parameter of the distribution) relates to changing cell size. The trend in the distributions of Figure 4.4 is evident and to a degree is predictable. If the cells of all resolution one spans could be grouped by twos without relocating any boundaries in the aggregation process, then the span distribution for resolution number two would be that for resolution one with the span axis labels halved. Achievement of this in practice would require that all resolution one spans be divisible by two and that horizontal and vertical spans not be interrelated. This will not likely ever be the case. Hence the relationship of span distributions under changing cell size is not absolutely predictable.

To monitor and analyze the effect of changing cell size on the span distribution a parameter of the distribution had to be selected. The mode of the "true" reference data set was interestingly related to detectable difference in horizontal and vertical structure under changing cell size. The modes of the aggregated data sets, however, converge to one for the aggregation which surpasses the mode of the reference data set. Larger resolution number data sets have modes of one and discriminatory power via a mode parameter is lost. The mean, however, varies with resolution number throughout and was found useful for the desired purpose.

## 4.2 Spans versus cell size

Each span distribution has a calculable mean at the particular resolution involved. The sample mean span was calculated from the estimated total span density function (as represented by the relative frequency distribution) by the conventional discrete formula

$$\bar{x} = (1/n)\Sigma x$$

recognizing that without knowledge of the specific form of the density function f(x) there could be no claim regarding unbiased maximum likelihood estimation. The calculated mean spans for each resolution number are plotted in Figure 4.5.

The relationship is logical. At resolution number one the span mean is the actual mean of the spatial structure. Map resolution one is a lower limit on the x-axis. As resolution number increases

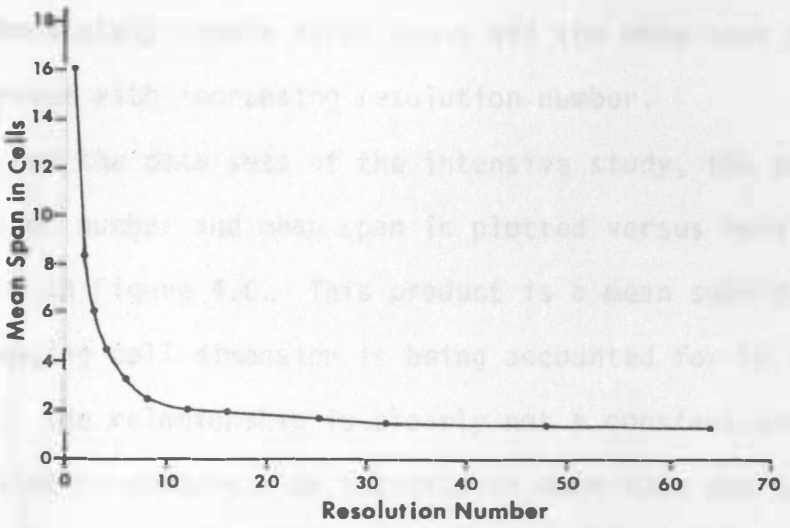



Figure 4.5. Span means versus resolution numbers.

even to the full dimension of the data set the number of cells would decrease to unity and the mean span would become one. A span mean of one is a lower limit on the y-axis. The hyperbolic nature follows the earlier argument that in the absence of boundary relocations during aggregation the distributions would change by simply dividing the resolution axis values by the resolution number. This would be equivalent to the product of resolution number and the mean span at that resolution equalling a constant (the span mean for resolution one).

Boundary relocations do occur during aggregations, however. As the resolution number increases and eventually surpasses the mean span of the reference data set, many smaller map units and even the thinner protrusions of the larger map units begin to disappear as they fail to dominate the larger cells. Elimination of small spans

will immediately create large spans and the mean span can be expected to increase with increasing resolution number.

For the data sets of the intensive study, the product of resolution number and mean span is plotted versus resolution number as shown in Figure 4.6. This product is a mean span distance since the changing cell dimension is being accounted for by the resolution number. The relationship is clearly not a constant over the range of resolution numbers. An increase in mean span due to boundary relocations and deletions is more than offsetting the decrease in mean due to increasing cell size alone. It is interesting that the net increase so closely fits a linear regression model.

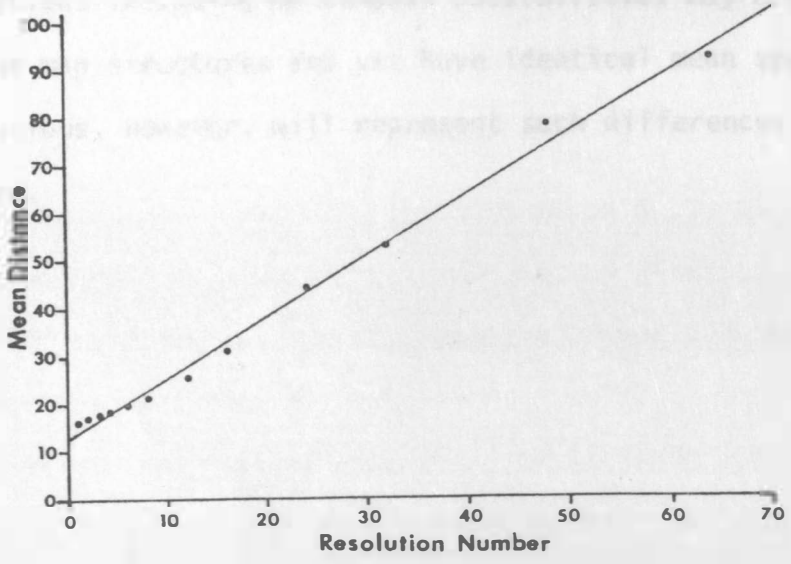



Figure 4.6. Mean Span Distance versus resolution number.

## 4.3 Data Characterization

A mathematical relationship for Figure 4.6 was not sought by standard statistical and/or curve fitting techniques since it was realized that such a relationship - even though clearly existent - would not likely hold for other map data sets. For example maps with multimodal span distributions would have mean distances that change in an altogether different pattern with changing resolution. This would quite possibly make the equivalent of Figure 4.6 non-linear and completely alter the relationship of errors to span distribution mean.

Since the span distribution characterizes the composition of map interboundary distances and spans correspond directly to the compact, sequential geocoding process, estimation of mapping error at a certain cell size should be possible from the relative frequencies of the span distribution rather than the mean span. Various distributions including multimodal possibilities may arise from different map structures and yet have identical mean spans. The span distributions, however, will represent such differences in map structure.

# CHAPTER 5. MAPPING ERROR AND THE SPAN DISTRIBUTION -A POSITIONAL AVERAGE MODEL

The system performance experiments discussed in chapter 3 demonstrated that even though individual mapping unit error for varying sizes and shapes may behave erratically with changing cell size, an average effect over the map is quite well behaved. In chapter 4 the mean span distance was found to logically relate to changing cell size. It was noted, however, that this relationship was probably unique to the particular data set being analyzed and the result not applicable to other data sets.

The usefulness of the span distribution itself to overcome the shortcomings of the mean span was noted. The practical relevance of the span distribution to predicting mapping error depends on the definition of a relationship between them. This chapter will introduce the pursuit of such a relationship.

## 5.1 The Size and Orientation of Cells

Several conceptual relationships between the span distribution and mapping error at a given cell size can be visualized. All depend on the span as a discrete interboundary distance and the cell size which must quantize that distance.

Consider one spatial dimension, an isolated span of n spatial units and a cell dimension of m spatial units. Let the cellularization begin in alignment with one end of the span. If m equals n, the cell would represent the span exactly and there would be no mapping error;

if the m unit cell becomes greater than 2n, there would be 100% mapping error; for m between 1 and 2n the error depends on the divisibility of the span by the cell.

Further consider m=n with end alignment. There would be no mapping error. If the spatial relationship were not end aligned but rather left to chance, there could result up to 49% mapping error; in the cases where two cells join near the midpoint of the span.

For a particular cell dimension of m units the factors which influence the error in mapping a span of length n units are the chance orientation of the cell or cells with respect to the span, and the m-n size relationship. In addition the interdependency of spans in the two spatial dimensions and the interplay of boundary adjustments between adjacent sequential spans of a map transect may warrant investigation.

## 5.2 An Orientation Study

From the discussion of the previous section it is apparent that the position of the cell or cells with respect to the span is a significant source of variation in the corresponding mapping error.

In the study of cell size effects, various cell sizes were generated by aggregation of cells in the reference map. The grouping of k by k cells into a larger cell was accomplished for all integer k which evenly divided the map dimensions. Every aggregation regardless of k began with row one, column one of reference data set. Hence each aggregation used in the study was one of many possible orientations

for that aggregation. In fact for an aggregation factor k there are exactly $k^2$ ways of spatially positioning the enlarged cellular network over the original map.

It was felt that some of the variations in error present in Figures 3.7 and 3.8, particularly for the smaller mapping units WEB, BNE and KRA, arose from the particular positional orientation of the aggregations applied to the data. To evaluate the potential magnitude of this variation, to seek any underlying average trend, and to gain insight into data structure relationship to error, an experiment was undertaken. Several simple closed figures were drawn on cellular reference background. A circle, square, ellipse and rectangle were used. The surrounding background was not considered of mapping interest and not included in mapping error calculations. For each integer aggregation k, the $k^2$ ways of aggregating were all observed and mapping error for each calculated. The $k^2$ observations of each aggregation k allowed calculation of a positional average mapping error. Errors arising from aggregation with respect to one common coordinate point were recorded separately to enable comparison of the average to the behavior observed with the procedure as used on the intensive study area.

Figure 5.1 contains the average and common-reference error graphs versus resolution number for the four simple regions analyzed. The suspicions regarding the presence of chance positional effects in Figures 3.7 and 3.8 appears justified. The error rate for average

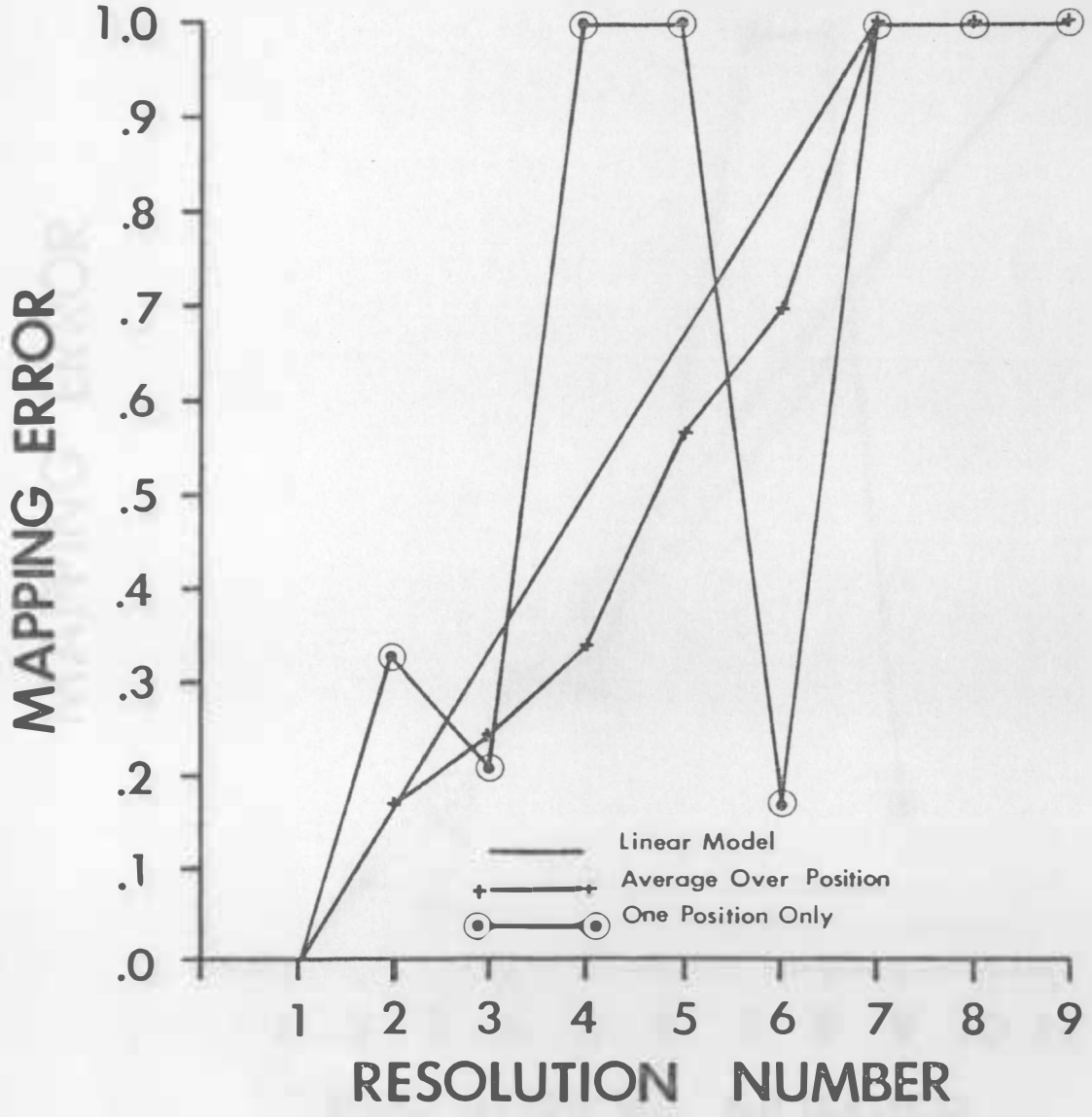



Figure 5.1.a. Mapping error for various resolution numbers in the case of a simple closed circle.

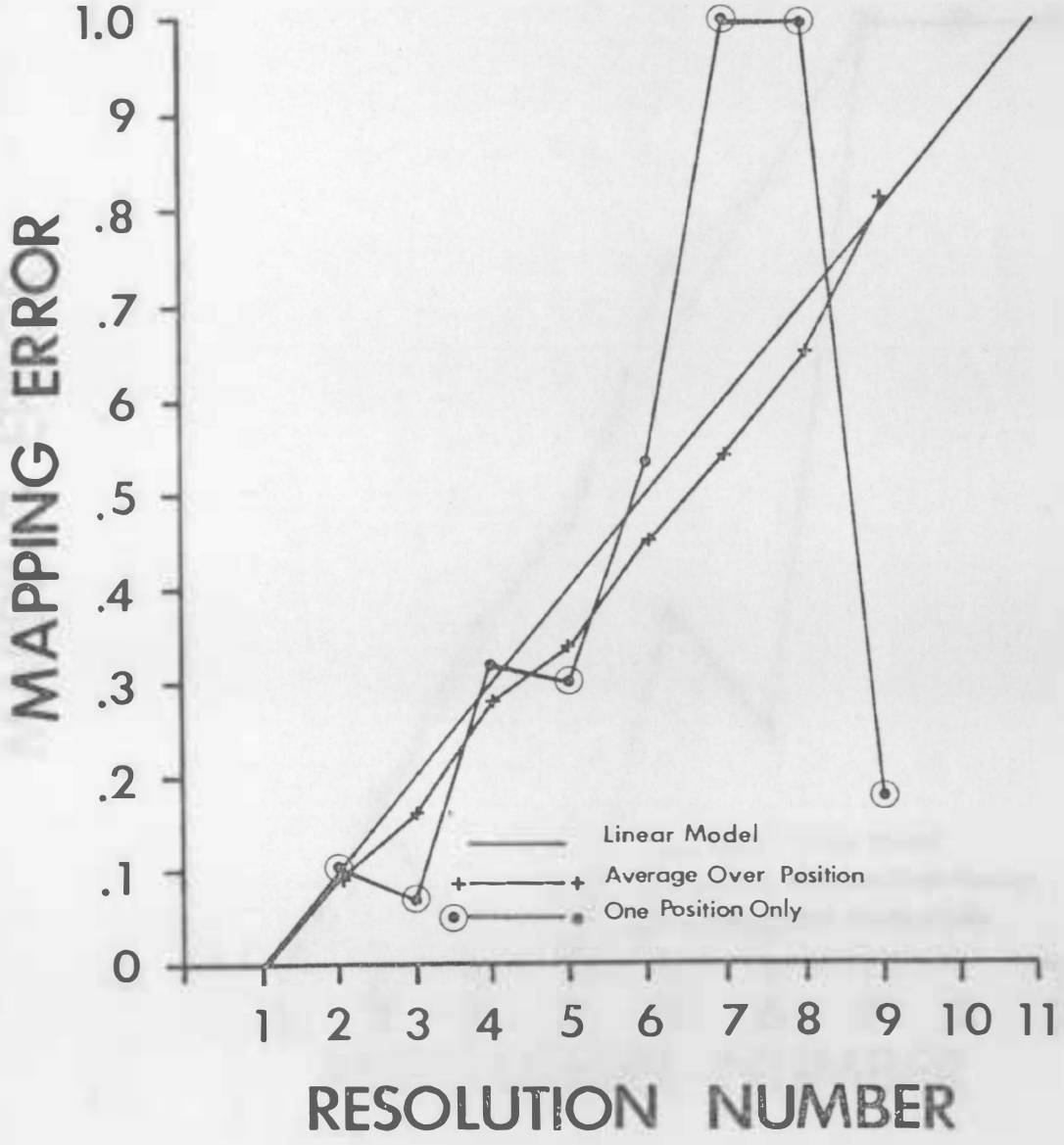



Figure 5.1.b. Mapping error for various resolution numbers in the case of a simple closed ellipse.

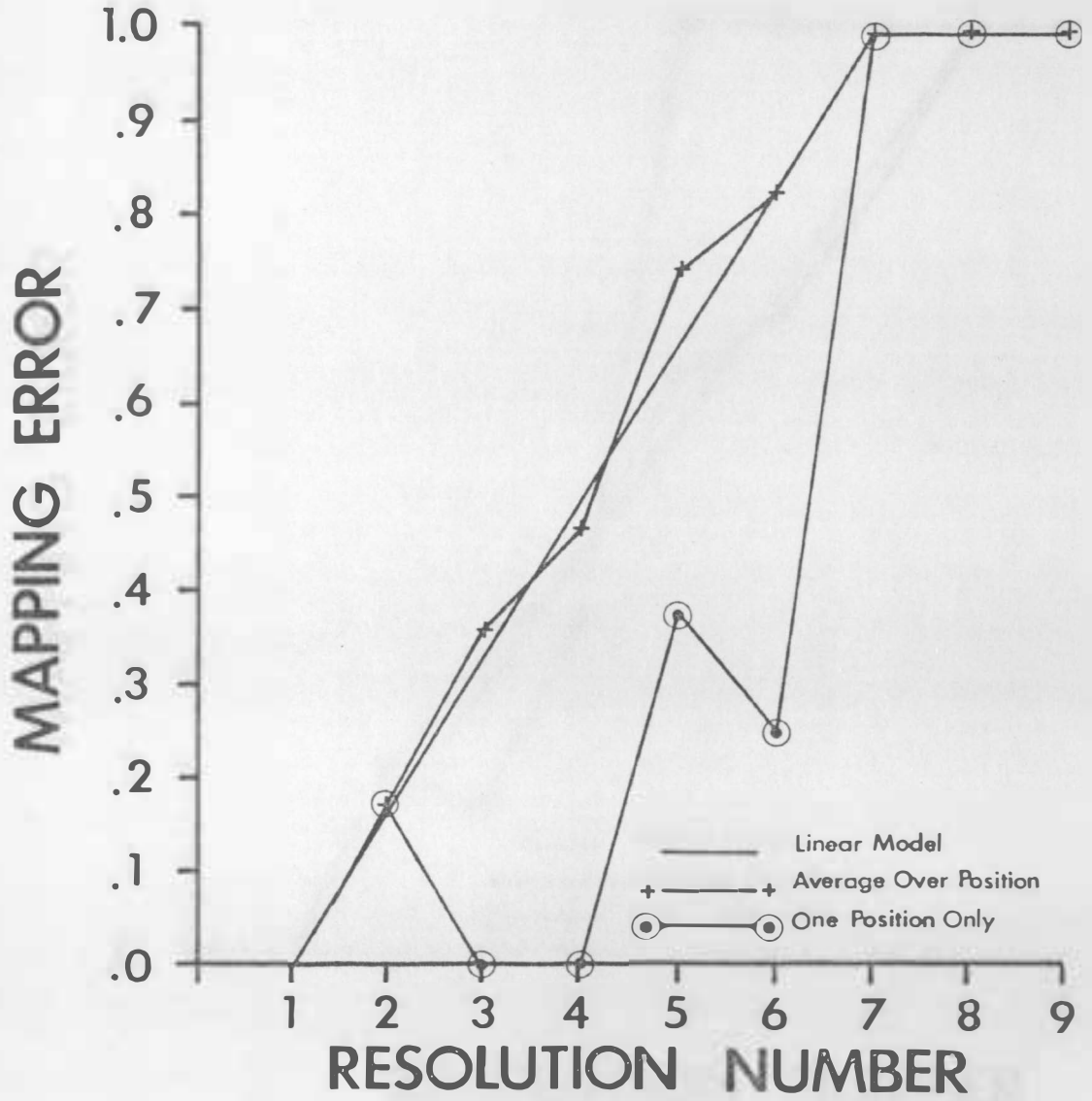



Figure 5.1.c. Mapping error for various resolution numbers in the case of a simple closed rectangle.

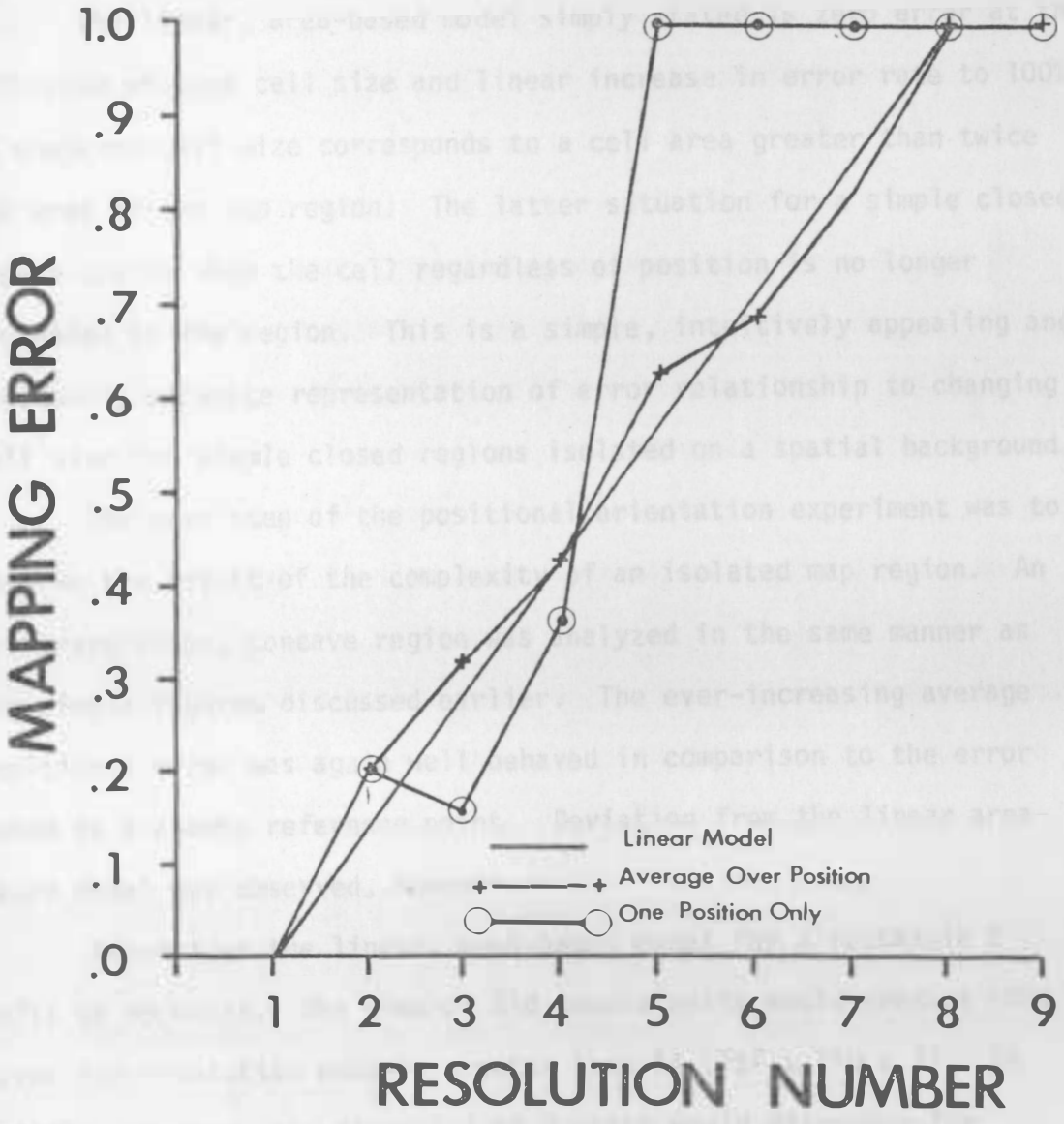



Figure 5.1.d. Mapping error for various resolution numbers in the case of a simple closed square.

positional aggregation is a much more well behaved and non-decreasing function. In fact the proximity to a linear area-based model is quite pleasing.

The linear, area-based model simply stated is zero error at the reference or base cell size and linear increase in error rate to 100% at whatever cell size corresponds to a cell area greater than twice the area of the map region. The latter situation for a simple closed region occurs when the cell regardless of position is no longer dominated by the region. This is a simple, intuitively appealing and reasonably accurate representation of error relationship to changing cell size for simple closed regions isolated on a spatial background.

The next step of the positional orientation experiment was to observe the effect of the complexity of an isolated map region. An arbitrary shape, concave region was analyzed in the same manner as the simple figures discussed earlier. The ever-increasing average positional error was again well behaved in comparison to the error based on a common reference point. Deviation from the linear area-based model was observed, however.

Reconsider the linear, area-based model for a rectangle 2 units by 55 units. The area of 110 square units would predict 100% error for resolution numbers greater than 14 ($15^2 \geq 110 \times 2$). In reality, however, the dimension of 2 units would disappear for resolution numbers greater than 5. Clearly the linear area-based model applies only to restricted cases.

It is asserted as self-evident that the averaging over the many possible aggregation positions of cells with respect to a single closed region is equivalent to the averaging over many randomly dispersed identical regions with respect to a single aggregation position. This extends the experiment results to the case of the intensive study map and allows the experimental results to be used to explain the error behavior in Figures 3.6 through 3.8. The conclusion is also drawn that area is as ambiguous as mean span in prediction of mapping error. As concluded in the mean span study, a summary parameter which is not uniquely related to the inter-boundary distance distribution cannot be expected to relate predictably to the mapping error in all cases.

The conclusion is also drawn that mapping error relates well to changing cell size if effects of positioning the cell can be averaged. This simply implies that analysis of mapping error versus cell size is justified over a map segment as observed in Figure 3.6 but not for isolated mapping units as in Figures 3.7 and 3.8.

## 5.3 The Span Distribution and Mapping Error

A conclusion of the previous section and also Chapter 4 was that interboundary distance cannot be usefully summarized by any parameter (region area) or statistic (mean span) which is not uniquely descriptive of the interboundary distance distribution. The distribution itself is uniquely characteristic of the corresponding map data.

Recall that the span distribution is discrete. Only fixed integer multiples of some linear spatial unit are present. Actual interboundary distances are continuous and may take on any real value. The span distribution must be derived with a sufficiently small spatial unit so as to avoid quantization impact on the interboundary distance characteristics of the map. "Sufficiently small" is not impossible to evaluate nor is it so small as to be prohibitively impractical. Discussion in later chapters should make this apparent.

The span distribution has ordinates of relative frequency versus abcissas of discrete span size. A graph of mapping error versus changing cell size has ordinates of percent error and abcissas of cell size. Earlier discussions suggested that physical relationships of cell size and span size are the source of mapping error. What is needed is a matrix of error components for m values (cell size) versus n values (span sizes). This matrix multiplied by the span distribution vector would yield the error prediction vector on cell size m. Figure 5.2 diagrams this proposed relationship between the span distribution and the mapping error. Mathematically

$$\overline{e}(m) = \overline{f}(n) \cdot \overline{g}(n,m)$$

The $\overline{g}(n,m)$ matrix would represent universal physical relationships between various cell and span sizes in terms of a corresponding mapping error. The span distribution $\overline{f}(n)$ would represent the unique character of the particular map---the actual combination of spans which comprise the map. The matrix product will yield a vector of error fractions versus cell sizes m for the map represented by $\overline{f}(n)$.

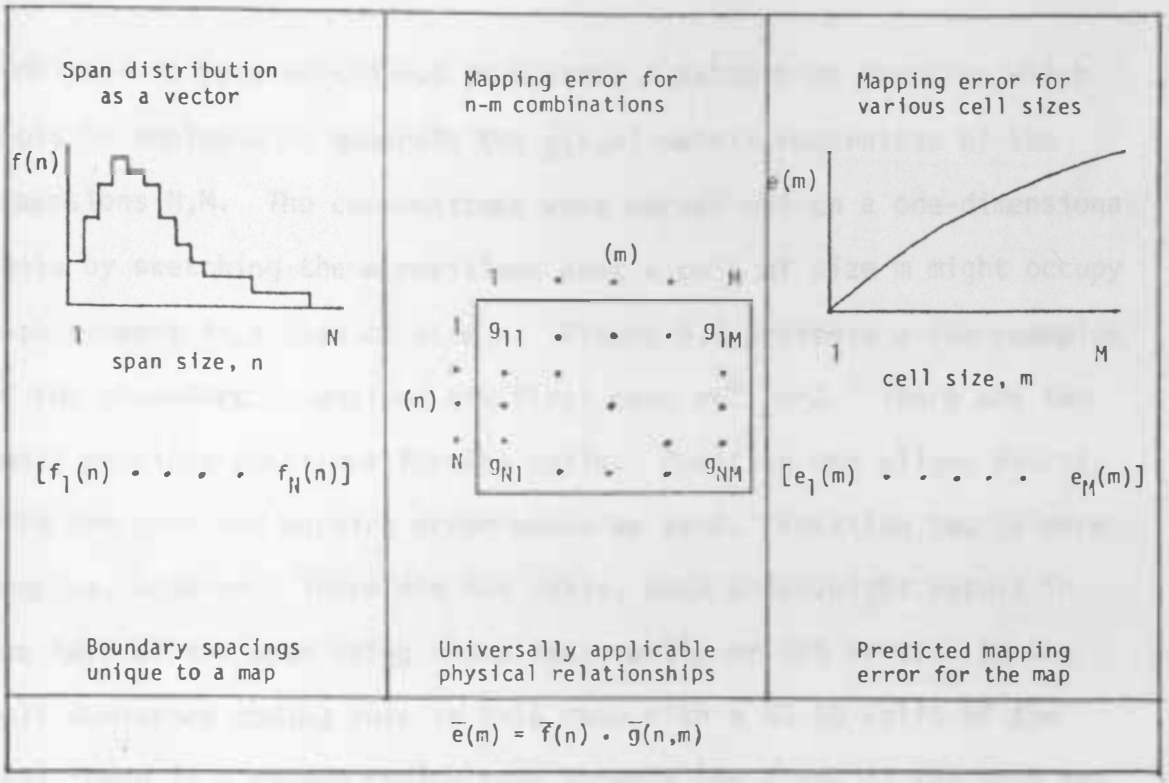



Figure 5.2 Components of the proposed mathematical relationship between span distribution and mapping error vectors.

## 5.4 A Positional Average Model

The problem at hand is to determine $\bar{g}(n,m)$. Earlier experiments pointed to the chance orientation of cells with respect to spans as a source of wide variation in mapping error. The experiments were encouraging, however, in the discovery of a well behaved positional average error. For this reason a positional average error fraction was considered a reasonable starting model for $\bar{g}(n,m)$.

The model was pursued by physically working out all cell positions for each n-m combination (span size-cell size). Enough

combinations were worked out to observe a pattern or function which could be employed to generate the $\bar{g}(n,m)$ matrix regardless of the dimensions N,M. The combinations were worked out on a one-dimensional basis by sketching the m positions that a cell of size m might occupy with respect to a span of size n. Figure 5.3 presents a few examples of the procedure. Consider the first case n=2, m=2. There are two (m=2) possible positions for the cells. Position one aligns exactly with the span and mapping error would be zero. Position two is more complex, however. There are two cells, each which might result in one half of the span being coded incorrectly or 50% error. By the cell dominance coding rule in this case with a 50-50 split of the cell there is a random choice made between the class of the span and the class of data outside the span. The calculation is (2 cells) (½ chance of error)(50% error)--- which is the 50% tabulated.

For the case n=2, m=4, position one, the entire span, 100%, might be error if the cell is coded to the data class outside the span. This can happen on the random tie breaker with a 50% chance. Hence ½ chance of 100% mapping error is the 50% tabulated. From the case n=2, m=5 it should become obvious that all m > 2n are 100% average error since the span is never capable of dominating the cell.

The procedure sketched in Figure 5.3 was continued for n=1,2, ...10 and m=2,3,...6. Average error was tabulated as fractions in reduced form. Observation of the column m=2, n=1,2,...10 revealed an obvious pattern in denominators which led to expression of all denominators as the value of the product m times n. Table 5.1 contains

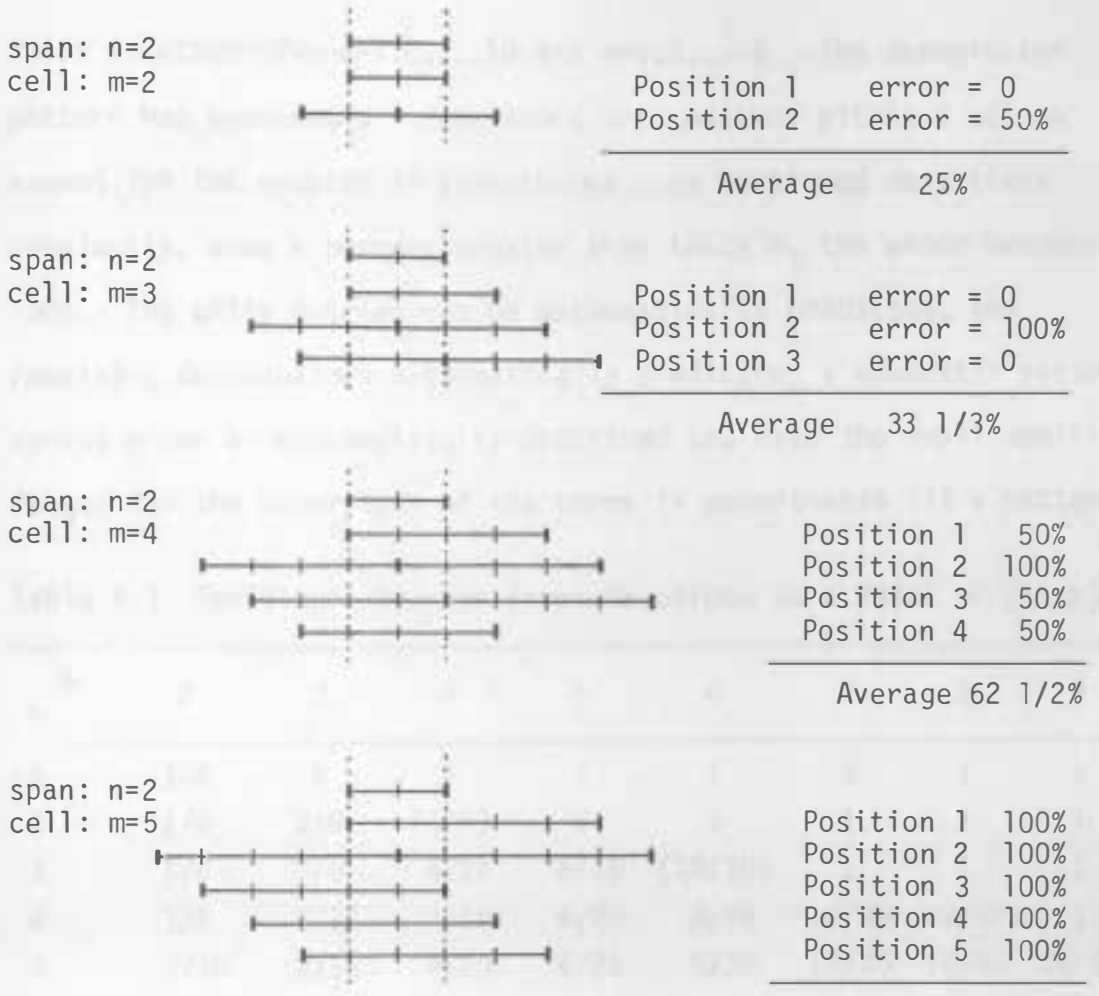



Figure 5.3 The procedure for observing $\bar{g}(n,m)$ by systematic sketches of the m positions of a cell of size m with respect to a span of size n. See text for discussion.

these fractions for n=1,2,...10 and m=2,3,...9. The denominator pattern has been noted. Numerators are constant within a column except for the entries in parentheses. As mentioned many times previously, when m becomes greater than twice n, the error becomes 100%. The unity entries can be mathematically predicted, the remaining denominators mathematically predicted, a numerator pattern versus m can be mathematically described and even the small additive factor for the numerators of the terms in parenthesis fit a pattern.

Table 5.1 Positional Average Error Fractions as a Model of $\bar{g}(n,m)$.

| n \ m | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1/2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1/4 | 2/6 | (5/8) | 1 | 1 | 1 | 1 | 1 |
| 3 | 1/6 | 2/9 | 4/12 | 6/15 | (12/18) | 1 | 1 | 1 |
| 4 | 1/8 | 2/12 | 4/16 | 6/20 | 9/24 | 12/28 | (22/32) | 1 |
| 5 | 1/10 | 2/15 | 4/20 | 6/25 | 9/30 | 12/35 | 16/40 | 20/45 |
| 6 | 1/12 | 2/18 | 4/24 | 6/30 | 9/36 | 12/42 | 16/48 | 20/54 |
| 7 | 1/14 | 2/21 | 4/28 | 6/35 | 9/42 | 12/49 | 16/56 | 20/63 |
| 8 | 1/16 | 2/24 | 4/32 | 6/40 | 9/48 | 12/56 | 16/64 | 20/72 |
| 9 | 1/18 | 2/27 | 4/36 | 6/45 | 9/54 | 12/63 | 16/72 | 20/81 |
| 10 | 1/20 | 2/30 | 4/40 | 6/50 | 9/60 | 12/70 | 16/80 | 20/90 |

The entire matrix can be generated to any dimensions required by the following rules:

(1) for $m > 2n$, $\bar{g}(n,m) = 1$

(2) for $m \leq 2n$ all denominators are $mn$

(3) for $m \leq 2n$ all numerators fit the pattern num(m) = num (m-1) + [m/2]

(4) additive numerator corrections for m = 2n are

$$\sum_{i=1}^{n-1} i$$

As an example consider the m=6, n=3 entry. Rule 1 does not apply. Rule 2 makes the denominator equal to 6 times 3 or 18. Rule 3 says that the numerator will be the numerator for m=5 , which is 6, plus the largest integer in m/2 which is 3 resulting in a numerator of 9. Rule 4 says that a numerator correction is required by adding (i)+(i+1)...(n-1) or 1+2=3. Therefore the numerator is 9+3 or 12 and the m=6, n=3 entry is 12/18.

Some physical bases for the entries may be noted. The source of the unity entries is physically obvious. The denominator m times n arises from averaging over m positions and mapping error being part or all of the n-unit span expressed as a fraction of n. The additive factors in the terms where m=2n arise from random tie breaking in assignment of cell dominance. A physical reason for the remaining numerator sequence is not immediately apparent.

## 5.5 Prediction of Mapping Error

With the rules defined for generating $\bar{g}(n,m)$ as a positional average error fraction matrix, it was a reasonably simple task to generate the matrix on a digital computer. The span distribution obtained from the base data set of the performance experiment in

chapter 3 encompassed spans to 120 units in length and the aggregations used in the study of changing cell size ranged from 2 to 64. Hence the $\bar{g}(n,m)$ matrix generated was 120 elements long (n) and 64 elements wide (m). The span distribution, $\bar{f}(n)$, a 120 element vector, was then multiplied by $\bar{g}(n,m)$ to predict the mapping error vector $\bar{e}(m)$ in accordance with Figure 5.2.

The predicted mapping error and the experimental mapping error for cell sizes up through 64 spatial units are compared in Figure 5.4. There is a consistant over-estimation of mapping error by consideration of positional average relationships between spans and cells in a one dimensional model.

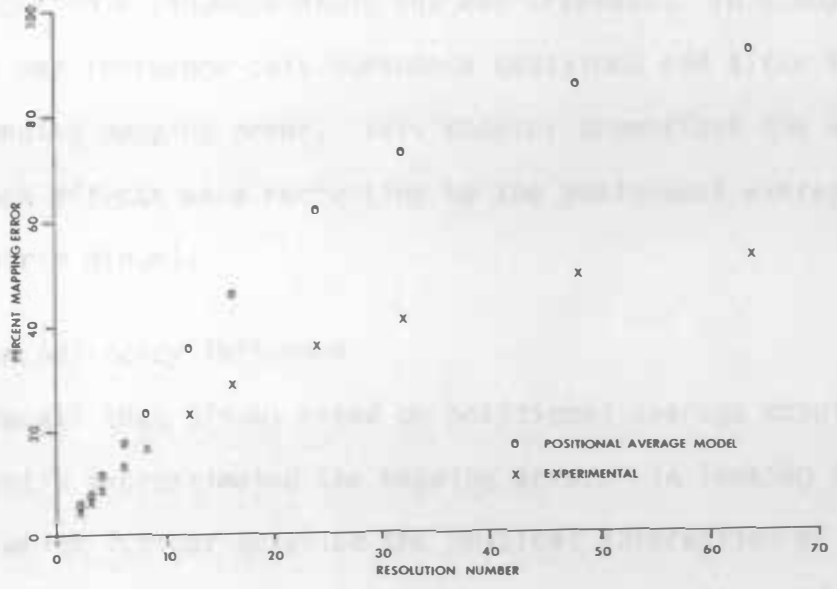



Figure 5.4 Predicted versus experimental mapping error. $\bar{g}(n,m)$ used was the positional average error fraction matrix.

# CHAPTER 6 MAPPING ERROR AND THE SPAN DISTRIBUTION - CORRECTION FOR SPAN ADJACENCIES

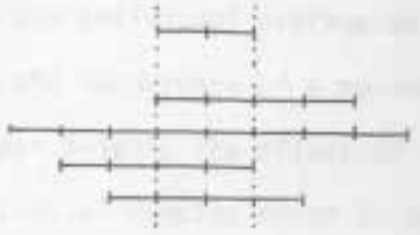
The positional average model of Chapter 5 was derived by observing isolated spans of n spatial units and the m positions of a cell of m spatial units. Isolation was introduced by consideration of the span with respect to a uniform, extensive background. In each case mapping error was considered in terms of the n units interior to the span and no consideration was given to correct mapping of the background. Cell dominance determined whether or not the units of the span were correctly or incorrectly mapped. In actual map applications, spans occur in a sequence along any map transect. This adjacency of spans may influence cell dominance decisions and alter the corresponding mapping error. This chapter summarizes the attempt to model such effects as a correction to the positional average mapping error matrix $\overline{g}(n,m)$.

## 6.1 Span Adjacency Influence

Recall that $\overline{g}(n,m)$ based on positional average mapping error consistently overestimated the mapping error. In looking for other factors which further describe the physical interaction of spans and cells, the factors considered must reduce the magnitude of the estimated mapping error. Consider Figure 6.1. Part A demonstrates the calculation of positional average mapping error when the region surrounding the span is large and homogeneous. In part B an adjacent span of one unit is assumed to be different from the background data

span: n=2

cells: m=4



| | |
|---|---|
| Position 1 | 50% |
| Position 2 | 100% |
| Position 3 | 50% |
| Position 4 | 50% |
| Average | 62½% |

A) the one dimension, positional average calculation.

spans: n=1 and n=2 adjacent

cell: m=4

positions 3 & 4

| | |
|---|---|
| Position 3 | 0% |
| Position 4 | 50% |

B) the one dimensional, positional average calculation when a one unit span is known to be adjacent (A reconsideration of positions 3 and 4 of part A)

Fig. 6.1 Impact of an adjacent span on calculation of positional average mapping error.

as well as different from the data associated with the two-unit span being analyzed. Only positions three and four from the calculation in part A are repeated to demonstrate the impact of adjacency. In position three the presence of the one-unit span causes the data class of the two-unit span to dominate the cell. The cell would be mapped as the data class of the two-unit span and no mapping error (in terms of the span itself) would occur. In part A the same cell position stood a 50-50 chance of totally misrepresenting the span--a 50% error. Clearly for position three, adjacency considerations

could reduce the positional average mapping error. Position four demonstrates the occurrence of a no-impact situation.

Consider briefly the effect of an adjacent span of two units. The calculations of mapping error in part A would not be altered. This would be true for adjacent spans of any size larger than one unit in this example.

The example does demonstrate the desired effect -- a reduction in mapping error magnitude when adjacency effects are considered.

## 6.2 Describing Adjacent Span Corrections

Considering Figure 6.1 part B, position three in comparison to part A, position three, there is a 50% error term removed. This correction is only valid, however, for the case where a one-unit span occurs adjacent to the two-unit span being analyzed. At present, there is no evidence pointing to interdependency of spans. There is no cause-effect relationship known that would dictate the relative frequencies of various spans occurring adjacent to a known span. The assumption is made that the likelihood of an adjacent span of one unit is estimated simply by the relative frequency of one-unit spans in the map, i.e. f(1). Simply stated the assumption is that spans are independent of one another. The 50% error correction for position three should then be weighed by the likelihood of the one-unit span actually occurring. The correction is then described as $\frac{1}{2}f(1)$.

The potential occurrence of multiple spans of corrective influence must also be considered. A notation for this situation is developed in conjunction with Figure 6.2. In both positions one and


Figure 6.2. Two possible positions of a six-unit cell with respect to a two-unit span.

two the error would be 100% if potential adjacent span influences are ignored. Any combination of adjacent spans or a single span which causes the two-unit span to be the dominant class within the six-unit cell or to be tied with another class for dominance of the six-unit cell, will restore a chance of correct mapping.

Position one for example would correctly map the two-unit span if immediately following the two-unit span there were three single-unit spans, $f^3(1)$. If there followed two two-unit spans, a three way tie would be broken by random chance and a one-third chance of correct mapping would arise from the two two-unit spans, $\frac{1}{3}f^2(2)$. Adjacent spans larger than two would dominate the cell and incorrectly map the two-unit span. The notation for correction potential for position three might then be $(1)f^3(1) + (1/3)f^2(2)$ which should be read "fully correct when three single-unit spans occur or one-third chance of fully correct when two two-unit spans occur." The general notation $f^j(k)$ is the likelihood of j sequential spans of size k. The f function is the span distribution and the assumption is independence of spans. The additional notation f(k1)f(k2) is consistent and simply represents the joint occurrence of different size spans, lengths k1 and k2. The notation for the likelihood of adjacent span combinations

occurring must be prefaced by a coefficient representing the mapping error eliminated by that combination. The notation for expressing correction terms is established.

## 6.3 The Correction Matrix

There should exist a matrix $\bar{h}(n,m)$ corresponding one-to-one with the derived $\bar{g}(n,m)$. The entries of $\bar{h}(n,m)$ are the subtractive correction factors which account for reduced mapping error with adjacency influence considered. To determine this matrix,observations of spans of n units in combination with cells of m units must be again made as was done in the analysis reported in Chapter 5. Using the established correction notation, average correction terms must be developed for each entry of $\bar{h}(n,m)$.

This procedure was pursued with hope that a generating pattern could again be defined. Some typical correction expressions obtained are organized in Table 6.1. From the discussion of Figure 6.2 and the entries of Table 6.1, it should be apparent that the larger the cell size for a given span size the greater the number of correction terms. This arises from the larger cell sizes having positions with respect to a span which extend further beyond the span than small cell sizes do. The further cell extension beyond the end of the span provides opportunity for multiple small adjacent spans or larger single spans to have a corrective influence on mapping error. These cases are the higher powers of unit spans and first power of a two-unit span in the m=5 column of Table 6.1, respectively.

Table 6.1 Correction expressions for selected m-n combinations

| n \ m | 3 | 4 | 5 |
|---|---|---|---|
| 2 | $\frac{1}{3}f(1)$ | $\frac{1}{4}f^2(1)+\frac{2}{2}f(1)$ | $\frac{1}{5}f^3(1)+\frac{4}{2}f^2(1)+\frac{2}{2}f(1)+f(2)$ |
| 3 | $\frac{2}{9}f(1)$ | $\frac{1}{6}f^2(1)+\frac{2}{3}f(1)$ | $\frac{2}{15}f^3(1)+\frac{4}{3}f^2(1)+\frac{2}{3}f(1)+\frac{2}{3}f(2)$ |
| 4 | $\frac{2}{12}f(1)$ | $\frac{1}{8}f^2(1)+\frac{2}{4}f(1)$ | $\frac{2}{20}f^3(1)+\frac{4}{4}f^2(1)+\frac{2}{4}f(1)+\frac{2}{4}f(2)$ |
| 5 | $\frac{2}{15}f(1)$ | $\frac{1}{10}f^2(1)+\frac{2}{5}f(1)$ | $\frac{2}{25}f^3(1)+\frac{4}{5}f^2(1)+\frac{2}{5}f(1)+\frac{2}{5}f(2)$ |
| 6 | $\frac{2}{18}f(1)$ | $\frac{1}{12}f^2(1)+\frac{2}{6}f(1)$ | $\frac{2}{30}f^3(1)+\frac{4}{6}f^2(1)+\frac{2}{6}f(1)+\frac{2}{6}f(2)$ |
| 7 | $\frac{2}{21}f(1)$ | $\frac{1}{14}f^2(1)+\frac{2}{7}f(1)$ | $\frac{2}{35}f^3(1)+\frac{4}{7}f^2(1)+\frac{2}{7}f(1)+\frac{2}{7}f(2)$ |
| general | $\frac{2}{mn}f(1)$ | $\frac{2}{mn}f^2(1)+\frac{2}{n}f(1)$ | $\frac{2}{mn}f^3(1)+\frac{4}{n}f^2(1)+\frac{2}{n}f(1)+\frac{2}{n}f(2)$ |
| positional average | $\frac{2}{nm^2}f(1)$ | $\frac{2}{nm^2}f^2(1)+\frac{2}{nm}f(1)$ | $\frac{2}{nm^2}f^3(1)+\frac{4}{nm}f^2(1)+\frac{2}{nm}f(1)+\frac{2}{nm}f(2)$ |

Further inspection of Table 6.1 reveals that n is simply a divisor of all terms in the general expressions and the positional average expressions. Therefore, the n contribution being known, the tabulation should actually be of m, the cell size, and various f(n) which contribute corrective effects. Table 6.2 represents this reorganization of the experimental results. The table entries are the coefficients of the f(n) at the column heading. If all rows for a certain value of m (see m=5) are combined into an expression the expression will be the entry in Table 6.1 for that m and the "positional

Table 6.2 Positional average error correction expressions as coefficients of $\bar{f}(n)$ for various cell sizes.

| m | $f^7(1)$ | $f^6(1)$ | $f^5(1)$ | $f^4(1)$ | $f^3(1)$ | $f^2(1)$ | $f(1)$ |
|---|---|---|---|---|---|---|---|
| 3 | | | | | | | $2/nm^2$ |
| 4 | | | | | | $2/nm^2$ | $2/nm$ |
| 5 | | | | | $2/nm^2$ | $4/nm$ | $2/nm$ |
| 6 | | | | $2/nm^2$ | $4/nm$ | $3/nm$ | $3/nm$ |
| 7 | | | $2/nm^2$ | $4/nm$ | $8/nm$ | $6/nm$ | $3/nm$ |
| 8 | | $2/nm^2$ | $4/nm$ | $8/nm$ | $10/nm$ | $7/nm$ | $4/nm$ |
| 9 | $2/nm^2$ | $4/nm$ | $8/nm$ | $14/nm$ | $10/nm$ | $8/nm$ | $4/nm$ |

| m | $f(2)$ | $[f^4(1)$ | $f^3(1)$ | $f^2(1)$ | $f(1)$ | $1]$ |
|---|---|---|---|---|---|---|
| 5 | | | | | | $2/nm$ |
| 6 | | | | | $4/nm$ | $13/3nm$ |
| 7 | | | | $2/nm$ | $4/3nm$ | $6/nm$ |
| 8 | | | $2/nm$ | $4/3nm$ | $6/nm$ | $7/nm$ |
| 9 | | $2/nm$ | $4/3nm$ | $6/nm$ | $11/nm$ | $8/nm$ |

| m | $f^2(2)$ | $[f^2(1)$ | $f(1)$ | $1]$ | $f^3(2)$ | $f(3)$ | $[f^2(1)$ | $f(1)$ | $1]$ |
|---|---|---|---|---|---|---|---|---|---|
| 7 | | | | $4/nm$ | | | | | $3/nm$ |
| 8 | | | $4/3nm$ | $7/nm$ | | | | $6/nm$ | $7/nm$ |
| 9 | | $4/3nm$ | $7/nm$ | $8/nm$ | $1/nm$ | | $3/nm$ | $11/nm$ | $10/nm$ |

| m | $f(3)$ | $f(2)$ | $f(4)$ | $[f(1)$ | $1$ | $]$ |
|---|---|---|---|---|---|---|
| 9 | $3/nm$ | | | | $4/nm$ | |

average" row. Recall that for every m row in Table 6.2 many values of n (span size) were observed to arrive at a general expression. This table represents an extensive amount of work which is almost certainly never error free at the first attempt. Note further than span combinations, larger spans and multiple spans enter the expressions as m increases. This implies that unless a pattern becomes apparent the expression terms for larger m cannot be known by any other method than manual sketching and tabulation of all combinations.

A decision had to be made to pursue some approximation in order to limit the geometric progression of effort required to expand the table. The f(n) are relative frequencies of spans in the input map. As mentioned earlier the f(n) must be derived with the spatial unit small enough so as to not impact inter-boundary distances. This implies a representative spread in abcissa values and by experience indirectly results in relative frequencies less than 0.20. If typical relative frequencies are indeed less than 0.20, then $f^j(n)$ for $j > 1$ or multiple span terms such as f(1) f(k) will have "event" relative frequencies small enough to be ignored. This approximation is implemented by selecting columns from Table 6.2 which have powers of one and single factors, i.e. f(1), f(2), f(3) etc.

Table 6.3 represents a reorganization of Table 6.2 to include only terms f(1), f(2), f(3) etc. Furthermore the m times n in the denominator is understood and dropped from the entries. The number pattern is expressed as functions of k where k is the largest integer in m/2, i.e. k=[m/2].

Table 6.3 Coefficients of $\overline{f}(n)$ for various m to generate first-order correction expressions for span adjacency effects.

| m | k | f(1) | f(2) | f(3) | f(4) | f(5) | f(6) | f(7) | f(8) |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 1 | $k-\frac{1}{3}$ | | | | | | | |
| 4 | 2 | k | | | | | | | |
| 5 | 2 | k | k | | | | | | |
| 6 | 3 | k | $(2k-1)-\frac{2}{3}$ | | | | | | |
| 7 | 3 | 2k | k | | | | | | |
| 8 | 4 | k | 2k-1 | 2k-1 | | | | | |
| 9 | 4 | k | 2k | $(3k-1)-\frac{3}{3}$ | k | | | | |
| 10 | 5 | k | 2k-1 | 3k-2 | 2k-1 | | | | |
| 11 | 5 | k | 2k | 2k-1 | 3k-1 | k | | | |
| 12 | 6 | k | 2k-1 | 3k-2 | $(4k-4)-\frac{4}{3}$ | 2k-1 | | | |
| 13 | 6 | k | 2k | 3k-1 | 4k-2 | 3k-1 | k | | |
| 14 | 7 | k | 2k-1 | 3k-2 | 4k-4 | 4k-4 | 2k-1 | | |
| 15 | 7 | k | 2k | 3k-1 | 4k-2 | $(5k-4)-\frac{5}{3}$ | 3k-1 | k | |
| 16 | 8 | k | 2k-1 | 3k-2 | 4k-4 | 5k-6 | 4k-4 | 2k-1 | |
| 17 | 8 | k | 2k | 3k-1 | 4k-2 | 5k-4 | 5k-4 | 3k-1 | k |

(all entries require 1/mn)

The Table 6.3 entries are organized as a matrix which matches m sizes to the factors of $\overline{f}(n)$ required to generate a correction expression. Call this matrix $\overline{c}(m,n)$. Multiplying $\overline{c}(m,n)$ by the vector span distribution $\overline{f}(n)$ will yield a vector $\overline{v}(m)$ which contains single term expressions containing n. If $\overline{v}(m)$ is expanded to two

dimensions by inserting values for n, the result will be the desired $\overline{h}(n,m)$, the subtractive correction matrix used to adjust the $\overline{g}(n,m)$ of Chapter 5.

## 6.4 Implementing the Correction

To implement the procedure outlined in the previous paragraph, some generating pattern must be found to create the $\overline{c}(m,n)$ matrix equivalent to Table 6.3. The small adjustment fractions 1/3, 2/3, 3/3 etc. can be easily generated as n/3 for entries at m=3n so they can be ignored temporarily in seeking a pattern. The form of all entries then becomes a coefficient of k and a subtractive number.

Table 6.4 summarizes the coefficients of k in the $\overline{c}(m,n)$ matrix. This pattern is easy to define. For a column of $\overline{f}(n)$ run zeros from m=1 to m=2n, count upward from 1 at m=2n+1 to m=3n and fill the column from m=3n to whatever limit is desired with the value at m=3n.

The generation of the subtractive numbers is not so obvious. Table 6.5 summarizes the subtractive numbers for the entries of $\overline{c}(m,n)$ in two groups, one for m even and one for m odd. A pattern for generating entries in one group can be used for the other and the groups combined to result in the subtractive number table.

First consider the constant column entries below the lower dividing line at m=3n. The differences between columns fits the pattern 1,1,2,2,3,3 etc. The last line, largest value of m desired, can be generated based on this difference pattern and the columns propagated upward to m=3n.

Table 6.4 Coefficients of K in the entries to the $\bar{c}(m,n)$ matrix.

| m | K | f(1) | f(2) | f(3) | f(4) | f(5) | f(6) | f(7) |
|---|---|---|---|---|---|---|---|---|
| 3 | 1 | 1 | | | | | | |
| 4 | 2 | 1 | | | | | | |
| 5 | 2 | 1 | 1 | | | | | |
| 6 | 3 | 1 | 2 | | | | | |
| 7 | 3 | 1 | 2 | 1 | | | | |
| 8 | 4 | 1 | 2 | 2 | | | | |
| 9 | 4 | 1 | 2 | 3 | 1 | | | |
| 10 | 5 | 1 | 2 | 3 | 2 | | | |
| 11 | 5 | 1 | 2 | 3 | 3 | 1 | | |
| 12 | 6 | 1 | 2 | 3 | 4 | 2 | | |
| 13 | 6 | 1 | 2 | 3 | 4 | 3 | 1 | |
| 14 | 7 | 1 | 2 | 3 | 4 | 4 | 2 | |
| 15 | 7 | 1 | 2 | 3 | 4 | 5 | 3 | 1 |
| 16 | 8 | 1 | 2 | 3 | 4 | 5 | 4 | 2 |
| 17 | 8 | 1 | 2 | 3 | 4 | 5 | 5 | 3 |

The wedge of entries between the dividing lines is generated from the difference sequence in entries down a column. The first entry is one in each column. This occurs for m=2 l + 2 if m is even or m=2 l + 3 if m is odd; the column heading is f(l). The differences down the column are 3,5,7,9 etc.

Utilizing the observed patterns, a computer program was written to generate the matrices equivalent to Tables 6.4 and 6.5. Multiplying

Table 6.5 Subtractive numbers for entries to the $\overline{c}(m,n)$ matrix.

| m | k | f(2) | f(3) | f(4) | f(5) | f(6) | f(7) | f(8) | f(9) |
|---|---|---|---|---|---|---|---|---|---|
| 7 | 3 | 0 | | | | | | | |
| 9 | 4 | 0 | 1 | | | | | | |
| 11 | 5 | 0 | 1 | 1 | | | | | |
| 13 | 6 | 0 | 1 | 2 | 1 | | | | |
| 15 | 7 | 0 | 1 | 2 | 4 | 1 | | | |
| 17 | 8 | 0 | 1 | 2 | 4 | 4 | 1 | | |
| 19 | 9 | 0 | 1 | 2 | 4 | 6 | 4 | 1 | |
| 21 | 10 | 0 | 1 | 2 | 4 | 6 | 9 | 4 | 1 |
| 23 | 11 | 0 | 1 | 2 | 4 | 6 | 9 | 9 | 4 |
| 25 | 12 | 0 | 1 | 2 | 4 | 6 | 9 | 12 | 9 |
| 27 | 13 | 0 | 1 | 2 | 4 | 6 | 9 | 12 | 16 |
| 4 | 2 | | | | | | | | |
| 6 | 3 | 1 | | | | | | | |
| 8 | 4 | 1 | 1 | | | | | | |
| 10 | 5 | 1 | 2 | 1 | | | | | |
| 12 | 6 | 1 | 2 | 4 | 1 | | | | |
| 14 | 7 | 1 | 2 | 4 | 4 | 1 | | | |
| 16 | 8 | 1 | 2 | 4 | 6 | 4 | 1 | | |
| 18 | 9 | 1 | 2 | 4 | 6 | 9 | 4 | 1 | |
| 20 | 10 | 1 | 2 | 4 | 6 | 9 | 9 | 4 | 1 |
| 22 | 11 | 1 | 2 | 4 | 6 | 9 | 12 | 9 | 4 |
| 24 | 12 | 1 | 2 | 4 | 6 | 9 | 12 | 16 | 9 |
| 26 | 13 | 1 | 2 | 4 | 6 | 9 | 12 | 16 | 16 |
| 28 | 14 | 1 | 2 | 4 | 6 | 9 | 12 | 16 | 20 |

the entries of Table 6.4 by k and subtracting the entries of Table 6.5 results in a numerical equivalent to Table 6.3 which was called $\overline{c}(m,n)$. This $\overline{c}(m,n)$ was then multiplied by $\overline{f}(n)$, the span

distribution, to obtain $\bar{v}(m)$. Since the $\bar{v}(m)$ vector is developed from Table 6.3 entries which include the common factor 1/mn, the vector can be expanded to two dimensions by including values of n in the entries and using n as the second dimension. The resultant is $\bar{h}(n,m)$ which was the desired correction matrix for $\bar{g}(n,m)$ from Chapter 5. The entries are all fractions representing the over-estimate of mapping error in $\bar{g}(n,m)$.

## 6.5 Applying the Correction

The correction may be handled two ways. The $\bar{h}(n,m)$ may be subtracted from $\bar{g}(n,m)$ and $\bar{e}(m)$ generated as in Chapter 5. Alternatively $\bar{f}(n)$ multiplied by $\bar{h}(n,m)$ will generate an error correction $\bar{e}_c(m)$. This error magnitude can be subtracted from the error model vector $\bar{e}(m)$. The second approach was taken to enable observation of the span adjacency error correction terms $\bar{e}_c(m)$. Percentage mapping error for the positional average model, the span adjacenty correction, the corrected model and the experiment itself are compared in Table 6.6. In Figure 6.3 the positional average error model and the experimental results are repeated from Figure 5.4 and the span-adjacency-corrected model included for comparison.

The span adjacency correction to the model appears to account for a substantial portion of the deviation between the experimental results and the uncorrected model predictions for cell sizes larger than 32 spatial units. With the correction however, the form of the predicted mapping error curve has been altered and no longer

Table 6.6 Mapping error comparisons for the positional average model, the span adjacency correction, the corrected model and the experimental results.

| cell size m | $\bar{e}(m)$ positional average | $\bar{e}_c(m)$ span adjacency correction | $\bar{e}(m)-\bar{e}_c(m)$ corrected model | observed experimental results |
|---|---|---|---|---|
| 2 | 5.33 | 0 | 5.3 | 3.8 |
| 3 | 7.14 | 0 | 7.1 | 5.8 |
| 4 | 10.94 | 0 | 10.9 | 8.8 |
| 6 | 17.00 | 0.17 | 16.8 | 12.8 |
| 8 | 23.06 | 0.48 | 22.5 | 16.0 |
| 12 | 35.16 | 1.89 | 33.3 | 22.6 |
| 16 | 45.41 | 4.22 | 41.2 | 28.0 |
| 24 | 61.18 | 10.23 | 50.9 | 35.4 |
| 32 | 72.26 | 17.39 | 54.9 | 40.4 |
| 48 | 85.16 | 32.10 | 53.1 | 48.0 |
| 64 | 91.52 | 44.22 | 47.3 | 52.3 |

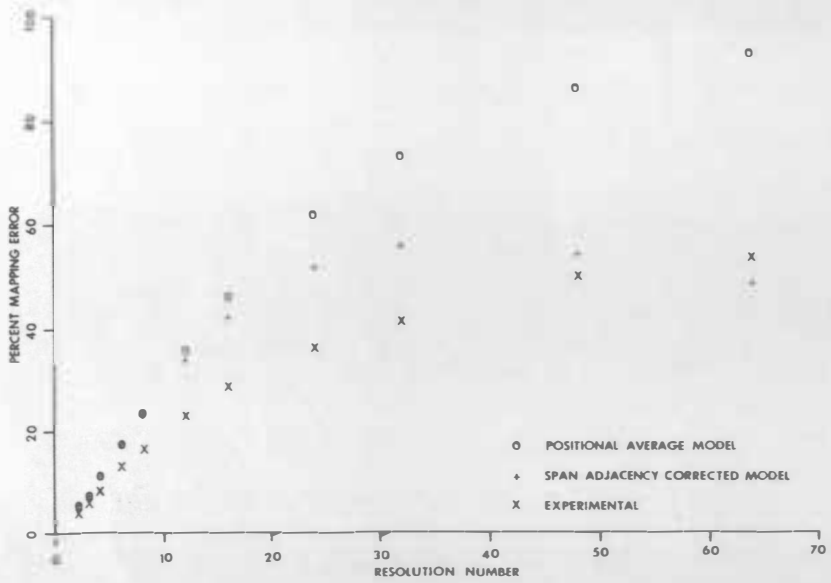



Figure 6.3 The experimentally observed mapping error compared to the positional-average model and the span-adjacency-corrected model.

"parallels" the appearance of the experimental results as it did before correction. Recall also that only the first order terms of span adjacency correction were utilized. If higher order terms and cross products had been included, the only possible effect would be an additional decrease in predicted mapping error ordinates throughout the cell size range. If the inclusion of higher order correction terms would bring the predicted error and experimentally observed error closer together in the cell size range 12 to 32, then the predicted error would probably far underestimate the experimentally observed error at larger cell sizes (beyond 32). This may be acceptable since two-dimensional interactions have not been included in the model.

# CHAPTER 7. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

## 7.1 Summary

The background provided on geoinformation systems outlined motivations, designs, comparisons, performance measurement techniques and suggested cell size selection bases for users of a grid system.

The Remote Sensing Institute System, AREAS, was designed and implemented by the author in accordance with techniques and system recommendations in the literature. That is, user concerns for costs, storage and analysis efficiency and product flexibility were continually incorporated into design decisions. It is believed that this user-centered implementation distinctively separates AREAS from other geoinformation systems which "do the job".

The basic and auxiliary data sets discussed in Chapter Three facilitated mapping and inventory accuracy analyses as measures of system performance with changing cell size. In Chapter Four the analysis of map characteristics is reported. Chapters Five and Six discuss the modelling of the relationship between a particular map and the system performance for that map.

## 7.2 Conclusions

The following observations and conclusions are made as a result of the research reported in this paper:

(1) In accordance with procedures and recommendations found in the literature, AREAS was developed with particular concern for the

user. Storage economy and operating efficiency have been jointly achieved in the system design.

(2) Although in the case of a single mapping unit, the behavior of mapping and inventory errors with increasing cell size may behave very erratically, an average affect over many mapping units in a map segment is apparently a well behaved non-decreasing function.

(3) The interboundary distance distribution does uniquely characterize a map.

(4) No summary parameter such as mean or mode of the distribution will ever provide a generally applicable prediction of performance since such parameters are not unique to a map.

(5) It appears that in the general case, the scanning of map distances in the two orthogonal directions corresponding to the cell dimensions, will yield distributions statistically dissimilar, hence implying that map sampling must be designed to randomly include transects in each dimension.

(6) A universal matrix model of the one-dimensional, average position of cells on a map does exist and may be generated to any desired dimensions as required.

(7) The influence of the adjacency of interboundary distances in one dimensional transects of a map is a significant, model-corrective factor. The first-order approximating model is also universal and generateable to whatever matrix dimensions are required.

(8) The total one-dimensional model of positional average corrected for first-order adjacency does not encompass enough of the

physical interrelationships of cell size to map interboundary distances to predict mapping performance accurately.

(9) The existance of universally applicable model components in the studies undertaken is most encouraging to the continued development of the model matrix required to transform the interboundary distance distribution into a prediction of mapping accuracy.

## 7.3 Recommendations

The long range objective of the ongoing research in applications of a cellular information system is to define a procedure for selecting and justifying cell size. The composition of such a procedure has begun to crystallize. Sampling of map interboundary distances will estimate the distribution. A matrix model of physical relationships between cell sizes and map distances will transform that distribution into an estimate of mapping accuracy with various cell sizes. With knowledge of the operating costs of a particular processing system, AREAS or other cellular systems, the trade off between accuracy and cost for various cell sizes should enable cell selection to be made according to user needs.

The research reported and conclusions drawn are contributory to the understanding of underlying physical relationships which must exist for the outlined procedure to ever become reality. The gap between the accomplishments outlined and the selection procedure envisioned will be filled when the following recommendations have been successfully investigated:

(1) The modelling work should pursue two dimensional interactions in a fashion parallel to the sequence of investigations in the one dimensional case reported. This might require consideration of a joint or two-dimensional interboundary distance distribution. Higher order adjacency correction terms may also be required to bring predicted mapping error into closer agreement with experimental results. These pursuits would promote the derivation of the universal matrix model required.

(2) As a follow-up to model development as suggested in recommendation one, a analogous development could be pursued for inventory performance. Alternatively, inventory performance may be predictable from mapping performance when that model is finalized.

(3) A sampling technique, or at least guidelines, must be defined to enable a user to estimate the required form of inter-boundary distance distribution to drive the matrix model. No single distribution model will apply and, furthermore, the need for an estimate of the distribution (as opposed to an estimate of a parameter of a particular distribution) does not lend itself to known sampling techniques with defineable confidence intervals. For these reasons, this component of the desired overall selection procedure may take the form of empirical guidelines.

(4) Users of cellular systems, AREAS included, need to analyze operating costs in a detail that enables processing cost estimate from a given cell size. This is not always a trivial task inasmuch as some processes depend on several other parameters as well.

(5) Finally, the summary recommendation is made that the research be continued toward achievement of the cell size selection procedure. The work performed thus far indicates that such an approach is feasible and the requisite physical relationships do exist.

# BIBLIOGRAPHY

[1] Joe D. Nichols, "Characteristics of Computerized Soil Maps" in Soil Science Society of America Proceedings, Vol. 39, pp. 927-932, 1975.

[2] Phillip A. McDonald and Jerry D. Lent, "MAPIT-A Computer Based Data Storage, Retrieval and Update System for the Wildland Manager" in Proceedings of the 38th Annual Meeting of American Society of Photogrammetry, Washington, D.C., March 12-17, pp. 370-397, 1972.

[3] W.A. Radlinski, "Modern Land Data Systems - A National Objective", Opening address to 1977 Annual Convention of American Society of Photogrammetry and American Congress of Surveying and Mapping, published - Photogrammetric Engineering and Remote Sensing, Vol. XLIII No. 7, pp. 887-890, July 1977.

[4] "IRIS - Illinois Resources Information System", University of Illinois Center for Advanced Computations Feasibility Study - Final Report, Urbana, Illinois, 1972.

[5] R.F. Tomlinson, "Geo-Information Systems and the Use of Computers in Handling Land Use Information" in Conference on Land Use Information and Classification, sponsored by Department of Interior of U.S. Geological Survey and the National Aeronautics and Space Administration, Washington, D.C., June 28-30, 1971.

[6] Nevin A. Bryant and Albert L. Zobrist, "IBIS: A Geographic Information System Based on Digital Image Processing and Image Raster Datatype" in Symposium Proceedings of Machine Processing of Remotely Sensed Data, Purdue University, pp. 1A-1, 1A-7, June 29-July 1, 1976.

[7] D. Steiner and T. Stanhope, "Data Base Development" Chapter 1 in Geographical Data Handling prepared for UNESCO/IGU Second Symposium on Geographicl Information Systems by International Geographical Union Commission on Geographical Data Sensing and Processing, pp. 36-103, August, 1972.

[8] R.F. Tomlinson, editor. Geographical Data Handling published for UNESCO/IGU Second Symposium on Geographical Information Systems by International Geographical Union Commission on Geographical Data Sensing and Processing, Ottawa, Ontario Canada, 1972.

[9] Richard L. Phillips, "Computer Graphics in Urban and Environmental Systems", Proceedings of IEEE, Vol. 62, No. 4., pp. 437-452, April 1974.

[10] Nevin A. Bryant and Albert L. Zobrist, "Integration of Socioeconomic Data and Remotely Sensed Imagery for Land Use Applications" in Proceedings of 2nd Annual Pecora Symposium sponsored by American Society of Photogrammetry and the U.S. Geological Survey, pp. 120-130, Oct. 25-29, 1976.

[11] "A Land Classification Method for Land Use Planning" by the Land Use Analysis Laboratory at Iowa State University, 1973.

[12] George Smith, Kris Van Gorkom, A.A. Dyer et al., Colorado Environmental Data Systems a final Report to the Colorado Department of Natural Resources by the College of Forestry and Natural Resources, Colorado State University, Fort Collins, Colorado, 1973.

[13] D. Sinton. "Introduction to Spatial Data Manipulation and Analysis" Chapter 8 in Geographical Data Handling prepared for UNESCO/IGU Second Symposium on Geographical Information Systems by International Geographical Union Commission on Geographical Data Sensing and Processing, p. 719, Aug. 1972.

[14] P. Switzer. "Estimation of the Accuracy of Qualitative Maps" in Display and Analysis of Spatial Data, NATO Advanced Study Institute, edited by John C. Davis and Michael McCullagh, John Wiley and Sons, New York, 1975.

[15] Dr. J.L. Van Genderen and B.F. Lock "Testing Land-Use Map Accuracy" in Photogrammetric Engineering and Remote Sensing, Vol. 43, No. 9, pp. 1135-1137, Sept. 1977.

[16] Michael R. Hord and William Brooner "Land-Use Map Accuracy Criteria" in Photogrammetric Engineering and Remote Sensing, Vol. 42, No. 5, pp. 671-677, May 1976.

[17] Michael E. Wehde, "THE OPERATION OF AREAS: Area REsource Analysis System", Report SDSU-RSI-78-10, 1978.

[18] Harry W. Smedes, "Land-Use Planning Aided by Computer Cellular Modelling/Mapping System to Combine Remote Sensing, Natural Resources, Social and Economic Data" in Proceedings of 9th International Symposium on Remote Sensing of Environment, Vol 1, pp. 289-291, April 1974.

[19] Michael E. Wehde "THE IMPLEMENTATION OF AREAS: Area REsource Analysis System", Report SDSU-RSI-78-09, 1978.

[20] Charles R. Meyers Jr., Richard C. Durfee and Thomas Tucker, "Computer Augmentation of Soil Survey Interpretation for Regional Planning Applications" Oak Ridge National Laboratory Report ORNL-NSF-EP-67, April 1974.

[21] Joe D. Nichols and Lindo J. Bartelli, "Computer-Generated Interpretive Soil Maps" in Journal of Soil and Water Conservation, Vol. 29(5), pp. 232-235, 1974.

[22] E.L. Amidon and G.S. Arin "Algorithmic Selection of the Best Method for Compressing Map Data Strings" in Communications of the Association for Computing Machinery, Vol. 14, No. 12, Dec. 1971.