A REVIEW OF THE ENVIRONMENTAL RESOURCE MAPPING SYSTEM AND A PROOF THAT IT IS IMPOSSIBLE TO WRITE A GENERAL ALGORITHM FOR ANALYSING INTERACTIONS BETWEEN ORGANISMS DISTRIBUTED AT LOCATIONS DESCRIBED BY A LOCATIONALLY LINKED DATABASE AND PHYSICAL PROPERTIES RECORDED WITHIN THE DATABASE.

BRYAN HALL.

Master of Applied Science Degree in Environmental Science
University of Western Sydney, Hawkesbury.

November 1994
Statement of Originality.

This thesis is submitted to satisfy part of the requirements for a Master of Applied Science Degree in Environmental Science at the University of Western Sydney Hawkesbury. The contents are my own work and have not been submitted to any other institution for any award.

Bryan Hall.

Acknowledgments.

In preparing this thesis I have been fortunate in having general assistance from family, friends and colleagues. The support from my parents over years of university education and throughout the many months spent in doing this project has been substantial.

I have benefited greatly from individuals who have encouraged me to never accept anything as being correct until such time as they could convince me that the arguments they presented were inevitably true.

For general deliberations I specifically thank:

* Charles Willock (Physicists) for interesting technical arguments and advice;

* Professor Colin Sutherland (Mathematician) for concise advice on counting, adding and sorting;

* Dr Iain Doust (Mathematician) for a lucid discussion on independence and probability; and.
* Dr Hendrick Grundling (Mathematician) for advice on applying the mathematical
group structure of the scales of measurement.

I thank Robert Blomfield for reading and suggesting improvements to some of
the manuscript and Dr Shelley Burgin for reading, editing and advice in turning a
large collection of disjointed notes and results into a thesis.
Table of Contents.

A REVIEW OF THE ENVIRONMENTAL RESOURCE MAPPING SYSTEM AND A PROOF THAT IT IS IMPOSSIBLE TO WRITE A GENERAL ALGORITHM FOR ANALYSING INTERACTIONS BETWEEN ORGANISMS DISTRIBUTED AT LOCATIONS DESCRIBED BY A LOCATIONALLY LINKED DATABASE AND PHYSICAL PROPERTIES RECORDED WITHIN THE DATABASE.  

Statement of Originality.  
Acknowledgments.  

ABSTRACT.  

Table of Contents.  

List of Figures and Tables.  
List Of Figures For Chapter 2.  
List Of Tables For Chapter 2.  
List Of Figures For Chapter 3.  
List Of Tables For Chapter 3.  
List Of Tables For Chapter 4.  
List Of Tables For Chapter 5.  
List Of Figures For Chapter 6.  
List Of Figures For Chapter 7.  
List Of Tables For Chapter 7.  

CHAPTER ONE: INTRODUCTION  
  Preamble.  
  Ecosystems, Resource Management, And Geographic Information Systems.  
  Geographic Information Systems Planning And Operation.  
  Thesis Outline.  

CHAPTER TWO: DERIVING INFERENCES FROM ECOLOGICAL DATABASES THAT DEFINE A PARTITION AND CONTAIN SINKS OF INFORMATION.  
  Outline of Chapter.  
  Cardinality: Stating Facts In Numerical Terms.  
  The Structure Of Locationally Linked Databases.  
  All Databases That Specify Location Consist Of A Finite Collection Of Location Elements.  
  Using A Map To Represent The World With A Finite Collection Of Location Elements.  
  Establishing A Mathematical Representation.  
  Scales Of Measurement.  
  Describing Location Elements Using Members Of Finite Sets.  
  P - The Set Of Measurable Properties For Each Location Element.  
  Ω – The Set Of All Possible Distinct Property Codings That Can Be Used To Describe Location Elements.  
  Partitioning Sets With Equivalence Relations.  
  Measuring The Environment At Each Member Of The Set Ω.  
Constructing Cases And Classes For Investigating The Distribution Of Organisms With Respect To Coded Properties.  
β And α The Sets Of Classes And Cases Respectively.  
Members Of The Sets Ω And β Need Conform Only To A Nominal Scale.  
Taking A Representative Sample Of Elements From The Set Ω.  
Decomposing The Set β To Establish The Effect Of Coded Properties On A Biological Organism.  
Sinks Of Information And Stochastic Processes.  
Breaking Down The Set Ω To Establish A Mathematical Representation To Determine Interactions Between Biological Organisms And Coded Properties.  
There May Exist Distinct Partitions Of Ω And β That Can Be Used To Aggregate Members Of Ω Identically.  
The Property Specification Required To Select A Particular Subset Of Location Elements Need Not Be Unique.  
Analysing For The Effect Of 1 Property When Specifying A Union Of Elements Of Ω.  

Insufficient Conditions For Determining Interactions Between Biological Organisms And Coded Properties.  

Establishing Prediction Using Data From A Locationally Linked Database.  

Supplement to Chapter 2: The Meaning Of Some Terms Used In Chapter 2  

CHAPTER THREE: GENERAL BACKGROUND ON E-RMS.  
Outline of Chapter.  
E-RMS A Personal Computer Software Package Intended To Convert Map Based Environmental Data Into Information.  
E-RMS Is A Raster Based Geographic Information System.  
E-RMS Is Divided Into Distinct Modules  
Database Specification Module.  
Data Import Export Module.  
Map Digitising Module.  
Map Display Analysis Module.  
Predictive Modelling Module.  
Tutorial Database Supplied With The E-RMS Package.  
Lessons In The Tutorial Package.  
Lesson Two.  
Lesson Three.  
Lesson Four.  
Predictive Modelling Module.  

CHAPTER FOUR: E-RMS REQUIRES LOCATION COORDINATES THAT DO NOT EXIST.  
Outline of Chapter.  
The Australian Geodetic Datum.  
Grid Zones, AMG Eastings And Northings.  
Coordinates Stated For The E-RMS Tutorial Database.  

CHAPTER FIVE: LIMITS TO RESOLUTION SET BY ATTAINABLE ACCURACY IN THE E-RMS DEMONSTRATION DATABASE.  
Outline of Chapter.  
Precision And Accuracy.  
Accuracy Required For Survey Data.  
Standards Of Map Accuracy.  
The Problem Of Stating Location.  
Surveying All Observation Points.  
Estimating Location From A Paper Map.  
Global Positioning Systems.
CHAPTER SIX: EFFECTING INDUCTIVE STATISTICAL ANALYSIS ON GEOGRAPHIC INFORMATION SYSTEM DATABASES AND ANALYTICAL MISTAKES MADE IN THE E-RMS SOFTWARE MANUAL.

Outline of Chapter.

Changing A Casual Observation Into Useful Information. 77
Recording Attributes About Information Records Within An E-RMS Database. 79
Classical Set Theory. 79
Evidence From Multiple Observations. 80
Randomness, Independence And Experiments. 81
Mistakes In Procedures Of Statistical Analysis Adopted In The E-RMS Software Manual. 82
Problems Arising From Tutorial Exercise Four: If a “species occurred at random in relation to elevation.” 82
Incorrect Statistical Analysis Procedures Adopted In The E-RMS Predictive Modelling Module. 84
How The E-RMS Statistical Analysis Model Operates. 85
Incorrect Use Of Correlation Analysis In The E-RMS Software Manual. 87
Introducing “Absolute Probabilities Of Occurrence”. 88
An Incorrect Application Of Bayes’ Theorem Based On An “Absolute Probability” Or “Relative Likelihood” Of Occurrence. 89
Applying Bayes’ Theorem To A Process That Is Stochastically Closed. 90
Taking A Random Selection Of Elements From The Set Of Location Elements. 90
Applying Bayes Theorem: Assuming The Mean Of A Population Approaches That Of A Sample. 94
Derived Inferences from Overlaying Maps. 95
E-RMS Determining Areas For Which Fire Potential Is High. 95
Accuracy When Overlaying Characteristic Regions On Thematic Maps. 95
E-RMS Using User Defined Categories Like Any Other Variable. 99

CHAPTER SEVEN: E-RMS USES A DEFECTIVE LAND SURFACE RECONSTRUCTION ALGORITHM FOR MODELLING FIRE INTENSITY.

Outline of Chapter.

Information - The Ability To Discriminate Between Alternatives . 101
Measurement And Uncertainty. 103
Choosing A Numerical Estimation Algorithm. 104

The E-RMS Land Surface Reconstruction Algorithm.

E-RMS Specifying The Slope And Aspect Of A Grid Cell. 107
Angles That Can Not Be Measured Coded In The E-RMS Demonstration Database. 108
Interpolating Between Contours. 110
Failure To Detect A Single Cliff In 520,000 Hectares Of Mountainous Territory. 111
Contradictory Results Produced By The E-RMS Land Surface Reconstruction Algorithm. 114
Incompatible Slope Elevation and aspect Data. 114

SYMMETRIC TESTS OF THE E-RMS LAND SURFACE RECONSTRUCTION ALGORITHM.

Procedures Followed To Carry Out The Tests. 118
Results Of Symmetric Tests On The E-RMS Land Surface Reconstruction Algorithm. 121
Test 1. 122
Test 2: 126
Test 3. 131
Test 4: 137
General Analysis Of The Reconstruction Test. 143
An Adequate Notice Of Data Quality In This Case Would Have To Warn A User That The Data Was Misleading And Wrong. 144
**ABSTRACT**

The Environmental Resource Mapping System (E-RMS) is a geographic information system (GIS) that is used by the New South Wales National Parks and Wildlife Service to assist in management of national parks. The package is available commercially from the Service and is used by other government departments for environmental management. E-RMS has also been present in Australian Universities and used for "academic work" for a number of years.

This thesis demonstrates that existing procedures for product quality and performance have not been followed in the production of the package and that the package and therefore much of the work undertaken with the package is fundamentally flawed. The E-RMS software contains and produces a number of serious mistakes. Problems identified in this thesis include:

- The package contains a defective surface reconstruction algorithm that combined with inadequate information, amongst other things, leaves enormous cliffs undetected, codes square kilometres of flat land as having aspect and exhibits substantial directional instability;
- The E-RMS Apsley Macleay database calculates fire intensity using a land surface reconstruction algorithm that can not discriminate a cliff of approximately 300 metres and adds vector quantities as scalars;
- The AMG Easting and Northing coordinates stated for E-RMS databases do not exist;
- The methods of "statistical analysis" explained in the software manual and demonstration tutorial exercises are consistently wrong; and
- Incorrect coding exists over square kilometres of the E-RMS Aspley Macleay George database prepared from maps scaled at 1:25,000.

As a result of these shortcomings I recommend that an inquiry be conducted to investigate:

- The technical feasibility of each project for which the E-RMS package has been used;
- The full extent and consequences of the failings inherent within the package; and
- The suitability of the E-RMS GIS package for the purposes for which it is sold.

Australian Standard 3898 requires that the purpose, functions and limitations of consumer software shall be described. To comply with this standard users of the package would have to be informed that:

- E-RMS databases are defined on coordinates that do not exist;
- The reconstruction algorithm produces mistakes of a serious and not fully known magnitude;
The "statistical analysis" procedures adopted for use in the E-RMS package are based on assumptions that have been shown to be false; and

It is impossible to write a general algorithm for analysing interactions between physical properties recorded in the database and biological organisms. Failure to consider the usefulness and extractable nature of information in any GIS database will inevitably lead to problems that may endanger the phenomena that the GIS is designed to protect.
Chapter One: Introduction

- Preamble
- Ecosystems Resource Management, And Geographic Information Systems
- GIS Planning & Operation
- Thesis Outline

Preamble

Geographic information systems are self-contained computer based systems whose purpose is to collect, store, manipulate, and present geographic data. In the Macquarie Dictionary (1991) geo is defined as a word element meaning "the earth" and graphy is a combining form denoting some processor form of drawing, representing, writing, recording or describing. Consequently, Geography is the name given to the study of the distribution and interactions between phenomenon distributed on the surface of the earth.

Substantial resources have been devoted to developing computer based data systems, because it is believed that these data systems will lead to greater understanding of geographical systems and the possibility of real time mathematical modeling of real or hypothetical scenarios. The specific requirements of individual map users vary greatly; however, apart from navigational purposes, maps are frequently used to compare and define geographic distributions for objects of interest.

Geographic information systems are designed so that an operator can overlay and analyse digitally reproduced maps containing attribute data for a region of interest. This is undertaken in an attempt to promote visual assessment of the distribution of thematic properties and to assist in management decisions or answer questions of scientific and/or social enquiry. It is unreasonable to draw inferences from the displays created by computer software packages without first having a rigorous understanding of the errors inherent in the system and the importance of precision and accuracy to the outcome of the questions for which answers are sought.

Ecosystems, Resource Management, And Geographic Information Systems.

The key idea behind the concept of an ecosystem is to study a manageable small biological unit. For convenience, an ecosystem is usually regarded as being isolated, invariably this is not strictly true. Its study is complicated because all the factors within an ecosystem interact and change one another.

By the very nature of a GIS it is the geographic element that is requisite, it is the locational element that distinguishes a GIS from other types of information systems. Clearly a GIS is not appropriate for studying very small ecosystems where the system is more or less well defined without reference to distant spatial phenomenon.

An important component of ecosystem management is the prediction of the effects of external influences (often human) on biological organisms. Geographic information systems can describe and sort physical parameters such as soil type, elevation,
rainfall, and ground slope very quickly and attempts have been made to model and describe the effects of various phenomena on biological organisms.

A survey of the Proceedings of the Australasian Urban and Regional information Systems Association for 1991 and 1992 (AURISA 91 and AURISA92) reveals that GIS are currently being used extensively for resource management in Australia and New Zealand. All Australian States are now building geographic information systems for forest management, water resources, agriculture, soil conservation, mineral resources, nature conservation, and national parks. State Government coordinating bodies have been developing jurisdiction wide GIS systems for specific applications. These applications are usually based on environmental and natural resources and natural resources data has been primarily the responsibility of individual government agencies (O'Callaghan and Garner; 1991).

Geographic Information Systems Planning And Operation.

Since GIS systems have "developed" to the stage where easy to use software packages can be bought for under $2000 and run on small personal computers, there has been a rush of interest and investment in such systems by operators with little or no experience in map making or mathematics. Some people have expressed concerns that the rapid growth of GIS technology has led to systems being developed and operated by individuals and organisations with insufficient technical expertise to extract reliable information.

Burrough (1987) claimed that: "Few users seem to realise the implications of error propagation, however, believing that the quality of the results of a GIS analysis are determined by the cartographic quality of the end product."

Aangeenbrug (1991) states that: "The rapid introduction of GIS has created a booster type atmosphere where critics have been scarce. The size of investments made and the urge to declare success before delivery have not been sufficiently tempered by a willingness to admit that GIS is a young, somewhat poorly defined and difficult technology"

Processing systems have always had a logic content, however, the advent of computers has afforded the opportunity to incorporate very large, complicated and often subtle logic structures. Owing to the increase in scale of the logic systems now possible there is a greater chance of human error at all stages of the life cycle (AS 3960; 1990).

It is not yet clear whether GIS systems can deliver reliable and useful information for the study of regional environmental problems. In order to address these issues it is essential to consider the structure of information as a tangible entity, the nature and purpose of measurement, the properties and capabilities of maps used to construct GIS systems, the amenability of GIS data to mathematical analysis, and issues associated with product quality. I have investigated the fundamental mathematical structure required for conducting measurements and deriving inferences from locationally linked databases, and reviewed and tested the New South Wales National Parks and Wildlife Services' Environmental Resource Mapping System (E-RMS). E-RMS is a geographic information systems produced by the National Parks and Wildlife Service to assist in managing the State's National Parks. The package is available as a commercial product from the Service and was purchased by the School of Science at
the University of Western Sydney Hawkesbury (UWSH) for use as an Environmental
modeling and planning tool.

**Thesis Outline**

This thesis is divided into eight chapters.

- **Chapter 1** (this chapter) contains very general background material on
geographic information systems and indicates that there have been a
substantial number of mistakes made in the construction and use of the E-RMS
geographic information system.

- **Chapter 2** establishes that it is not possible to write a general algorithm for
analysing interactions between biological organisms and properties recorded at
various locations described by a locationally linked database; the problem can
not be solved, even approximately. The fact that it is not possible to write a
general algorithm for analysing interactions between biological organisms and
coded properties is very important because it follows immediately that all
current attempts to write such an algorithm are invalid. Concurrently with the
growing use of GIS technology for environmental management there have
been many attempts to write such algorithms. The E-RMS package contains
one such flawed attempt.

- **Chapter 3** contains general background information about E-RMS. E-RMS is
a raster based GIS that is divided into distinct modules and contains a
demonstration tutorial database covering the Apsley Macleay Range.

- **Chapter 4** establishes that the "AMG" easting and northing coordinates stated
for the E-RMS tutorial database do not exist and that it is not possible to enter
correct AMG coordinates into E-RMS databases.

- **Chapter 5** establishes that the stated precision for the E-RMS tutorial
database (pixels of size 100 m * 100 m) is beyond the limits imposed by
attainable accuracy. That is to say, an accuracy of 100 m * 100 m for reported
location can not, using available resources, be expected to be achieved in
practise.

- **Chapter 6** is concerned with effecting mathematical analysis on GIS
databases and analytical mistakes made in the E-RMS software manual.
Amongst other things, it is established that E-RMS can not be used on its own
as a stand alone GIS, the procedures of statistical analysis adopted in the E-
RMS software manual are consistently wrong, correlation analysis is used
incorrectly, Bayes' Rule is applied incorrectly using entities described as being
"absolute probabilities of occurrence" and "relative likelihoods of occurrence",
the application of Bayes Rule described requires the false assumption that the
mean of a population approaches that of a sample, and that attention needs to
be given to accuracy when overlaying characteristic regions on thematic maps.

- **Chapter 7** is concerned with the E-RMS land surface reconstruction algorithm
and fire intensity modeling applications. Chapter 7 contains slope, aspect and
elevation pictures that were screen captured directly from screen images
produced by the E-RMS tutorial database. These figures, and others,
demonstrate square kilometres of flat land coded as having an aspect, incorrect slope and/or elevation coding along half-kilometre fronts, and that the land surface reconstruction algorithm is directionally unstable. Specifically, the fire intensity model described in the E-RMS software manual uses categories of slope that the reconstruction algorithm can not distinguish between, relies on a reconstruction that will not detect a cliff of 274 metres, adds vector quantities as scalars and will code a rise of 100 metres over a horizontally projected distance of 100 metres as having a "slope" of 3 to 6 degrees.

- **Chapter 8** contains a summary of facts established in this thesis and presents a short number of recommendations for actions to address many of the problems in the software that underpins some of the states most important environmental databases.
Chapter Two: Deriving Inferences From Databases That Define A Partition and Contain Sinks Of Information

- Outline of Chapter
- The Structure Of Locationally Linked Databases
- Establishing A Mathematical Representation
- Determining Interactions Between Biological Organisms & Coded Properties - Insufficient Conditions
- Establishing Prediction Using Data From A Locationally Linked Database
- Supplement to Chapter 2: The Meaning Of Some Terms Used In Chapter 2

Outline of Chapter

This chapter establishes that it is not possible to write a general algorithm for analysing interactions between biological organisms and properties recorded at various locations described by a locationally linked database; the problem can not be solved, even approximately. In order to establish this fact this chapter demonstrates that any attempt to write such an algorithm meets with intractable problems. This result has substantial implications for analysing pattern in such spatial data. The arguments presented are comprehensive, exhaustive, and at times abstract. A summary of the core arguments of this chapter is given below.

- An algorithm sufficient to permit the construction of a numerical representation for analysing interactions between biological organisms and particular properties recorded at various locations described by a locationally linked database need not exist because:
  - any attempt to use the coded properties (p_i for i=1 to N) to define an equivalence relation on any locationally linked database defines a partition of the database;
  - a consistent description of location elements need conform only to a nominal scale of measurement;
  - to take a representative sample it is necessary to invoke a measure of central tendency, measures of central tendency need not exist for members of sets conforming only to a nominal scale, and therefore it may not be possible to take a representative sample of location elements from the locationally linked database;
  - there can be a large number of fatal properties;
  - the distribution and identity of fatal properties is frequently unknown,
  - different organisms have different non-linear (catastrophic) responses to certain properties;
some organisms require properties that are fatal for other organisms;

the property \( p_i \) may be fatal for a particular organism and the absence of the property \( p_i \) at a location element may imply the absence of other properties;

it is not generally possible to use a stochastic process operating over the set of location elements to test hypotheses regarding the effect of any given property \( (p_i) \) on an organism;

the property specification required to select any particular subset of location elements need not be unique; and

the inclusion of an OR statement in a selection procedure does not guarantee that a particular location element can be reasonably assigned to a class defined by a combination of the specified properties for comparison with the selected set.

Since it is not generally possible to use a stochastic process operating over the set of location elements to test a hypotheses regarding the effect of property \( p_i \) on an organism, one has to know how the particular organism will respond to coded properties prior to writing the algorithm for determining interactions between the organism and coded properties.

The distribution of organisms is frequently limited by conditions that are sub optimal rather than lethal. A transplant experiment may be necessary to account for historical factors responsible for determining the distribution of an organism. The Supplement to this chapter explains the meaning of some technical terms used in this chapter.

**Cardinality: Stating Facts In Numerical Terms.**

To formulate information in a scientific manner it is essential to have basic facts stated in numerical terms. However, it is not necessary to enumerate each unit in the universe in order to arrive at an acceptable estimate for the total. A carefully designed sample may provide the necessary information (Raj; 1968). It is possible to enumerate only those characteristics of entities for which there exists a correspondence between the characteristic and the concept of Cardinality. It is reasonable to write an algorithm for analysing interactions between entities distributed at various locations described by a locationally linked database and physical properties recorded within the database only when it is known exactly what can and what can not be counted, or represented by numbers, and how the said entities can be analysed. In many environmental management applications distributed entities of interest are biological organisms and the recorded properties describe physical characteristics present at the section of the environment described by each location element.

**The Structure Of Locationally Linked Databases**

**All Databases That Specify Location Consist Of A Finite Collection Of Location Elements.**

According to the Collins Dictionary of Mathematics (1989) the concept of space embraces a set of objects to which may be attached associated attributes, together with relation(s) defined on that set. A straight line defined between pairs of objects is one such relation.
A spatial examination of an entity must address itself to either analysing a variable that assumes values at different locations and/or analysing properties assigned to a particular spatial location. This entails a statement of location with respect to an imposed coordinate system, physical properties and topological characteristics at the stated location. Topological characteristics are significant because they address the spatial nature of geographic descriptions and specifically deal with issues associated with connectedness of points and areas.

The number of elements any finite adding machine can address is finite, therefore, the method of digital representation must be finite. A finite set of location elements, denoted by $\oplus$, can be constructed to represent all distinct location elements addressable to a locationally linked database. Without loss of generality the set $\oplus$, of location elements, can be said to have $M^2$ members.

**Using A Map To Represent The World With A Finite Collection Of Location Elements.**

In order to represent the real world using the finite set of location elements it is necessary to propose a map, an ordered pair, made up of a transformation with a given co-domain, that takes a location in the real world (W) and maps the location to a particular location element addressable to the database. To effect reasonable analysis such a map would have to be well defined.

<table>
<thead>
<tr>
<th>World</th>
<th>$f:W \rightarrow \oplus$</th>
<th>Database</th>
</tr>
</thead>
</table>

Figure 2.1: The map $f: W \rightarrow \oplus$ takes locations in the real world and maps them to particular location elements addressable to the database.

Conclusions derived about the real world W, using the representation $\oplus$, can be extended to the real world only if the transformation $f:W \rightarrow \oplus$ is invertible, at least approximately. The analysis that follows is independent of whether or not such a map exists and therefore establishes that even if it is possible to construct an $f:W \rightarrow \oplus$ that is exactly invertible it is still not possible to write a general algorithm for analysing interactions between biological organisms and properties recorded at various locations described by a locationally linked database.

**Establishing A Mathematical Representation**

**Scales Of Measurement.**

Scales of measurement based on elements of the Real Number Line are possible only because there exists an isomorphism between what can be done with measurable properties of objects and what can be done with numbers (Stevens; 1946). When measuring characteristics of objects, experimental operations are performed for classifying (determining equality), for rank-ordering, and for determining when differences and when ratios between the aspects of objects are equal. The empirical operations performed and the characteristics of the property being measured determine the type of measuring scale attained (Stevens; 1946).
Table 2.1 describes the group structure and permissible statistics for each of 4 scales of measurement that can be erected when using a representation based on elements of the Real Number Line. The mathematical group structure of a scale is the determined by the collection of algebraic functions which leave the scale form invariant. For a statistic to have any meaning when applying any particular scale the statistic must be invariant under all the transformations permissible for the scales listed mathematical group structure.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Basic Empirical operations</th>
<th>Mathematical Group Structure</th>
<th>Permissible Statistics (invariantive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Determination of Equality</td>
<td>Permutation Group: $x(i) = f(x)$</td>
<td>• Number of cases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f(x)$ means any one-to-one substitution.</td>
<td>• Mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Contingency correlation</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Determination of greater or less</td>
<td>Isotonic group: $x(i) = f(x)$</td>
<td>• Median Percentiles</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f(x)$ means any monotonic increasing function.</td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>Determination of equality of intervals or differences</td>
<td>General Linear Group $x(i) = ax + b$</td>
<td>• Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Rank-order correlation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Product-moment correlation</td>
</tr>
<tr>
<td>Ratio</td>
<td>Determination of equality of ratios</td>
<td>Similarity Group $x(i) = ax$</td>
<td>• Coefficient of variation</td>
</tr>
</tbody>
</table>

Table 2.1: The scales of measurement (Birkhoff in Stevens; 1946).

The permutation group includes the isotonic, general linear, and similarity groups as subgroups. Therefore, any statistic applicable when using the nominal scale is automatically applicable when using an ordinal, interval or ratio scale. Similarly, any statistic applicable when using an ordinal scale can be applied when using the interval or ratio scale and any statistic applicable when using the interval scale is applicable when using the ratio scale.

In order to take ratios of measurable characteristics of objects in any self-consistent manner it is essential that the ratio scale be used. This scale requires an absolute zero. Measurement of temperature in degrees Celsius serves as a specific example. Given a temperature $x \degree C$ it is not reasonable to consider a temperature $2x \degree C$; for if $x$ be 10
0°C then 2\(\times x\) corresponds to 20 0°C, but if \(x\) be -10 0°C then 2\(\times x\) 0°C corresponds to a temperature of -20 0°C! The problem arises since measurements of temperature in degrees Celsius conform only to the interval scale.

**Describing Location Elements Using Members Of Finite Sets.**

**P** - The Set Of Measurable Properties For Each Location Element.

Every number is represented by a symbol and therefore to describe the distribution of phenomena that can be represented by numbers, at any location \(L_i\), a member of \(\Theta\), it is necessary and sufficient to ascertain the presence or absence of an associated symbol. Properties that can be represented by numerals, at any \(L_i\), are determined to be either present or absent. Properties coded as present or absent at each location element must be drawn from a finite set of properties, this set can be denoted by the symbol \(P\). Without loss of generality \(P\) can be said to have \(N^3\) members. Each member of the set \(P\) can be uniquely identified using 1 of the symbols \(p_i\), where \(i\) takes all integral values between 1 and \(N\).

Generally it will not be possible to impose an ordering on elements of the set \(P\). This can be seen by considering an example for which the properties \(p_1\), \(p_2\) and \(p_3\) represent a value for a measurable property of a plant, an elevation value and a particular class of land aspect respectively. It is possible to define a property of magnitude only on those sets endowed with an ordering. In the case where it is not possible to define a property of magnitude it follows immediately that it is not possible to define a metric.

**Ω** - The Set Of All Possible Distinct Property Codings That Can Be Used To Describe Location Elements.

The set of \(N\) distinct properties can be used to recursively generate a set of all possible distinct property codings that can be used to describe location elements. This set can be denoted by the symbol \(Ω\). In order to generate the set \(Ω\), a rule for generating \(Ω\) must be established. This can be achieved using the characteristic function \(g: P \rightarrow Ω\), where \(g\) maps property \(p_i\) to the \(i\)th member of an \(N\) dimensional tuple and assigns to the \(i\)th member of the tuple the value 0 to denote the presence of property \(p_i\) and 1 to denote the absence of property \(p_i\). Each member of the set \(Ω\) is a word of length \(N\). It is clear that \(Ω\) contains \(2^N\) members. Each member of \(Ω\) can be uniquely represented by one of the first \(2^N\) binary numbers and/or one of symbols \(w_i\), where \(i\) takes all integral values between 1 and \(2^N\).

<table>
<thead>
<tr>
<th>(P): The Set of all distinct properties</th>
<th>(g : P \rightarrow Ω)</th>
<th>(Ω): The set of all possible distinct property codings.</th>
</tr>
</thead>
</table>

Figure 2.2: The characteristic function \(g: P \rightarrow Ω\) that constructs \(Ω\), the finite set of all distinct property codings, from \(P\), the finite set of available properties.

The property coding for each member of \(\Theta\) is described uniquely by one of the members of \(Ω\). The map \(ψ: Θ \rightarrow Ω\), that assigns to each member of \(Θ\) the member of \(Ω\) that describes the distribution of properties at the member of \(Θ\) is an injective function from \(Θ\) to \(Ω\). In general the map \(ψ: Θ \rightarrow Ω\) will not be surjective and it is to be expected that many members of the set \(Ω\) will not have any members of the set.
mapped to them by $\psi: \Theta \rightarrow \Omega$.

<table>
<thead>
<tr>
<th>$\Theta$</th>
<th>The set of all distinct location elements</th>
<th>$\psi: \Theta \rightarrow \Omega$</th>
<th>$\Omega$: The set of all possible distinct property codings.</th>
</tr>
</thead>
</table>

Figure 2.3: The injective function $\psi: \Theta \rightarrow \Omega$ that maps every member of the finite set of location elements to a member of $\Omega$, the set of all distinct property codings.

**Partitioning Sets With Equivalence Relations.**

If $\sim$ is an equivalence relation on any set $\mathcal{Z}$ then the family of all equivalence classes is a partition of $\mathcal{Z}$. Additionally, if $\{S_\alpha\}$ is a partition of $\mathcal{Z}$ then there exists an equivalence relation on $\mathcal{Z}$ whose equivalence classes are the $\{S_\alpha\}$ (Rotman; 1965). Because the mapping $\psi: \Theta \rightarrow \Omega$ is an injective function from $\Theta$ to $\Omega$, any partition of $\Omega$ can be used to partition the set of all location elements, $\Theta$. The properties $p_i$, for $i=1$ to $N$, can be used to construct a partition the set $\Omega$ of all possible words of length $N$.

**Measuring The Environment At Each Member Of The Set $\Theta$.**

When measuring with the nominal scale it is possible to identify either a most numerous class or a number of most numerous classes. The group structure permits that subject to certain conditions, a contingency correlation can be defined which allows hypotheses regarding the distribution of cases among the classes to be tested. In order to do this it is necessary to identify exactly which entities are to be considered as classes and exactly which entities are to be considered as cases.

**Constructing Cases And Classes For Investigating The Distribution Of Organisms With Respect To Coded Properties.**

$\Omega$ is the set of all possible distinct property codings that can be assigned to location elements, members of the set $\Theta$. $\Omega$ consists of words of length $N$ letters. To investigate interactions between biological organisms and coded properties it is necessary to split the set of $N$ letters into $2$ sets of letters:

1. Letters that describe measurable properties of the biological organism (at any given location) about which information concerning interactions between it and other coded properties is sought. The organism of interest can be called organism $Z$.

2. Letters that describe other measurable features for each location element.

Without loss of generality, it can be claimed that $X^4$ of the $N$ recorded properties are measurable properties describing organism $Z$ and that $N-X$ of the properties are measurable properties describing entities other than organism $Z$.

In order to prove that no general algorithm can be written for analysing interactions between the measurable properties of biological organisms and other coded properties it is necessary to demonstrate only that intractable problems arise when the simplest possible measuring scale is applied to the organism. The simplest measurement that can be made for a biological organism can be achieved by applying the nominal scale using $2$ classes. Class $0$ organism coded present and class $1$ organism coded absent.
could be used. This approach has the added advantage of keeping the argument and notation simple.

β And α The Sets Of Classes And Cases Respectively.

Using such an approach, the following two sets can be generated:

1. The set α consisting of a 1 dimensional tuple in which a 0 is placed to denote the presence of organism Z and a 1 to denote the absence of organism Z at a specified location element. The set α has two members.

2. The set β consisting of all distinct property codings using the remaining N-1 coded properties. The set β can be generated in an analogous way to the set Ω. The set β has $2^{N/2}$ members.

Two additional mappings can now be proposed:

1. The map $♣: ⊕→α$

2. The map $♦: ⊕→β$.

Both of the mappings $♣: ⊕→α$ and $♦: ⊕→β$ are injective functions from $⊕$ to $α$ and $⊕$ to $β$ respectively. Therefore any partition of $α$ or $β$ can be used to partition $⊕$

| $⊕$: The set of all distinct location elements | $♣: ⊕→α$ | $α$: The set of all possible cases. |
| $♦: ⊕→β$ | $β$: The set of all possible classes. |

The two members of $α$ would be referred to as the cases: case 1, organism Z present and case 2 organism Z absent. The number of classes is the number of members of $β$ constituting the range of $♦:⊕→β$; frequently the number of classes will be completely described by a proper subset of $β$. The most numerous class, or most numerous classes, are defined by the member or members of the set $β$ that have at least as many members of $⊕$ mapped to them by the injective function $♦:⊕→β$ as any other single member of $β$. The group structure permits that subject to certain conditions, a contingency correlation can be defined which allows hypotheses regarding the distribution of cases (members of the set $α$) among the classes (a subset of members of the set $β$) to be tested. For example, it may be possible to test the hypothesis that organism Z is distributed at random and therefore the distribution of members of the set $α$ is independent of any, and all, the classes defined by elements of the set $β$. It may be possible to test such an hypothesis by comparing the assignment of members of $⊕$ to elements of $α$ and the assignment of members of $⊕$ to members of $β$, with the expected assignment if elements of $α$ were assigned using a random walk over the set $⊕$.

Members Of The Sets $Ω$ And $β$ Need Conform Only To A Nominal Scale.
The ordinal scale is applied to place objects in a rank ordering scheme. Interval scales are defined by imposing the constraint of linearity on an ordinal scale. Ratio scales can be defined when it is possible to define the four relations of equality, rank-order, equality of intervals and equality of ratios (Stevens; 1946). It is known that members of the set P need not conform to an ordinal scale and therefore there is no requirement that members of either of the sets Ω or β conform to an ordinal scale. Any statistic requiring an ordinal, interval or ratio scale need not be applicable to elements of the sets Ω or β. Because elements of the set Ω or β need be measurable only with the nominal scale it will not generally be possible to define a property of magnitude on elements of the sets Ω or β.

**Taking A Representative Sample Of Elements From The Set Θ.**

Generally it will not be possible to take the average or standard deviation of elements of the sets Ω or β. In order to take a "representative sample" it is necessary to invoke some measure of central tendency. Importantly, the average and standard deviation measures of central tendency can not generally be applied to elements of Θ. In the case where it is not possible to establish a measure of central tendency it is impossible to take a "representative sample".

**Decomposing The Set β To Establish The Effect Of Coded Properties On A Biological Organism.**

In order to establish the effect (potential or actual) of properties p_i (for i=1 to N-1), on a biological organism Z, at a location described by a member of Θ, it is necessary to establish an operation for decomposing the member of β that describes the location element to its component letters. Essentially, in order to determine interactions (potential or actual) between the biological organism Z and particular properties it is necessary to factor elements of β in some reasonable manner.

Any attempt at such a decomposition must consider the issue of what the Property codes actually represent and how they are likely to impact on the biological organism Z. The properties describe physical characteristics of the environment in which organisms live. The concept of an organisms niche is well established and arises from the elementary observation that organisms are not distributed at random throughout the worlds ecosystems. Each of the properties mentioned can be thought of as a dosage. Weber's general quantitative assertion of psychophysics holds that there exists a constant C such that two stimuli S and S+ΔS will be physiologically indistinguishable unless the ratio ΔS/S>C. C is frequently approximately equal to 1/30 (Resnikoff; 1989). Organisms respond catastrophically to dosages associated with any selective physical phenomenon. This observation forms the basis of the LD50; or logdose 50 statistic. The presence or absence of some property codes (represented by letters) may be immediately fatal; regardless of how suitable the rest of the character string is. Due to the catastrophic response of organisms to some properties it may be impossible to decompose a word to determine the individual effects of coded properties on an organism.

**Sinks Of Information And Stochastic Processes.**
Once a fatal property, for a particular organism, is present at a given location, it is impossible to say anything about the potential or actual effect of other recorded properties on the organism. A fatal property acts as a sink of information.

It is apparent that even if a property \( p_i \) has no effect on organism \( Z \), organism \( Z \) will not generally be distributed independently of location elements containing property \( p_i \). Therefore, it is not generally possible to use a stochastic process operating over the set \( \Theta \) to test a hypotheses regarding the effect of property \( p_i \) on organism \( Z \).

The absence of any adequate decomposition, for members of \( \beta \), means that it may be impossible to construct a numeric representation to investigate the interaction between all properties \( p_i \) (for \( i=1 \) to \( N \)) and the biological organism \( Z \) distributed at locations described by members of \( \Theta \).

**Breaking Down The Set \( \Theta \) To Establish A Mathematical Representation To Determine Interactions Between Biological Organisms And Coded Properties.**

In order to construct a numeric representation directed at establishing potential or actual relationships between the organism \( Z \) at various elements of the set \( \Theta \) and particular properties \( p_i \), it is necessary to apply a filtration to elements of the set \( \Theta \) and exclude from consideration any elements of the set \( \Theta \) at which a fatal attribute, for the particular organism, is present (whether coded or not). If the location of all fatal properties was known then it would definitely be possible to apply such a filtration to elements of the set \( \Theta \). However, in the case where the property \( p_i \) is fatal for a particular organism and the absence of the property \( p_i \) implies the absence of property \( p_k \) at all location elements, it is impossible, from the set \( \Theta \), to say anything about the interaction between Property \( p_k \) and the particular organism. Given this problem, the fact that any attempt to establish an equivalence relation on members of the sets \( \Omega \) or \( \beta \) defines a partition of these sets, and the observation that measures of central tendency need not exist for either \( \beta \) or \( \Omega \) it may be impossible to take a statistically valid representative sample of location elements from the set \( \Theta \). The distribution of fatal properties at both the sampled and unsampled sites must be accounted for.

**There May Exist Distinct Partitions Of \( \Omega \) And \( \beta \) That Can Be Used To Aggregate Members Of \( \Theta \) Identically.**

It is observed that the mappings \( \psi: \Theta \rightarrow \Omega \) need not be bijective and therefore there may exist distinct partitions of \( \Omega \) and \( \beta \) that can be used to aggregate members of \( \Theta \) identically. In the case where a subset of location elements mapped by \( \psi: \Theta \rightarrow \Omega \) to elements of \( \Omega \) having Properties \( p_1 \) AND \( p_2 \) AND \( p_3 \) coded as present is sought a partition of \( \Omega \) can be defined by the equivalence relation that relates elements of \( \Omega \) having identical coding for properties \( p_1 \), \( p_2 \), and \( p_3 \). In this case the partition consists of a family of eight equivalence classes. These eight equivalence classes can be denoted by the symbols \( e_1, e_2, e_3, ..., e_8 \). A distinct partition of \( \Omega \) could be obtained by the family of 8 equivalence classes having identical property coding for \( p_4 \), \( p_5 \) and \( p_6 \). These eight equivalence classes can be denoted by the symbols \( f_1, f_2, f_3, ..., f_8 \). In such a case each member of \( \Theta \) can be assigned uniquely to both one \( e \) and one \( f \) class. Since \( \psi: \Theta \rightarrow \Omega \) is only an injective function there may exist a 1:1 correspondence between the members of \( \Theta \) assigned to each \( e \) class and the members of \( \Theta \) assigned to each \( f \) class.
The Property Specification Required To Select A Particular Subset Of Location Elements Need Not Be Unique.

An obvious corollary to the fact there may exist distinct partitions of $\Omega$ that can be used to aggregate members of $\oplus$ identically is the fact that the property specification required to select a particular subset of location elements need not be unique. A statement to select a subset of location elements mapped by the injective mapping $\psi: \oplus \rightarrow \Omega$ to members of $\Omega$ having up to N specified properties coded as present always selects a well defined subset of location elements. For example, if $N=10$ the statement to select all location elements mapped by $\psi: \oplus \rightarrow \Omega$ to elements of $\Omega$ having properties $p_1$, $p_2$ and $p_3$ coded as present would always select a well defined subset of location elements. Such a subset of location elements can be denoted by the symbol $\varphi$. It would not be at all difficult to construct a database in which the subset $\varphi$ was equivalent to the set of location elements mapped by $\psi: \oplus \rightarrow \Omega$ to elements of $\Omega$ having property $p_i$ (where $i$ is any single specified integer between 1 and N) coded as present. In such circumstances there may be no reasonable way of assigning location elements not in the set $\varphi$ to classes established by a combination of the specified properties used to select $\varphi$. If the property coding required to select a particular subset of location elements is not unique then it may be impossible to analyse the database to investigate the effects (potential or actual) of coded properties on an organism.

Analysing For The Effect Of 1 Property When Specifying A Union Of Elements Of $\Omega$.

Boolean operators acting on members of $\Omega$ are not associative. For example, the statement to select all location elements mapped by $\psi: \oplus \rightarrow \Omega$ to members of $\Omega$ having properties $p_1$ AND $p_2$ OR $p_3$ coded as present is ambiguous and can be given an analytical meaning only by imposing an order of operations. The Axiom of Choice, which permits the construction of a set containing exactly one element of each of a given family of sets, requires that the sets used for the construction be disjoint. Sets defined by a union of elements of $\Omega$ are not disjoint and therefore in the case where a selected subset of location elements is determined by a union of elements of $\Omega$ it will not generally be possible to analyse for the effect of 1 specified property on an organism.

Determining Interactions Between Biological Organisms & Coded Properties - Insufficient Conditions

An algorithm sufficient to permit the construction of a numerical representation for analysing interactions between biological organisms and particular properties recorded at various locations described by a locationally linked database need not exist because:

- any attempt to use the coded properties ($p_i$ for $i=1$ to $N$) to define an equivalence relation on any locationally linked database defines a partition of the database;
- members of the set $\Omega$ and $\beta$ need be measurable only with the nominal scale of measurement;
• it may not be possible to take a representative sample of location elements from the locationally linked database;

• there can be a large number of fatal properties;

• the distribution and identity of fatal properties is frequently unknown;

• different organisms have different non-linear (catastrophic) responses to certain properties;

• some organisms require properties that are fatal for other organisms;

• the property $p_i$ may be fatal for a particular organism and the absence of the property $p_i$ at a location element may imply the absence of other properties;

• it is not generally possible to use a stochastic process operating over the set $\oplus$ to test hypotheses regarding the effect of property $p_i$ on organism $Z$;

• the injective functions $\psi: \oplus \rightarrow \Omega$ and $\diamondsuit: \oplus \rightarrow \beta$ are not generally surjective and therefore there may exist many members of the sets $\Omega$ and $\beta$ that do not have any members of $\oplus$ mapped to them;

• distinct partitions of $\Omega$ and $\beta$ may be able to aggregate members of $\oplus$ identically; and

• the inclusion of an OR statement in a selection procedure does not guarantee that a particular location element can be reasonably assigned to a class defined by a combination of the specified properties for comparison with the selected set.

It follows immediately from the observation that some organisms require properties that are fatal for other organisms that in order to write an algorithm for examining interactions (potential or actual) between a biological organism and coded properties it is necessary to have knowledge of both the organism’s physiology and the distribution of fatal properties. It is not generally possible to use a stochastic process operating over the set $\oplus$ to test hypotheses regarding the effect of property $p_i$ on organism $Z$ and therefore one has to know how the organism will respond to coded properties prior to writing the algorithm for determining interactions between the organism and coded properties.

Apart from the theoretical problems associated with attempting to apply mathematical analysis methods to such databases there are practical problems:

• It is observed that the distribution of organisms is frequently limited by conditions that are suboptimal rather than lethal (Begon, 1990). Consequently, a small change in environmental characteristics may expand or restrict the area over which an organism disperses; and

• Examining the present distribution of a species to determine factors responsible for an organism’s prosperity does not consider historical factors that may have determined the distribution of the organism. In order to account for historical changes a transplant experiment may be necessary.
Establishing Prediction Using Data From A Locationally Linked Database

The success or failure of managing data of mega- or giga-byte proportions depends on whether or not the accuracy and content of each data entry is adequate and appropriate to give consistent results for the types of question being asked. Attempts to estimate the change of an ecosystem over time can not be mounted in any reasonable manner until the relationship between the rate at which the information becomes obsolete and the scale at which the information describes the ecosystem is established.

The natural environment is susceptible to external influences and is "governed by rules" that are not rigorously understood. It is a characteristic of mathematical thought that it can have no success where generalisations can not be made (Peirce; 1895). Unfortunately, because the construction of a GIS database defines a partition, and data that has been aggregated to define a partition either can not be, or can not readily be generalised, prediction is practically impossible. In order to generalise the data it is usually necessary to establish cause and effect and this must be done statistically. In order to statistically interrogate the database for the effects of coded properties on organisms, a filtration (the existence of which can not be guaranteed) must be applied, and this inevitably means that much of the data must be rejected. Additionally, conclusions derived from filtered data may not be valid when extended over the entire database. The assumption that more data necessarily increases predictive ability does not hold under circumstances where sinks of information exist.

Supplement to Chapter 2: The Meaning Of Some Terms Used In Chapter 2

I substantially attempted to apply standard definitions; as published in the Collins Dictionary of Mathematics. To suit their purpose, the meaning of some terms defined by authors varies, therefore explanation of some terms serves to prevent confusion.

- **Bijection:** A bijection is an operation that associates two sets in such a way that each member of the codomain T, is paired with exactly one member of the domain D. A bijection is both injective and surjective and has an inverse.

- **Cardinality and Equivalent Sets:** Two sets are equivalent if an isomorphism is possible between their respective members. The cardinal number of a set is the number of elements in the set. For example, the Cardinality of the set \{dog, cat, mouse, house\} is 4. The cardinal number of a set is a characteristic that one set has in common with all equivalent sets.

- **Domain:** A domain is the set of input values for a relation or function.

- **Function:** A function f:X→Y is a relation from X to Y such that:
  1. For each x ∈ X, there exists some y ∈ Y with (x,y) ∈ f.
  2. For each x ∈ X, the above y is unique.
  The element y ∈ Y is the image of x under f, and is ordinarily written f(x). With such notation, a function f is the subset of X*Y consisting of all pairs (x,f(x)). If 2 does not hold the f is only a relation (Rotman; 1965).

- **Group:** A Group is a set on which a binary add operation is defined. A non empty set X on which the binary add operation is defined is called a group,
only if, the following 4 rules are satisfied (Schmidt; 1966):

1. The product of any two elements of X itself belongs to X.
2. The symbolic multiplication is associative; that is to say, for any three elements $A, B, C$ belonging to X, $(A*B)*C = A*(B*C)$.
3. There exists an element I such that $A*I = A$ for all members of the set X.
4. There is an element I, satisfying requirement 3 such that for each $A$ in X there exists an $X$ in X with $A*X = I$.

- **Injective Mapping**: An injective mapping is an operation that associates two sets in such a way that different members of the domain (D) are all paired with members of the codomain (T). For an injective mapping, there is no requirement all members of the codomain T be paired with members of the domain D. That is to say, the image of the mapping operation may be a proper subset of the codomain.

- **Isomorphism**: An isomorphism is a one-to-one correspondence between the elements of two or more sets that preserves the structural properties of the domain.

- **Mapping**: A mapping is defined from a domain to a codomain, whenever a rule is established by which an entity $e$, a member of the domain of the map, is associated with an image $e'$, a member of the codomain of the map.

- **Magnitude**: Magnitude is a scalar quantity and consists of a number assigned to a quantity enabling comparisons to be made on the basis of the ratio of any two such quantities.

- **Surjective Mapping**: A surjective mapping is an operation that associates two sets in such a way that every member of the codomain T, is the image of at least one member of the domain D. Hence, the range of any surjective mapping is the entire codomain T.

- **Well-Posed Problem**: A problem that is well formulated and for which, under appropriate conditions, a unique solution, can be shown to exist.

---

1. A general algorithm is an algorithm that must be able to answer any well posed question concerning interactions between all coded properties and all organisms distributed at locations described by a locationally linked database.

2. Where M is a positive integer.

3. Where N is a positive integer

4. Where X is a positive integer less than N.
Chapter Three: General Background On E-RMS

- Outline of Chapter
- E-RMS
- E-RMS Is A Raster Based Geographic Information System.
- E-RMS Is Divided Into Distinct Modules
- Tutorial Database Supplied With The E-RMS Package
- Predictive Modelling Module

Outline of Chapter

This Chapter explains that E-RMS is a raster based GIS which is divided into a Database Specification Module, a Data Import Export Module, a Map Digitising Module, a Map Display Analysis Module, and a Predictive Modelling Module. The demonstration tutorial database covering the Apsley Macleay Range is provided with 4 tutorial exercises.

E-RMS A Personal Computer Software Package Intended To Convert Map Based Environmental Data Into Information

The software manual accompanying E-RMS identifies the system as a microcomputer software package facilitating the storage, retrieval, and analysis of map based environmental data. E-RMS is designed to operate on IBM PC or PS/2 or closely compatible computers. At UWSH the system is loaded on a 486 IBM compatible computer. E-RMS is entirely menu driven and the software manual advises that minimal user training is required.

The software manual states the following:

"The capabilities of E-RMS extend far beyond the standard data storage and retrieval functions offered by most GIS products. The basic philosophy behind the development of E-RMS has been to provide an easy to use system for converting raw map based environmental data into information of direct value to environmental resource managers. E-RMS therefore offers a wide range of tools for data analysis, model building and extrapolation."

E-RMS Is A Raster Based Geographic Information System.

E-RMS is a raster based GIS and the "size" of grid cells is specified by the user. Each grid cell can have information on a wide range of attributes associated with it. Figure 3.1 is reproduced from the software manual in an attempt to communicate unambiguously the manner in which the database is structured and its intended purpose. The presence or absence of each category is recorded for each grid cell in the database.

Figure 3.1. The description of location elements within an E-RMS database. (E-RMS software manual).
**E-RMS Is Divided Into Distinct Modules**

The package is described as being divided into 4 distinct modules: a database specification module, a data import/export module, a map digitising module, and a map display analysis module. A tutorial accompanies the E-RMS package and it is possible to purchase an additional predictive modelling module.

**Database Specification Module**

Using the Database Specification Module it is possible to assign a name to a database being constructed. For example; if one were constructing a GIS for the Blue Mountains National Park then one could specify the name "Blue Mountains National Park" in the database specification module. All E-RMS databases must define a rectangular matrix of square grid cells and it is claimed that the associated grid cell matrix is normally aligned with the Australian Map Grid (AMG) or the Universal Transverse Mercator (UTM) coordinate system.

According to the E-RMS software manual it is essential to specify:

- The grid cell size and this specifies the width of each grid cell in metres;
- The number of rows and this is the number of rows along the north-south axis of the database rectangle;
- The number of columns which is the number of grid cells along the east-west axis of the database rectangle;
The easting of the origin, that is, the easting of the southwest corner of the database rectangle. This is usually stated in AGM or UTM coordinates and must be specified to metre resolution; and

The Northing of the origin in AGM or UTM coordinates, this too must be specified to metre resolution (E-RMS Software Manual).

**Data Import Export Module**

The Data Import Export Module allows grid cell data to be imported (exported) into (from) an E-RMS database from (to) foreign files of varying format.

**Map Digitising Module.**

The Map Digitising Module is used to directly digitise data into specific locations within the database. This is usually done by an operator attaching a map to a "digitising board" and activating a cursor (similar in concept to a computer mouse) to communicate to the computer that a specific item of interest is at a particular location on the map. Digitised information may be thought of as being stored in coverages as shown in Figure 3.2.

Figure 3.2: Digitised information is stored in coverages (E-RMS Software Manual).

Once contour lines have been entered into the E-RMS database, either through the map digitising module or via the data import/export module, the software can be instructed to interpolate slope, elevation, and aspect data for every pixel in the database.

**Map Display Analysis Module**

The Map Display Analysis module acts as an interface between an E-RMS database and its user. The map display analysis module is entirely menu driven and easy to use because at each step the user simply selects one of several listed alternatives. Data can
be retrieved in both map or report form. New features can be created by combining existing data categories and analysis of relationships between data can be attempted.

**Predictive Modelling Module**

The Predictive Modelling module enables decision tree models to be built. An example is given in which fire intensity is predicted throughout a database area using slope and vegetation variables present in the database. Additionally, by analysing correlations between variables the module can derive its own models via a process described as "machine learning".

**Tutorial Database Supplied With The E-RMS Package**

The tutorial database provided with the package uses a sample database covering the Apsley Macleay George System, South East of Armidale. This database is a reproduction of a more comprehensive one used by the National Parks and Wildlife Service (NSW) to manage reserves in the Apsley Macleay George System.

The following is recorded in the database specification module:

<table>
<thead>
<tr>
<th>NAME</th>
<th>APSLEY-MACLEAY GEORGE SYSTEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRID CELL SIZE (metres)</td>
<td>100</td>
</tr>
<tr>
<td>NUMBER OF COLUMNS</td>
<td>656</td>
</tr>
<tr>
<td>NUMBER OF ROWS</td>
<td>790</td>
</tr>
<tr>
<td>ORIGIN AMG EASTING (metres)</td>
<td>375000</td>
</tr>
<tr>
<td>ORIGIN AMG NORTHING (metres)</td>
<td>545000</td>
</tr>
</tbody>
</table>

**Lessons In The Tutorial Package**

The tutorial package contains 4 lessons.

**Lesson 1**

The purpose of the first lesson is to introduce the general E-RMS menu system and to generate and manipulate simple maps on the screen.

**Lesson Two**

The second lesson introduces the capabilities of the system to display more than one map layer at a time. In this lesson a method of looking at relationships between variables is introduced. The variables are specified as a base (primary) map layer and a top (secondary) map layer. A map showing the distribution of a particular plant species can be "layed over the top of" an elevation or rainfall map.

In the example given, rainforest was selected as the base layer. For the secondary map layer a new category was derived by merging all the categories that indicated a fire had occurred between 1980 and 1988. This new category was named "All Fires Since 1980." The next operation involves simply plotting the map with both the base and primary map layer.
Lesson Three.

This lesson explains how to derive new features by "Boolean overlaying" and proximity transformations. In the example given, areas of high fire potential were derived by "Boolean overlaying" and this was done by selecting pixels coded with particular classes of vegetation, slope, aspect, and fire history. The feature so created is named "HIGH FIRE POTENTIAL".

Lesson Four.

Lesson four explains how to solve a more complex "real life" problem. In this case the habitat of a rare species of tree Eucalyptus michaeliana is investigated. By analysing the environmental attributes of the sites at which the plant is known to exist in the area described by the database, an attempt is made to predict where else the species is likely to occur throughout the database area. This is done by calling up a selection of frequency histograms each containing two distributions. There are 2 separate histograms called up in this example (a total of 4 distributions) and they describe elevation and rainfall characteristics respectively.

Predictive Modelling Module

All predictive models built with the predictive modelling module are decision tree models. The decision tree model given in the software manual estimates fire intensity. A modelled variable in E-RMS resembles all other variables in that it defines a number of categories. For example, the categories of the FIRE POTENTIAL model are LOW, MODERATE, and HIGH. Once such a model has been run each pixel in the database is coded as having categories of the model present or absent. In the fire intensity model each grid cell would be coded as having a LOW, MODERATE, or
HIGH, fire potential. A modelled variable can be displayed and analysed using the map display analysis module and can "be used as a branching variable in further models" (E-RMS Software Manual).
Chapter Four: E-RMS Requires Location Coordinates That Do Not Exist.

- Outline of Chapter
- The Australian Geodetic Datum
- Coordinates Stated For The E-RMS Tutorial Database

Outline of Chapter

This chapter establishes that the "AMG Coordinates" specified for the E-RMS tutorial database can not be reconciled with the coordinates described in The Australian Map Grid Technical Manual Special Publication Number 7 and/or The Australian Geodetic Datum Special Publication Number 10. It is further established that it is not possible to correctly enter AMG Coordinates into E-RMS databases.

The Australian Geodetic Datum

In 1966, all the geodetic surveys in Australia and New Guinea were adjusted and placed on a new Australian Geodetic Datum (AGD). The new grid was named The Australian Map Grid and is frequently denoted and referred to as the AMG. At the time, the Australian map grid covered Australia and its territories, with the exception of the Australian Antarctic Territory and sub-antarctic islands. The AMG is based on the metric UTM system and coordinates on the AMG are derived from the projection of the Transverse Mercator lines of latitude and longitude onto the Australian Geodetic Datum. In 1966 the coordinates were known to be correct to less than 1 mm anywhere in a grid zone (National Mapping Council of Australia; 1972).

Grid Zones, AMG Eastings And Northings.

In the AMG each zone is 6 degrees wide and contains 1/2 degree overlaps at each intersection. These zones are numbered from 47, with central meridian 99 degrees east to zone 58 with central meridian 165 degrees east. Each zone has a true origin, defined by the intersection of its central meridian with the equator and a false origin from which AMG eastings and northings are quoted. The false origin is defined to be 500,000 metres west of the true origin and 10,000,000 metres south of the true origin (National Mapping Council of Australia; 1972).

A pair of grid coordinates on any given datum and grid may be used to uniquely define a point on the earths' surface. It is necessary to specify the Grid zone, the easting, and the northing of the location. The Australian Geodetic Datum Special Publication number 10 sites the AMG coordinates for "the Lion" survey station as being:

<table>
<thead>
<tr>
<th>Station</th>
<th>Zone</th>
<th>Eastings</th>
<th>Northings</th>
</tr>
</thead>
<tbody>
<tr>
<td>LION</td>
<td>56</td>
<td>497 345.431</td>
<td>6 852 369.368</td>
</tr>
</tbody>
</table>

Table 4.1: Coordinates for "The Lion" geodetic station.
The eastings and northings are stated in metres. This location is in New South Wales, approximately 19 kilometres north west of Kyogle. From the coordinates of "The Lion" it is apparent that it is approximately 3150 kilometres south of the equator and 6950 kilometres north of the false origin.

Each zone of the AMG is divided into 100 kilometre squares. The AMG holds the convenience of rectangular coordinates by restricting the size of grid squares.

**Coordinates Stated For The E-RMS Tutorial Database**

The E-RMS database specification module does not permit the entry of a grid zone and therefore it is impossible to enter unique AMG coordinates using the E-RMS software package.

<table>
<thead>
<tr>
<th>NAME</th>
<th>APSLEY MACLEAY GEORGE SYSTEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORIGIN AMG EASTING (metres)</td>
<td>375000</td>
</tr>
<tr>
<td>ORIGIN AMG NORTHING (metres)</td>
<td>545000</td>
</tr>
</tbody>
</table>

Table 4.2: AMG Easting and Northing coordinates stated in the E-RMS Tutorial Database

An AMG northing of 545000 metres would have to be approximately nine and a half thousand kilometres south of the equator and this is outside the range of the region in which the AMG is defined. Additionally, an AMG northing of 5,450,000 (in case a zero was left off accidentally) is approximately 1400 kilometres south of "The Lion" geodetic station and this is in the vicinity of the Tasman Sea or the Indian Ocean.

By examining the New South Wales topographic map Kangaroo Flat; 9335-1V-S and comparing the coordinates described on the map with those recorded in the tutorial database it seems most likely that the coordinates recorded on the sides of the topographic map have been mistaken for AMG Easting and Northing Coordinates. The coordinates written on the side of the map are Grid Reference Coordinates not AMG Easting and Northing Coordinates. On the map Kangaroo Flat; 9335-1V-S it is stated:

"Grid: Australian Map Grid (U.T.M.) in 60 zones. Grid lines are shown at intervals of 1000 metres from False Origin - Zone 56 Integrated Survey Grid information shown along border in grey - Zone 56/2."

"If reporting beyond 180 in any direction, prefix grid zone designations, as 56JML208526. Before giving a grid reference, civilian users should state the number and name of this map. Eg. 9335-1V-S Kangaroo Flat 208526".

When stated on their own grid reference coordinate numbers do not uniquely define a location within the region of geographical space described by the AMG.
Chapter Five: Limits To Resolution Set By Attainable Accuracy In The E-RMS Demonstration Database.

- Outline of Chapter
- Precision and Accuracy
- Accuracy Required For Survey Data
- Standards Of Map Accuracy
- The Problem Of Stating Location
- Conversion Of Analogue To Digital Data
- Resolution Stated For E-RMS Tutorial Database

Outline of Chapter

This Chapter establishes that the stated precision for the E-RMS tutorial database (pixels of size 100 m * 100 m) is beyond the limits imposed by attainable accuracy. This is achieved by considering the accuracy of survey data, The National Map Accuracy Standards, methods available for determining location and processes available for converting map/location data to digital format.

Precision and Accuracy

The precision of a record of observation is related to the extent to which the record specifies that the phenomena being observed exists within some abstract interval. Accuracy can be defined by the distribution of record errors. Resolution sets a lower bound on accuracy and it is generally considered good practise to ensure recorded precision falls within limits set by attainable accuracy.

Digital databases can specify locations and report data very precisely and this can create false illusions of data accuracy. Wolf (1993) confirmed that it is important to record all significant figures in measurement. Failing to record the last significant digit is a waste of the extra time and resources spent in obtaining the added accuracy implied by the extra digit. Conversely, when more than the number of significant figures is recorded, a misleading and false level of accuracy is implied.

In order to be able to study ecological communities it is essential to collect and store information about them. If a GIS is to be used it is of paramount importance that the limit of resolution of the entire system be well known. In short, there is no point in constructing a GIS that is designed to operate at a level of resolution greater than the accuracy at which location can, or will be, reported.

Accuracy Required For Survey Data

The accuracy required for survey data is that there be no plotable error in the survey data (Bannister and Raymond; 1984). A line can be hand drawn on a sheet of paper to within 0.25 millimetres and consequently if a survey is to be undertaken at 1 in 2000 scale all measurements must be sufficiently accurate to ensure that the relative
positions of any point with respect to any other point in the survey can be stated to an accuracy within 0.25 millimetres at survey scale, at 1:2000 this represents 50 centimetres.

The zero dimension of any conventional paper map is defined as 0.25 millimetres, however, it is not feasible to scrutinise a map at this level of detail and extract reliable information.

<table>
<thead>
<tr>
<th>Map Scale</th>
<th>1:1,000</th>
<th>1:4,000</th>
<th>1:10,000</th>
<th>1:25,000</th>
<th>1:50,000</th>
<th>1:100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum mapping unit. (metres)</td>
<td>0.25</td>
<td>1</td>
<td>2.5</td>
<td>6.25</td>
<td>12.5</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 1: Minimum mapping unit for maps of various scales.

**Standards Of Map Accuracy**

The Australian National Standards of Map Accuracy are reproduced below.

National Mapping Council of Australia.

Standards of Map Accuracy.


1. Horizontal Accuracy.

The horizontal accuracy of standard published maps shall be consistent within the criterion that not more than 10% of points tested shall be in error by more than 0.5 millimetres.

This limit of accuracy shall apply in all cases to positions of well defined points only. "Well defined" points are those that are easily visible or recoverable on the ground. In general what is "well defined" will also be determined by what is plottable on the scale of the map within 0.25 millimetres.

2. Vertical Accuracy.

Vertical accuracy, as applied to the contours of standard published maps shall be consistent with the criterion that not more than 10% of elevations tested shall be in error by more than one-half the contour interval. In checking elevations taken from the map, the apparent vertical error may be decreased by assuming a horizontal displacement with the permissible horizontal error for a map of that scale.

The following quote is recorded in the minutes of the First Meeting of the working party to deal with standard topographic mapping specifications for all scales. 5-7/9/1982 (CMA Bathurst, 1982):

"Lt Col Harrison pointed out to the delegates the importance of the map accuracy statement as a quality control device in mapping and as an aid to the map user. He stressed that its inclusion on maps should be considered a professional obligation."
The errors in maps produced in accordance with the above standard are, in general, normally distributed. It should also be noted that the error in the horizontal position of contours on the flat two dimensional map is not independent of the "lie of the land". For a fixed contour interval the horizontal error associated with contour location depends on the rate of change of land elevation in any particular direction. Consequently, it follows that contour lines are subject to greater horizontal locational errors in flat and undulating country than in steep mountainous terrain.

The Problem Of Stating Location

Because the AMG is defined so precisely, the errors associated with statements of location are relatively constant: that is to say the error inherent in stating location does not necessarily decrease simply because a given database covers a smaller area. In terms of using GIS for environmental analysis there are a few ways in which one could attempt to state the location of an observed phenomenon.

Surveying All Observation Points.

The location of all reporting sites or locations where observations are made could be surveyed in and referenced from known locations such as geodetic stations. Whilst such methods could be highly accurate (within 2 mm in any grid zone) they are extremely time consuming and expensive, and therefore of little benefit.

Estimating Location From A Paper Map.

Another alternative is to simply estimate location off a paper map. This method is somewhat problematic and subject to varying degrees of inaccuracy, dependant upon the scale of the map used to make the location statement, the characteristics of the terrain and vegetation cover in which the location exists, and the skill of the individual reading the map. It is generally not possible to state location to within 100 metres off a map scaled at 1:25,000 without going to an extraordinary amount of effort. At 1:25,000 scale 100 metres on the ground is represented by 4 mm on the map.

Global Positioning Systems.

Global Positioning Systems (GPS) are able to state location on the ground to within a few centimetres. However, this technological capability does not, at present, resolve all the inherent difficulties associated with stating location.

The term GPS is used generically to describe satellite based location systems. Both the American Defence Department and the former Soviet Union have developed such systems over a number of years. Specifically the American system is known as the Navigation Satellite Timing and Ranging Global Positioning System (Navstar GPS). The Russian system is known as Glonass; an abbreviation of Global Navigation Satellite System. There are a number of other electronic satellite navigation systems, however, Navstar and Glonass are distinguished by their aim of providing a continuously available global locational system (Ackroyd and Lorimer; 1990).

The Navstar GPS system has three operational parts: the space segment, the ground control segment, and the operator segment. In 1990 it was envisaged that the completed system would consist of 21 active satellites and 3 inactive "spare"
The satellites have an orbiting period of approximately 12 hours at a radius of approximately 20,000 kilometres from the earth (Ackroyd and Lorimer; 1990).

Each GPS satellite transmits microwave signals at two distinct frequencies and these are modulated by a code unique to each satellite. Each satellite "knows" its own position in its earth orbit. This is possible because the time evolution of the orbit path is stated in advance and because the satellite is fitted with an atomic clock. Orbit location can be calculated from known initial conditions. Corrective information can be sent to the satellites to prevent inaccuracies from accumulating.

GPS signals can be used to determine position by either a single receiver operating in "point positioning mode" or by the use of differential networks. These require multiple units set up at known locations with respect to each other. By comparing measurements from each receiver it is possible to detect a relative change of a single millimetre in horizontal position. Because it is not possible to track a satellite below the horizon, vertical deviations can be detected only when they are approximately one centimetre (Tralli; 1993).

In order to calculate 3-dimensional position it is essential to track 3 satellites with an unobstructed line of sight. Because the satellites are travelling at enormous speed (approximately 3000 metres per second), the atomic clock of a fourth satellite must be exploited to correct the timing signals. However an unobstructed line of sight can not be guaranteed at all locations; particularly in mountainous terrain and/or locations with extensive vegetative coverage.

The atmosphere contains a succession of ionised layers beginning at an elevation of around 50 to 100 kilometres and extending several hundreds of kilometres upward; these ionised layers are known collectively as the ionosphere. In the ionosphere uncharged molecules absorb high energy (short wavelength) electromagnetic radiation from the sun and become ionised and consequently the ionosphere contains a large number of free charged ions and electrons. Additional electric activity occurs throughout the entire atmospheric system. Electromagnetic activity interferes with the time required for the microwave signals transmitted from the satellites to reach the GPS instruments. Water vapour and clouds can also lead to signal delay. These effects can be corrected for and may be significant, depending on the manner in which the receiver is operated and the application it is being used for. It is generally true that the more expensive instruments have superior correction capabilities than cheaper hand held units. The price of small hand held GPS units is dropping all the time and they can now be purchased from outdoor and marine stores for under $2000. These new systems are coming onto the market so rapidly that their 95% accuracy is not well established (Personal Communication: Australian Army Survey Regiment; 1993).

The performance of GPS systems is also not independent of the satellite constellation above the GPS detector at the time of the observation. Performance can be expected to be optimal when four or more satellites are spread out in a more-or-less "rectangular" constellation. At present, particularly in the Southern Hemisphere, there are insufficient satellites in orbit to guarantee such optimal conditions can be continuously met.

For security reasons the US Department of defence interferes with the time on the satellite clocks and this introduces "errors" into the GPS signals. This practise is known as Selective Availability. These "errors" make it "impossible" for civilian
users to gauge the precise position of satellites at any given time and this means that
previous locational accuracy is no longer attainable when operating with a single
instrument. This has the effect of increasing the error associated with a locational
statement from approximately twenty metres on the ground and 30 metres vertically
to approximately 100 metres on the ground and 150 metres vertically (all stated at the
95% confidence level). It is possible to "get around" the problems imposed by
selective availability by using differential networks with one receiver placed at a
known location. Provided the two instruments are referencing off the same satellites
the associated error vector should be the same for both instruments (Tralli; 1993).

Global positioning systems can provide astonishingly accurate and generally reliable
information. However, this accuracy comes at a price and can be guaranteed only
under optimal conditions. It is possible for a $50,000 system with some
microcomputer post-processing to state latitude, longitude, and elevation to within
about 10 millimetres (Tralli; 1993). It is not realistic to describe the conditions in a
heavily wooded valley as being optimal for GPS operation and ecologists would
generally not be prepared to take the time required to get a reliable signal and use
differential techniques with very expensive GPS units to be able to report location at
the level required to reliably correspond an observation on the ground to a 100 by 100
metre "location" within the database. Additionally, in the case where GPS systems do
not report locations in AMG coordinates, coordinate conversion is required.

**Conversion Of Analogue To Digital Data**

In order to convert analogue information to digital format it is necessary to adopt
some form of manual entry (such as hand digitising) or to exploit automated scanning
technologies. According to Cimon et al (1990), competition amongst developers of
GIS has resulted in efficient, easy to use software being made available on small
computers and consequently manual digitising of maps is now the most common
method of registering geographical information on a computer system for further
analysis.

Data entered from manual digitisers can be entered in either stream or point options.
When stream entry is operating, data is entered automatically as the digitising cursor
is guided across a trajectory of interest on the map. During point entry the location is
recorded only when the operator activates the cursor; generally by depressing an
active button on the digitiser. It is generally true that the intervening section between
each entered point (whether recorded automatically or only at the operator's
instruction) is filled in by a straight ray joining the two adjacent points.

Bolstad et al (1987) conducted an experiment to determine the nature of the error
inherent in the digitising process. In the experiment polygons were digitised twice off
source documents which were scaled at 1:25000 and 1:23600. The maximum
discrepancy between the two vector representations of the polygons was then
examined. For 6 sites it was found that the maximum deviation was 20 metres (as a
representation on the ground) or less.

Digitising errors occur on top of the specified error in the map and consequently the
error can now be expected to be over twice that present in the original source
document. To ensure high levels of accuracy in digital data acquisition it is essential
to acquire the digital information directly via photogrammetric methods. By the use of
such methods it is possible to maintain the high precision of photogrammetric
methods and process at a level of 5 to 10 micrometres at the scale of the photography. At 1:25000 scale this represents a faithful reproduction of the photographic image of the ground at the 12 to 25 centimetre level (Taib and Trinder; 1991).

**Resolution Stated For E-RMS Tutorial Database**

It is observed that the E-RMS tutorial database covers an area of approximately 80 kilometres north - south by 65 kilometres east - west and is constructed with a grid cell size of 100m by 100m.

<table>
<thead>
<tr>
<th>NAME</th>
<th>APSLEY-MACLEAY GEORGE SYSTEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRID CELL SIZE (metres)</td>
<td>100</td>
</tr>
<tr>
<td>NUMBER OF COLUMNS</td>
<td>656</td>
</tr>
<tr>
<td>NUMBER OF ROWS</td>
<td>790</td>
</tr>
</tbody>
</table>

Table 5.2: Entries in E-RMS Tutorial Database (E-RMS software manual).

Any point P existing at some location on the ground in 3-dimensional space is located within the database by establishing its projection onto a horizontal XY plane. Given the north - south displacement from the top to the bottom of the database area and the east - west displacement across the database area it is apparent that the diagonal distance across the database area is approximately 102 kilometres. The claim that it is possible to locate the projection of a location into the correct box at all locations on the database requires that it be possible to state the horizontal projection of a location with a certainty of better than 1 part in 1000 along the diagonal of the database.

**Stating Location At The Limit Of Resolution Of The E-RMS Tutorial Database?**

Given the problems of stating location dealt with, the errors inherent in hand digitising data from topographic maps, the observation that the AMG coordinates stated for E-RMS databases do not exist and the fact that this software package cannot readjust maps to compensate for deviations in magnetic and grid north, an accuracy of 100 metres by 100 metres will not be achieved in practice. An total error in hand digitising, including map registration, from a paper topographic map of plus or minus 1 millimetre represents a location error of plus or minus 32.5 metres (25 + 12.5 metres; National Map Accuracy Standards). Since these errors are additive the potential uncertainty associated with locating an entity (by hand digitising) in the database, under the optimistic circumstances outlined, is 75 metres and this is over half the width of a 100 metre box.

- Outline of Chapter
- Changing A Casual Observation Into Useful Information.
- Classical Set Theory
- Evidence From Multiple Observations
- Mistakes In Procedures Of Statistical Analysis
- Deriving Inferences from Overlaying Maps

Outline of Chapter

This chapter is concerned with effecting inductive statistical mathematical analysis on geographic information system databases and analytical mistakes made in the E-RMS software manual. It is established that E-RMS has no provision for storing information about the methods used for data collection and abstraction and therefore cannot be used on its own as a GIS. Analysis from chapter 2, combined with considerations of conditional probabilities arising from set theory show that the procedures of statistical analysis adopted in the E-RMS software manual are consistently wrong, correlation analysis is used incorrectly, Bayes' Rule is applied incorrectly using entities described as being "Absolute Probabilities of Occurrence" and "Relative Likelihoods of Occurrence", the application of Bayes Rule described requires the false assumption that the mean of a population approaches that of a sample, and that attention needs to be given to accuracy when overlaying characteristic regions on thematic maps.

Changing a Casual Observation Into Useful Information

To change a casual observation into useful information or data requires the detailed reporting of at least the following attributes of the observation:

- Theme - The phenomenon or object being observed or measured must be recorded in some measurable or defined units;
- Location - The simple observation of a phenomenon without a record of the location of the phenomenon rarely generates useful information; and
- Time - A record of observation for which there is no concurrent record of time has minimal information content (Sinton; 1978).

Any observation that does not record explicitly or implicitly all of the above attributes is inadequate in its information content and therefore should not be part of a geographic information system. Data for which no record of precision or reliability
exists should be suspect. Once such data are entered into a computer it is assumed that they are accurate to the stated level of precision (Sinton; 1978).

At least three important types of data generalisation commonly take place:

- **Aggregation** - This usually involves the definition of spatial location and the counting of thematic characteristics for that location. Generated data is interval in nature and may be manipulable by mathematical or statistical techniques;
- **Classification** - In this case observations are matched and categorised with other observations exhibiting like characteristics; and
- **Induction** - This is the process by which a series of sample measurements are generalised to include a much larger group of phenomenon or locations assumed to have the same characteristics as the sample measurements (Sinton; 1978).

These procedures of abstraction and generalisation significantly affect the utility of data for analytic purposes. Certain types of detail present in the original data may be lost. It is important to establish the extent and characteristics of the detail lost in the process of generalisation as this affects the nature of the thematic content of the information (Sinton; 1978).

**Recording Attributes About Information Records Within An E-RMS Database.**

The E-RMS package has no provision, within the software, to state the origin of source documents, or to describe the procedures used to aggregate or classify the data used to construct the map. It is impossible to properly analyse data unless there is a record of the methods used in aggregating and recording the data and consequently the E-RMS package can not be used on its own as a GIS. Sinton (1978) has argued that data for which there is no attribute record should not be part of a GIS.

**Classical Set Theory**

A set can be thought of as a collection, possibly infinite, of distinct objects, that is treated as an entity in its own right, and with identity dependent only on its member elements. A well defined set is a set in which an element can be unambiguously categorised as either being in, or not in, the set. Because the recorded presence, or absence, of any property can be determined "exactly" (Chapter 2), the instruction to select a subset of location elements at which the property \( p_i \) is coded as present always selects a well defined subset of location elements.

Unfortunately the act of treating a set as an entity in its own right can, and within the context of GIS often does, mean that it is not possible to distinguish between members within the set. Complications arise because not all "Biological Sets" are well defined. Whitehead (1961) has suggested that classification is a halfway house between the immediate concreteness of the individual thing and the complete abstraction of mathematical notions. If the classification system has categories that are well defined then the variation that occurs between categories is discontinuous, that is each element is present in one category or the other.

**Evidence From Multiple Observations**
The theory of Mathematical Statistics is concerned with obtaining all and only the conclusions for which multiple observations are evidence. Mathematical statistics is not merely the handling of facts stated in numerical terms (Kaplan; 1961). Experimental work is frequently carried out under conditions that are well controlled and consequently observations can be made accurately. However, in many locationally linked databases the material is inherently variable and consequently the results have to be treated statistically.

The testing of an appropriate hypothesis relating to measurable characteristics is central to statistical decision making and consequently when applying statistical methods, it is essential to carefully and precisely define the problem to be solved. Inductive statistics is based on the mathematics of probability. By associating a property string, as outlined in previous discussion, a sorting process is being imposed on each location element. The location elements may be sorted in a variety of different ways according to the properties (letters) of interest at the time. A set structure is associated with classifications defined by the properties of interest. It would always be the case that a subset of location elements could be defined by a statement similar to: "the subset A consists of all location elements having elevation 100 metres coded as present". Additionally, a subset B may be defined as the set of all location elements having 200 millimetre rainfall coded as present.

The purpose of inductive statistics is to provide methods for making statistical inferences about a population based on a collection of sampled individuals. Because inductive statistical inferences are determined on a probability basis the inferences are probabilistic. Before any inductive statistical analysis can be conducted on a locationally linked database it is necessary to establish exactly how a random experiment could be conducted on such a database and the outcome that would be expected from a long series of trials of the random experiment.

**Randomness, Independence And Experiments**

An event occurs "at random" in the case where the circumstances required to bring about the event exist "in the hands of chance". In such cases no explicit cause exists and consequently it is not possible to directly attribute a cause or collection of contributory processes.

Events are said to be independent in the case where the occurrence of any one of the events, or occurrence of any combination of the events, has no bearing on the occurrence of any of the other events, or any combination of any of the other events. It is instructive to define an experiment as a process having the following properties:

- It is performed according to a set of rules that determines the performance completely;
- It can be repeated arbitrarily often; and
- The results of each performance depend on "chance". That is on influences that can not be controlled and therefore can not be uniquely predicted (Kreyszig; 1988).

A single performance of an experiment is referred to as a trial and the result as the outcome of the trial. It is a matter of common observation that most random experiments exhibit statistical regularity. That is to say, the relative frequency of an
event E selected from a finite sample space in a long sequence of trials approaches some constant value. Because of this convergence a number \( P(E) \) is postulated and defined as the probability of the event E from the random experiment. The statement E has probability \( P(E) \) should be understood to mean that over a long series of trials it is to be expected that the relative frequency of E will converge to \( P(E) \) (Kreyszig; 1988).

E-RMS tutorial exercise 4 demonstrates an incorrect understanding of random processes and incorrectly attempts to apply a stochastic process operating over the set of location elements to infer information about factors responsible for determining the distribution of Eucalyptus michaeliana.

**Mistakes In Procedures Of Statistical Analysis Adopted In The E-RMS Software Manual**

Problems Arising From Tutorial Exercise Four: If a "species occurred at random in relation to elevation."

In tutorial lesson 4 the environmental attributes for the rare plant *Eucalyptus michaeliana* are analysed in an attempt to predict where else the species is likely to occur throughout the area described by the database.

There are 218 plant survey sites within the area described by the database and these sites are selected as "the domain." An attempt is made to compare the attributes present at the known distribution of *Eucalyptus michaeliana* with the attributes of all of the surveyed sites. By noticing particular attributes associated with locations where *Eucalyptus michaeliana* has been found an attempt is made to "predict" other locations in the database where *Eucalyptus michaeliana* may occur.

Grey and yellow frequency histograms are constructed. The grey histogram represents the frequency distribution of elevation categories throughout the domain (the sites surveyed for rare species). The frequency distribution of the elevation at the 13 sites where *Eucalyptus michaeliana* was found is represented by a yellow histogram. It is then stated that:

"If the species occurred at random in relation to elevation you would expect the yellow histogram to match the grey histogram."

Using the word random in this way is wrong. For an event to occur at random it is essential that the event have no attributable cause. To talk about the distribution of a species being "random with respect to elevation" demonstrates an incorrect understanding of random processes. The distribution of an entity may occur at random over a domain; an entity can not correctly be said to be distributed at random with respect to some other entity.

In the case where the trees were present over the entire sampling domain the distribution of the trees would be uniform with respect to all the stated attributes. The distribution of trees is not random and the problem of non-independence arises. If it is given that a tree's health is in no way dependent on elevation then it can still not necessarily be expected that the trees will be uniformly distributed with respect to elevation. The reason for this is that whilst elevation may have nothing to do with the tree's life-cycle, rainfall may well. There is no guarantee that the rainfall will be
uniformly distributed with respect to elevation. It is not at all difficult to argue that the distribution of species is not directly influenced by elevation. For example, if it were possible to simulate every characteristic present at one elevation (say 100 metres) with that at another elevation (say 1000 metres) then there is no apparent reason why species that previously occupied 100 metre elevation sites, but not 1000 metre elevation sites, could not survive at the modified 1000 metre sites. This is in contrast to soil moisture which affects organisms directly. Given the problems outlined; a histogram such as that presented is misleading.

Ultimately to have success doing any sort of inductive statistical analysis it is essential to have some method of knowing the distribution of fatal attributes for an organism, applying an appropriate filtration and examining the family or families of equivalence classes defined by the coded properties.

**Incorrect Statistical Analysis Procedures Adopted In The E-RMS Predictive Modelling Module.**

The predictive modelling module contains a general algorithm for examining the "statistical relationships" between the distribution of properties described in the database and the distribution of particular organisms. It has already been established that there exist intractable problems associated with writing such a general algorithm (Chapter 2); the problem can not be solved, even approximately. The model attempts to investigate these "statistical relationships" by examining correlations between the distribution of properties and the distribution of an organism of interest. There is no attempt made to discuss the need for, or to apply, a filtration to the surveyed sites prior to undertaking any of this analysis. The predictive modelling module attempts to derive its own models and this is frequently done so that ground surveyed data can be extended across unsurveyed areas. The correlations used are correlations between observations at the surveyed sites and variables having complete geographical coverage. The model produced is a decision tree model relating ground surveyed data to one or more remotely mapped variables.

**How The E-RMS Statistical Analysis Model Operates**

The method by which the model operates is outlined briefly below.

Consider Figure 6.1 to represent the database. The ticks and crosses on the diagram indicate the location of all surveyed sites. A tick is used to denote the presence of the item of interest (say species X) and a cross is used to denote the absence of the item of interest (species X again). It is possible to examine the distribution of an attribute for which complete geographical coverage is known and look for a correlation between the ticks (or the crosses) and estimate whether the presence of the ticks, or the crosses, is more likely when the attribute for which complete geographical coverage is known is present or absent. For example, if 200 metre elevation were present at 10% of the sampling sites and species X was always present at all the sites having 200 metre elevation it may be inferred that species X was "statistically" likely to be present at all the other sites in the database having 200 metre elevation. On the other hand, suppose the other 90% of the surveyed sites had 500 metre elevation and species X was not present at any of them; then it could be "statistically" inferred that at all the other locations in the database having 500 metre elevation species X was not present.
Figure 6.1: Ticks denote the recorded presence of an item of interest and crosses denote its recorded absence at surveyed sites scattered throughout the area described by the database.

The model applies a stopping criterion that determines how categories of a variable combine to form branches in a model. As an example, consider the distribution of an organism in relation to the variable rainfall:

Divide the rainfall variable into 3 categories: (0-200 mm rainfall annually, 200-500 mm rainfall annually, over 500 mm rainfall annually) and suppose that each category of rainfall is present over exactly one-third of the database area. Suppose every occurrence of the organism (say 200 occurrences) was in the category of rainfall 200-500 mm, then, the setting of an appropriate stopping criterion would allow the induction algorithm to determine that the organism did not prosper at sites not having 200-500 millimetre rainfall. Having determined this, the elevation profile of the sites at which the organism was present could be examined.

**Incorrect Use Of Correlation Analysis In The E-RMS Software Manual**

The software manual states the following:

"Stopping Criterion (Statistical Significance). Before a model is derived by induction the user must specify a stopping criterion. This stopping criterion is expressed in terms of a statistical significance level (if you have no knowledge of statistics then consult any basic statistical textbook for an introduction to the concept of statistical significance). For example a significance level of 0.05 indicates that the branching patterns in a derived model has a less that 5% probability of being based on chance correlations between the training and predictor data sets. In other words you can be at least 95% confident that the branching structure of the model reflects real rather than chance correlations."

When correlation analysis is used indiscriminately, like this, false conclusions can result. Incorrect use of correlation analysis, more than any other factor, has probably led many people to feel that a clever statistician can use data to support both sides of an argument (Kenkel; 1984).
An important theory in economics states that, when other things are equal, house values are negatively correlated with the distance of the house from the central business district. However, for most cities, a collection of houses sampled at random indicates that houses in the city cost more than houses near the central business district. That is, the data indicates a positive correlation between house price and house distance from the central business district. This spurious correlation arises because other things are not equal. Generally, houses in the suburbs are larger, have larger yards, and other attractive features. Once these considerations are taken into account, the theory can be shown to be true (Kenkel; 1984).

From the preceding discussion it may be seen that because the attributes are not assigned independently and assigning attributes partitions the database into a family of equivalence classes "all other things are not equal." It is incorrect to discuss correlation analysis in the manner described in the software manual because it ignores all the variables that may influence a single variable other than one specific selected variable. In order to attempt to determine the correct relationship between variables it is necessary to build a model that includes the interactions between all the variables. Additionally, discussion in Chapter 2 has pointed out that there need not be a unique collection of attributes that can be used to aggregate a particular subset of location elements. Consequently, the outcome of the branching analysis attempted could depend on the order in which the variables are selected and analysed for "statistical relationships".

**Introducing Absolute Probabilities Of Occurrence**

The notion of probability adopted in the software manual can not be reconciled with that of stochastic regularity. The "DERIVE MODEL BY INDUCTION" option of the predictive modelling module is used to derive a decision tree model. As described above, the process proceeds by correlating the distribution of remotely mapped variables with ground surveyed data. Frequently, there is complete geographical coverage for remotely mapped variables and correlations between such variables and ground surveyed data for different species of limited distribution are used to extend or interpolate ground surveyed data across unsurveyed areas.

At this point the "logic" of the software manual becomes extremely difficult to follow and this is at least partly because of an attempt to define a distinction between entities described as being the "absolute probabilities of occurrence" and "relative likelihood's of occurrence". The concept of an "absolute probability of occurrence" is not one which is defined in the Collins Dictionary of Mathematics and does not appear to be known to the subject of mathematics.

**An Incorrect Application Of Bayes Theorem Based On An Absolute Probability Or Relative Likelihood Of Occurrence.**

The problem that is being addressed is that of estimating the "probability" that an entity would be located on a pixel containing a particular attribute. An example of what is being attempted is to calculate the probability that species X would be located on a pixel having the attribute 500 metre elevation.

In order to calculate the "probability that the pixel on which species X was observed to occur possesses a particular attribute" a method based on an interpretation of Bayes rule is used. The software manual states:
"Bayes rule ... requires an estimate of the prior probability for each of the categories. The prior probability of a category is the probability of that category occurring at any grid cell in the database area before knowing anything about the predictors at that grid cell. A category's prior probability is therefore the proportion of the database covered by that category."

**Applying Bayes Theorem To A Process That Is Stochastically Closed.**

This statement appears to be based on a number of incorrect assumptions. Bayes Rule is used to revise an estimate for the probability of an event happening as new information becomes available. Bayes rule can be stated as (Kenkel; 1984):

\[
\text{Prob(Hypothesis | Fact)} = \frac{\text{Prob (Fact | Hypothesis) * Prob (Hypothesis)}}{\text{Prob (Fact)}}.
\]

It is implicitly assumed that the probability of the fact differs from zero. This is a reasonable assumption because if it were false then the fact would never have arisen in the first place. As an example, Bayes rule can be used to update the probability of the sum of two die throws equalling a particular number between 1 and 12 once the first number rolled becomes known.

**Taking A Random Selection Of Elements From The Set Of Location Elements.**

By recognising that the database consists of a finite collection of distinct location elements \((L_1, L_2, L_3, ..., L_M)\), each of which has a single member of \(\Omega\) associated with it, it is apparent that it is possible to develop a method for selecting a location element at random. Because the number of location elements is finite and they are all distinct, it is possible to uniquely associate an integer between 1 and M, where M is the number of location elements described by the set \(\Theta\), with each location element. A method of randomly selecting an integer between 1 and M could be applied and this method would allow selection of a location element at random. Having selected a location element in this random manner the properties coded as present at that location element could be observed and recorded.

Over a long series of such trials, for each property \(p_i\) (for \(i = 1\) to \(N\)), one would expect the proportion of location elements selected containing the Property \(p_i\) coded as present, to converge to the relative frequency of location elements coded as having the Property \(p_i\) present. That is to say, if 50% of the location elements have the property \(p_i\) coded as present, then over a long series of trials, the proportion of location elements selected having the property \(p_i\) coded as present should converge to 50%.

The fact that the inherent property description of each location element, in any locationally linked database, can be uniquely described by an element of a constructed set \(\Omega\) (Chapter 2) leads to some most inconvenient results.

For abstract random independent events A and B \(P(A \text{ and } B) = P(A) * P(B)\). When selecting location elements from the database, on which equivalence classes are defined by a partition of the set \(\Omega\), as described in chapter 2, independence can never be established. For any selected location element it does not make any sense to consider the probability of any property being coded as present. The property is either coded as present or it is not; there is no probability about it. Having selected a location element the properties coded as present are completely described by the
member of $\Omega$ that defines the class to which the location element belongs; the process is stochastically closed. Under such circumstances the statement required to establish the probability of a joint event is:

$$P(p_i \text{ and } p_j \text{ coded as present together}) = P(p_i \text{ coded as present}) * P(p_j \text{ being coded as present when selecting location elements from the subset of } \oplus \text{ having } p_i \text{ coded as present}).$$

Because the probability of two properties being simultaneously coded as present, at any location element, depends on the intersection of subsets of $\Omega$ it is not ambiguous to consider the probability of obtaining a location element exhibiting two properties $p_i$ and $p_j$ coded as present simultaneously. That is:

$$P(p_i \text{ and } p_j \text{ together}) = P(p_i \text{ coded as present}) * P(p_j \text{ being coded as present when selecting location elements from the subset of } \oplus \text{ having } p_i \text{ coded as present}) = P(p_j \text{ coded as present}) * P(p_i \text{ being coded as present when selecting location elements from the subset of } \oplus \text{ having } p_j \text{ coded as present}).$$

The above formula can be extended to establish the probability of obtaining a location element at which any number of properties between 1 and $N$ are simultaneously coded as present.

For independent events, $A$ and event $B$, from the same sample space, the probability of event $A$ and/or event $B$ occurring is obtained by establishing the simple algebraic sum of the probability of event $A$ occurring and event $B$ occurring. That is, $P(A \text{ and/or } B) = P(A) + P(B)$. Again, the general result for independent events, does not hold when taking a random selection of elements from the set $\oplus$. From set theory it is know that:

$$P(A) + P(B) = P(A \cup B) + P(A \cap B);$$

It follows that:

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The probability of obtaining a location element at which at least one of the two properties, $p_i$ and or $p_j$, are coded as present is determined by proportion of location elements having the property $p_i$ and/or the property $p_j$ coded as present minus the proportion of location elements having both the properties $p_i$ and $p_j$ coded as being present.

From the preceding discussion it is apparent that when using a random method to select a pixel on the database the probability of selecting a pixel having a particular characteristic (say 200 millimetre rainfall) is given by the relative frequency of that attribute. To calculate the probability of selecting a pixel containing two characteristics simultaneously it is essential to examine the number of location elements coded as having both the properties present simultaneously. In addition, probability only has any meaning before location elements have been selected. Once a location element has been selected it is absurd to talk about the probability of it having a particular attribute - it either does have that attribute or it does not. Once an individual grid cell has been chosen the whole notion of probability no longer has any meaning. The Probability function has "collapsed" and has been replaced by a certain
outcome. For any particular grid cell an attribute is present or absent with "probability" zero or one. Bayes Rule can not be applied to a process that is stochastically closed.

Only when location elements are selected at random is a category's prior probability equal to the proportion of the database covered by that category. It is known that organisms are not distributed at random and consequently a category's prior probability for an organism can not correctly be said to be given by the proportion of the database covered by that category. Furthermore, once the "prior probability" of one category has been established attempts are made to estimate the "prior probability" of the next category. It is assumed that the "prior probability" of the additional category is the probability of that category occurring at any grid cell in the area described by the database before knowing anything about the predictors (other categories) at that grid cell. Such an assumption does not take any account of the partitioning process and consequently the assumption is wrong. Because it does not make any sense to speak about the probability of a selected pixel or collection of pixels having particular attributes, attempts to compare attributes selected with attributes that should have been selected, based on a random prior probability of occurrence, seems to be a most unreasonable thing to do. Extending ground surveyed data in this way; using correlations between remotely mapped variables, can not be done in a probabilistic sense. The ground surveyed data is either similar at the other locations or it is not; there is no probability about it. This problem is, in part, exactly that of using statistics to say something about the part from the whole. Inductive statistics can be used only to ascertain properties of the whole from that of a sample.

**Applying Bayes Theorem: Assuming The Mean Of A Population Approaches That Of A Sample**

The application of Bayes rule from limited survey information, to infer the distribution of an entire population, requires the assumption that the mean of the population tends to (or is distributed about) that of the sample. One of the most common errors when using Bayesian Statistics is this assumption: that the mean of a population approaches that of a sample (Goode; 1962). Such an assumption is false and is the converse of the true situation: subject to certain conditions the mean of a collection of sampled properties tends to that of the population. Attempts to broaden the use of descriptive statistics from the sample on which they were based to cover the entire sample space are subject to sampling error. It follows that even if independence could be assumed, conclusions drawn from a surveyed sample applying Bayes' Rule, in the manner described in the E-RMS software manual, may turn out to be false when applied to the entire population described over the database; particularly in the case where biological selection is occurring.

**Deriving Inferences from Overlaying Maps**

**E-RMS Determining Areas For Which Fire Potential Is High.**

In tutorial exercise 3 areas of high fire potential are derived by Boolean overlaying vegetation, slope, aspect and fire history. The vegetation categories selected as exhibiting high fire potential are open forest, woodland, woodland/cleared, scrub, lantana, and river terrace. In the overlay operation the selected categories of vegetation are combined with slope and land aspect data. The feature so created is
name "HIGH FIRE POTENTIAL" and is then saved as a "USER - DEFINED FEATURE."

Accuracy When Overlaying Characteristic Regions On Thematic Maps.

Instruction 17 of tutorial exercise 3 says:

"You could go on specifying additional overlay variables until the cows come home (there is no limit in E-RMS). This will do for the current example so select NO MORE VARIABLES REQUIRED."

Whilst one can go on overlaying categories "until the cows come home" there comes a time at which the outcome from overlaying additional categories no longer resembles the outcome the operator intended. It is unreasonable to continue overlaying maps without examining the quality of the information being obtained.

Macdougal (1975) explained that in maps where characteristic regions are represented at least two types of accuracy are important, the precision (the upper bound on accuracy) with which the boundary lines are located and the extent to which the soils on the ground represent the regions coded.

The lower limit to the accuracy of a map overlay operation consists of the sum of the allowable positioning errors and the product of the purities of the constituent factor maps, plus errors which are introduced in the assembly of the final overlay. The sum can be expressed as:

$$A_{\text{min}} = f(\Sigma H(i),\Pi P_i,e);$$

where:

- $A_{\text{min}}$ is the probability that a specified combination actually occurs at the indicated location on the overlay map, $H$ is the allowable positioning error, $P$ is the purity on the $i$th map, $e$ is the error introduced in the assembly of the overlay.

The possible magnitude of each of these terms in a typical application can be quite large. For example, in an overlay of six maps, each with an allowable horizontal error of 0.5 mm (equivalent to Australian Topographic National Map Accuracy Standards) and a purity of 0.8 (a good soils map), the possible horizontal error in the location of boundaries is 3 mm, and the purity of the overlay is 0.806, or 0.21. On the basis of purity alone, one could argue that such an overlay map is not significantly different from a random map. The soil mapping criteria of the US Soil Conservation Service explicitly states that regions on detailed large scale maps may include up to 15% of a substantially different soil type, and over 15% may be soils with different management requirements (Macdougal; 1975). However Macdougal (1975) does claim that it is unlikely that the errors in the constituent maps are so independent and systematic that map overlays are this inaccurate, particularly in the location of boundaries. In optimal circumstances it may be reasonable to assume that the purity of the overlay is that of the least accurate map in the combination and the inaccuracy in boundary location is the average boundary error ($H(\bar{\text{bar}})$). In this case the upper limit to the overlay accuracy is

$$A_{\text{max}} = f( H(\bar{\text{bar}}),P_{\text{min} \text{,} \text{e}})$$

If the errors in the source maps (for example vegetation and soils maps) do not occur independently, in this case possibly because the vegetation determines the characteristics of the soils and the soil in turn directs the growth of vegetation, then
the categories are not independent. Information in the vegetation map makes it possible to infer and recover information about soil properties. Consequently, the amount of new information gained by overlaying a soils map is significantly less than may have at first been expected. Additionally, the fact that maps contain information that is not independent of other maps immediately imposes constraints on the types of questions that one can expect to get meaningful answers to. Without knowing how the information contained in each map is linked to the information in other maps it may be quite simple for an operator to ask a wrong question. Finally, because not all biological sets are well defined, maps in which regions are coded, such as soils maps, often do not have precise boundaries between regions.

Macdougal (1975) concludes by saying:

"It is clear from this discussion that some overlay maps may indeed differ little from random maps, and that most overlay maps contain more error than their compilers probably realise. ... Most important, the compilers of overlay maps should attempt to estimate their accuracy, and present this in the legend or the accompanying text."

To interrogate a soil series map with logical (Boolean) operators it is essential to invoke the erroneous assumption that the soil series is a homogeneous unit of productivity exhibiting consistent properties with respect to other soil series. It has been demonstrated that this assumption may lead to significant errors in estimates of potential productivity and according to Gersmehl (1980) this has profound implications for prime land delimitation, tax assessment, zoning administration, and other policies that presently rely on published soil productivity ratings.

In order to test the question: can soil series serve as a functional unit for comparisons of productivity; Gersmehl (1980) selected pairs of counties for which soil surveys were completed within a few years of each other (so chosen to minimise the effect of historical changes or soil survey methods). For an arbitrarily selected pair of common soils, tables of estimated yields in the County soil surveys were used to prepare a table of average yields for the major crops in the region. Comparisons were made only after controlling for slope, erosion, class, surface texture, and level of management.

A summary of Gersmehl's (1980) conclusions follows:

"The obvious conclusion is that an attempt to use the relative productivity of two soil series in one place as a predictor of yields elsewhere is likely to yield considerable error. Any regional map of soil productivity that uses a series as a unit should therefore be viewed with suspicion, if not rejected outright.

....

All ... sources of environmental variation may contribute to differences in the estimation of relative productivity of two soils in different places. Attempts to determine the precise reason for a given discrepancy are undoubtedly significant but are beside the point of this paper, which is simply to demonstrate that maps of soil productivity based upon relative rankings of soil series in one place are inevitably questionable in other places.

....
To summarise, any regional map of soil productivity that uses the series as a unit is on methodological quicksand. Moreover any policy of tax assessment, zoning, or land-use planning that depends on such general indices of productivity may inflict serious inequity on property holders. Finally, any regional assessment of land-use potential that is based on such an inequitable map may lead to significant misallocation of public resources.

E-RMS Using User Defined Categories Like Any Other Variable.

Considering the user defined variable, high fire potential, instruction 29 of tutorial exercise 3 states: "User defined categories within this variable can then be used like the categories of any other variable in subsequent analysis."

Logical operators were used in the construction of the variable high fire potential and it is essential that the restrictions imposed by the logical operators used in the construction of the variable high fire potential not be contradicted by instructions used in subsequent analysis. Unfortunately, once a variable such as high fire potential has been created, there is no provision within E-RMS to explicitly state exactly how the variable was constructed.

1 It is possible to ask a wrong question if a false assumption is implicit in the statement of the question.
Chapter Seven: E-RMS Uses A Defective Land Surface Reconstruction Algorithm For Modelling Fire Intensity.

- Outline of Chapter
- Information - The Ability To Discriminate Between Alternatives
- Measurement And Uncertainty
- Choosing A Numerical Estimation Algorithm
- The E-RMS Land Surface Reconstruction Algorithm
- Symmetric Tests of the E-RMS Land Surface Reconstruction Algorithm
- Estimating Fire Intensity Using Data Derived from the E-RMS land surface reconstruction algorithm

Outline of Chapter

This Chapter is concerned with the E-RMS land surface reconstruction algorithm and fire intensity modelling applications. This chapter contains slope, aspect and elevation pictures that were screen captured directly from screen images produced by the E-RMS Tutorial Database. These figures demonstrate square kilometres of flat land coded as having an aspect and incorrect slope and/or elevation coding along half-kilometre fronts. A symmetric test of the reconstruction algorithm comprehensively demonstrates that the slope and elevation the reconstruction algorithm assigns depends on the direction assigned north in the database. Specifically, the fire intensity model described in the E-RMS software manual uses categories of slope that the reconstruction algorithm can not distinguish between, relies on a reconstruction that will not detect a cliff of 274 metres, adds vector quantities as scalars and will code a rise of 100 metres over a horizontally projected distance of 100 metres as having a "slope" of 3 to 6 degrees.

Information - The Ability To Discriminate Between Alternatives

There exists a fundamental distinction between data and information. Consider the symbols A, B, C as representing individual pieces of data; the amount of information we can extract or convey using these symbols depends on the structure imposed on the way the symbols can be presented. Suppose that these symbols are transmitted in groups of three and that each transmission block must contain all three of the symbols, that is, one A and one B and one C. A transmission block can be represented by the notation ( , , ).

If no ordering property in the presentation of the symbols A, B, C can be recognised then a transmission block can not be used to convey any information. The reason for this is that the transmission signal is constant as (A, B, C) is defined as being equivalent to (C, B, A), (B, C, A), (A, C, B), (B, A, C), and (ACB). Alternatively, if it is possible to recognise the obvious ordering property then transmitting the symbols A, B, C as a block imposes the restriction that (A, B, C) is different from any other
distinct rearrangement of the symbols A,B,C. The same three symbols can be used to convey 6 discrete messages on the one transmission. So in the first case data transmission only is possible, while in the second case the transmission of the same amount of data (the three symbols A,B,C) provides the ability to discern information.

There are two distinct kinds of information theory that are of concern here. The first is statistical information theory which is concerned with what happens over a long series of uncertain situations where information must be transmitted through some sort of communication channel. In 1948 Shannon published two papers on "A Mathematical Theory of Communication" in the Bell Systems Technical Journal and these papers laid the foundation for the mathematical theory of information. Semantic information theory addresses what we actually mean by the symbols we invoke to communicate (Hintikka; 1970).

Information must be structured in a coherent way and it is possible to extract useful information from data only when both the structure of the information being sought and the structure of the data from which the information must be extracted are understood.

Geographic information systems for environmental analysis generally have the capacity to interpolate slope, aspect and elevation data from contour lines. This problem represents an attempt to reconstruct a physical entity from incomplete information. In this case the image is a series of sampled points (contour lines). When information is not sufficient to permit an exact reconstruction of an entity, it is impossible to separate the constructed image from the series of instructions that created it. In any data processing operation conservation of information is the best outcome that can be achieved.

**Measurement And Uncertainty**

Consider a measuring rod of unit length on which N equal spaced intervals are defined. In order to observe a change, \( \Delta t \), in a measured variable, it is necessary that \( t \) and \( t+\Delta t \) lie on opposite sides of a division on the measuring rule. In order to guarantee that \( \Delta t \) be observable it is necessary that \( \Delta t > 1/N \). This statement asserts nothing more than if it is desired to measure with an accuracy of 1 part in N then there must exist at least N distinguishable states (Resnikoff; 1989). No numerical estimation method can compensate for or correct a problem due to inadequate information. Therefore, before embarking on any attempt to code information about ground slope, aspect, or elevation from contours it is necessary to consider the information available.

**Choosing A Numerical Estimation Algorithm**

All reconstructions that can be put into digital format must reconstruct by some sort of numerical method. It is essential to establish good ground control procedures in any photogrammetric mapping operation. It is impossible for the accuracy of a finished map to be any better than the ground control upon which it is based. Many maps that have been processed to exacting standards have failed to pass field inspection; simply because the ground control was inadequate. Subject to varying conditions, the cost of establishing ground control for photogrammetric mapping can be expected to range between 20\% and 50 \% of the total mapping cost (Wolf; 1983). In the case where slope, elevation and aspect data are being estimated from contours it is easy to check
how faithful the reconstruction of the image is. If the correspondence between reality and the numerical method's estimate is unsatisfactory then either:

- A more satisfactory reconstruction algorithm should be sought; and/or
- The reconstruction algorithm should be provided with additional information; or
- The attempted reconstruction should be abandoned (Kreyszig, 1988).

**The E-RMS Land Surface Reconstruction Algorithm**

- E-RMS Specifying The Slope And Aspect Of A Grid Cell.
- Angles That Can Not Be Measured Coded In The E-RMS Demonstration Database.
- Interpolating Between Contours.
- Failure To Detect A Single Cliff In 520,000 Hectares Of Mountainous Territory.
- Contradictory Results Produced By The E-RMS Land Surface Reconstruction Algorithm. Incompatible Slope Elevation and aspect Data.

The E-RMS software manual does not explain in any detail how the terrain reconstruction algorithm calculates elevation, slope and aspect from contour lines. All the information provided about the "least squares plane" in the software manual is reproduced in the description and picture for Figure 7.1.

Figure 7.1: Reproduction of picture describing E-RMS reconstruction algorithm. A "least squares" plane is fitted to the elevation data for the eight points demonstrated. The plane is then used to calculate the elevation, slope and aspect at the centre of the grid cell (E-RMS Software Manual).
From Figure 7.1 it appears that the algorithm samples the height of each of the
contours in eight directions from an angle of zero degrees (north) through to 360
degrees in even increments of 45 degrees. An averaging process is then performed to
arrive at an estimate for the height at each pixel.

In Figure 7.2 contour lines are shown. It is apparent that the vertical trajectory up the
page does not cross the 100 metre contour line shown. The horizontal trajectory
across the page detects both the 100 metre and the 10 metre contour line.
If one had to choose between the linear height estimate provided by the vertical trajectory and that provided by the horizontal trajectory, at the intersection of the two trajectories, then the most reasonable guess would be that provided by the horizontal trajectory. A linear interpolation between the 100 metre and 10 metre elevation values along the horizontal trajectory would estimate the height at the intersection point as being approximately 50 metres. The approach taken by the E-RMS reconstruction algorithm is to estimate the height by taking some sort of an average between the height given by the vertical and horizontal trajectory. For example, the height could be estimated as 20 metres $((10+10+10+50)/4$ metres). Such a method fails to take into account the greater information content of the horizontal trajectory. Information makes it possible to distinguish between alternatives and consequently the sample values provided by the horizontal trajectory have a higher information content than the sample values provided by the vertical trajectory. Such a method will tend to "average out" all the sharp changes in slope and elevation and in so doing "flatten the database". The application of a simple arithmetic average in a situation such as this is mistaken. Least squares functions and orthogonal projections are generally used to solve equations that are inconsistent only in very limited circumstances (Hill; 1986). A least squares approach assumes that the error or variability in measuring the dependent variable is minimal in comparison to the independent variable. The assumption of linearity can not be sustained under circumstances where it is required to reconstruct a land surface from contour lines.

**E-RMS Specifying The Slope And Aspect Of A Grid Cell.**

The E-RMS land surface reconstruction algorithm specifies the slope and aspect of a grid cell. The software manual never makes it clear precisely what is meant by such a description. Only in the special case where the height of each corner of the raster lies
on the plane defined by the heights at the other 3 corners is it possible to unambiguously talk about the slope and aspect of a gridcell. If the ground elevation at the fourth point does not lie on the said plane then, unless a selection rule is imposed, it is possible to obtain different planes with many slopes and aspects for the one raster. It seems reasonable to assume that subject to some restrictions, aspect is the direction of the normal to the tangent plane defined by the "slope" assigned to the grid cell. Unfortunately, the interpolated slope and aspect data calculated by the E-RMS land surface reconstruction algorithm can not be reconciled with the calculated elevation profile

**Angles That Can Not Be Measured Coded In The E-RMS Demonstration Database.**

The best available topographic maps for the Apsley Macleay National Park are scaled at 1:25,000 and this is as shown in the catalogue Medium Scale Mapping New South Wales 1993 (CALM; 1993). The 1:25,000 maps that could have been used to construct the database have a contour interval of 10 metres.

Since the raster size stated for the E-RMS Apsley Macleay database is 100 metres by 100 metres the reconstruction can, at best, sample the ground height at 100 metre intervals. It requires only elementary geometry to determine categories of slope that can be differentiated under such circumstances. Angles that can be differentiated using a fixed 10 metre contour interval over a horizontally projected distance of 100 metres are given by the formula \( \alpha = \tan^{-1}(n \times 10/100) \); where \( n \) is an integer greater than zero.

<table>
<thead>
<tr>
<th>n</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation change (M)</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td>100</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>Angle ( \alpha ) (degrees)</td>
<td>5.7</td>
<td>11.3</td>
<td>16.7</td>
<td>21.8</td>
<td>45.0</td>
<td>63.4</td>
<td>71.6</td>
</tr>
</tbody>
</table>

Table 1: Some values of \( n \), with their corresponding change in height over 100 metres and corresponding angle.

It is apparent that the first category of slope that can be differentiated is 6 degrees. The largest category of slope that is coded in the E-RMS demonstration database is 70 degrees. Taking such a category of slope as defining a cliff it is simple to show that the minimum detectible cliff would have to be at least 275 metres (\( 100 \times \tan 70 \)). For comparison, the air-craft warning beacon on top of the Sydney Harbour Bridge is 141 metres above mean seal level and this is just over half the height of a shear drop required to be described as being "a cliff" when sampling heights 100 metres horizontally apart (Anon; 1982).

The rise of 10 metres over a distance of 100 metres corresponds to an angle of 5.7 degrees. This can not be reconciled with the fact that the minimum stated category of slope for the E-RMS demonstration database is 0 to 1 degrees. To code categories of slope as being 0 to 1 degree, 1 to 3 degrees, and 3 to 6 degrees at 100 metre resolution and commercially supply the package without a specific caveat is to imply that the
reconstruction has distinguished between these categories at the 100 metre level. From the preceding discussion it is clear that such a claim is false.

**Interpolating Between Contours.**

Interpolation between contours may be used in some circumstances to code lower categories of slope than can be measured. The use of such techniques introduces coding error, as opposed to coding mistakes. When interpolating between contours it is not possible to distinguish between alternatives that can not be measured. The information content of an interpolated reconstruction is no greater than that which can actually be measured from the original map. This statement asserts nothing more than in the case where there exists N distinguishable states for measuring an entity, and M separate states are defined, where M>N, the assignment of an entity to at least M-N of the states can not be determined exactly.

In practise, the assignment of an angle to a class that can not be measured may be determined statistically. For example, when reconstructing with a 10 metre contour interval and coding slopes over 100 metre intervals it may be possible to code categories of slope 0 to 3 degrees and 3 to 6 degrees and estimate (frequently from additional information) that 90% of entities coded as having slope 3 to 6 degrees indeed have such a slope. Importantly, in any one instance, it is not possible to say if the slope at any single location is 0 to 3 degrees or 3 to 6 degrees. In order to interpolate categories of slope 0 to 1, 1 to 3, and 3 to 6 degrees from a 10 metre contour interval, over a projected distance of 100 metres, it is necessary to assign 3 classes under circumstances where it is possible to measure only 1.

When using a 10 metre contour interval it is possible to differentiate a slope of 0 to 1 degree from a slope of 1 to 3 degrees only when sampling at intervals of 573 metres; this is an interpolation over half a kilometre. Coding a slope of 0 to 1 degree under such circumstances involves a linear interpolation beyond the original information by a factor of over 5. This simply can not be done in a mountain range such as that "described by" the E-RMS tutorial database. The variance of the terrain is too great to reasonably assign "the slope of a grid cell" to one of the 3 non-distinguishable classes between 0 and 6 degrees. Furthermore, in the E-RMS Apsley Macleay database, the elevation is partitioned into 100 metre categories. If an angular reconstruction over 100 metres is attempted from 100 metre elevation categories then the minimum potential uncertainty in the first stateable slope is over 45 degrees.

**Failure To Detect A Single Cliff In 520,000 Hectares Of Mountainous Territory.**

Figure 7.3 and 7.4 demonstrate that the tutorial database has not detected any cliffs over an area of approximately 5,100 square kilometres. The region described is known to be rugged and mountainous. When adopting a linear reconstruction (a straight line as implied by coding a constant angle over a fixed distance) over 100 metres the minimum sized cliff that can be detected allowing concessions that are both computationally and theoretically impossible is, as previously stated, approximately 275 metres and therefore it is not surprising that no cliffs have been detected. In any photogrammetric mapping operation ground testing of the final product is essential. The categories of slope formulated and coded in this tutorial package could not be expected to survive any serious ground testing operation.
The fact that cliffs of over 270 metres can escape detection raises questions about the representation being obtained and the ability to discriminate between alternatives at the ten kilometre by ten kilometre level - let alone the 100 metre by 100 metre level. Large cliffs can act as drainage points and have substantial affects on local microclimate for kilometres around. Consequently, it would be false to assume that a rainfall interpolation algorithm run over a database such as this has any ecologically...
meaningful relationship with soil moisture content or local humidity. Large cliffs cast shadows and this changes the intensity of incident solar radiation, thereby changing temperature variance and average temperature.

Contradictory Results Produced By The E-RMS Land Surface Reconstruction Algorithm. Incompatible Slope Elevation and aspect Data.

A close-up look at the reconstructed slope data reveals that every now and again a distinct slope just seems to "pop-up" in amongst a field of fairly similar slopes. An enlargement of a particular region where this is occurring is shown in Figure 7.5. At the coordinates shown in the picture it can be seen that 3 successive grid cells are assigned 3 successive categories of elevation and that the category of slope assigned is not 40 to 70 degrees. Because there are no grid cells coded as "having a slope" of over 70 degrees (Figure 7.3) it follows that the maximum category of slope that can be assigned to these grid cells is less than 40 degrees. Such a slope assignment can not be reconciled with the geometric fact that a vertical rise of 100 metres over a horizontally projected distance of 100 metres represents an angular gradient corresponding to 45 degrees.

Figure 7.5. Close up of E-RMS tutorial database slope reconstruction and elevation.

Figure 7.6. Close up of E-RMS tutorial database slope reconstruction and elevation. Figure 7.6 demonstrates a pixel which is coded as having "slope" 3 to 6 degrees and from the surrounding elevation it is apparent that the change in elevation across a 100 metre by 100 metre pixel is of the order of 100 metres. This represents an angle of approximately 45 degrees.
Figure 7.7. E-RMS tutorial database containing flat land with aspect.

Figure 7.7 demonstrates that the reconstruction algorithm applied to the tutorial database has reconstructed pixels as having both a slope between zero and one degree and an aspect. These failures are not just isolated events and occur over square kilometres of the database.
Symmetric Tests of the E-RMS Land Surface Reconstruction Algorithm

- Procedures Followed To Carry Out The Tests.
- Results Of Symmetric Tests On The E-RMS Land Surface Reconstruction Algorithm.
- Test 1.
- Test 2:
- Test 3.
- Test 4:
- General Analysis Of The Reconstruction Test.
- An Adequate Notice Of Data Quality In This Case Would Have To Warn A User That The Data Was Misleading And Wrong.

I ran a series of tests to further investigate the performance of the E-RMS package's aspect, slope and elevation reconstruction algorithm.

Procedures Followed To Carry Out The Tests.

In order to do these tests a database was set up as per the instructions in the E-RMS software manual. The following information was entered into the database specification module:

<table>
<thead>
<tr>
<th>NAME</th>
<th>UWSH CAMPUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRID CELL SIZE (metres)</td>
<td>10</td>
</tr>
<tr>
<td>NUMBER OF COLUMNS</td>
<td>704</td>
</tr>
<tr>
<td>NUMBER OF ROWS</td>
<td>500</td>
</tr>
<tr>
<td>ORIGIN AMG EASTING (metres)</td>
<td>88000</td>
</tr>
<tr>
<td>ORIGIN AMG NORTHING (metres)</td>
<td>75000</td>
</tr>
</tbody>
</table>

Table 2: Entries in Database Specification Module for UWSH Campus.

The AMG coordinates stated for the database are not AMG coordinates. It is not possible to enter unambiguous AMG coordinates into the E-RMS database specification module. The purpose of stating the above coordinates was simply to allow the database to define a rectangle in geographical space from rectangular coordinates.
A pair of compasses were used to inscribe two circles onto a sheet of uncreased A3 Reflex copying paper. The corners of the paper sheet were then used to define an "exact rectangle", as stated in the database specification module. Having attached the A3 sheet of paper to the digitising board the circular "contour lines" on the A3 sheet of paper were digitised into the database. A total of four distinct tests were run. Each trial specified the circular "contour lines" as being at different heights. In each trial the larger circle was defined as being at the lower height and the smaller circle defined as being at the higher height. The specific details of the tests are set out below.

The contours were digitised using a conventional Calcomp 33600 large format digitiser. The digitiser has a high resolution: up to 100 lines per millimetre and an accuracy of 0.254 millimetres (Calcomp; 1991). It was not feasible to determine how accurately the contours were digitised into the database since the software does not have any routine for checking this. However, to maximise accuracy a new sheet of paper that had never been folded was used. Hand digitising of contours involves a long series of repetitive operations and thus, despite care, mistakes, systematic, and random errors may occur. Any comments regarding the slope and aspect data interpolated by the E-RMS software must therefore be considered within the limitations imposed by the constraints of the experimental method.

Once the contours were entered into the database the commands necessary to instruct the software package to interpolate intervening elevation, slope and aspect data for each pixel in the database were executed. The categories of aspect, elevation and slope made available to the machine are indicated on the print outs presented with the results of each calculation.

Test 1: Attempted aspect, elevation, and slope reconstruction from two concentric circles of diameter approximately 200\(^1\) and 23\(^2\) pixels across. The largest circle was defined as having an elevation of 1 metre and the smallest circle an elevation of 10 metres.

Test 2: Attempted aspect, elevation, and slope reconstruction from two concentric circles of diameter approximately 200 and 23 pixels across. The largest circle was defined as having an elevation of 10 metres and the smallest circle an elevation of 100 metres.

Test 3: Attempted aspect, elevation, and slope reconstruction from two concentric circles of diameter approximately 200 and 23 pixels across. The largest circle was defined as having an elevation of 50 metres and the smallest circle an elevation of 500 metres.

Test 4: Attempted aspect, elevation, and slope reconstruction from two concentric circles of diameter approximately 200 and 23 pixels across. The largest circle was defined as having an elevation of 100 metres and the smallest circle an elevation of 1000 metres.

The tests were selected because they provide the database with contour lines having circular symmetry and information that has such symmetry can be used to evaluate the directional stability of a reconstruction algorithm.

**Results Of Symmetric Tests On The E-RMS Land Surface Reconstruction Algorithm.**
Figure 7.8: This figure contains circular contours digitised from the circles inscribed on the A3 sheet of paper. The cross hatching represents the gridlines for the "map" shown as a line feature. This figure demonstrates the case where the larger circle is defined as having an elevation of 1 metre and the smaller circle an elevation of 10 metres. By counting the number of grid lines east-west and north-south along the database it was possible to calculate the diameter of each of the inscribed circles at map scale.

**Test 1.**

Attempted aspect, elevation, and slope reconstruction from two concentric circles of diameter approximately 200 and 23 pixels across. The largest circle was defined as having an elevation of 1 metre and the smallest circle an elevation of 10 metres.
Figure 7.9: Interpolated elevation data. Calculated when the large and small concentric circles were defined as having an elevation of 1 and 10 metres respectively. The two circular contour lines are shown as black line features.

The elevation profile produced by the machine for this reconstruction demonstrates a major reconstruction failure. The presence of the two isolated green patches indicated as having 5 metre elevation suggests that the reconstruction algorithm is accumulating errors as it attempts to reconstruct. Inside the smaller circle the elevation was reconstructed as being 6 metres and this suggests another reconstruction failure because the entire area is enclosed by a 10 metre contour interval.

At locations shown 2 metre elevation was coded. From the elevation profile reconstructed for these contour intervals it is clear that the elevation is being reconstructed differently along different trajectories from the origin of the two circles. The most probable explanation for this is that at certain locations the reconstruction algorithm can not "observe" the smaller circle. The directional failure occurred despite the fact that the two circles were separated by a "constant" linear distance. This demonstrates the algorithms non-uniform sampling and interpretation of information on a "flat two-dimensional" database (or sheet of paper), partitioned into square pixels.

Figure 7.10: Close up of circular section of interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 1 and 10 metres respectively. The two circular contour lines are shown as line features.
From Figures 7.11 and 7.12 it is apparent that no aspect or slope data was transferred by the E-RMS software to the map display analysis module. Despite repeated attempts it was not possible to get any aspect or slope data forwarded from the map digitising module to the map display analysis module for this reconstruction. The machine did not appear to determine that the pixels in the database did not have an aspect. Rather, the software program flagged some sort of an operating error and failed to respond to further instructions. It was necessary to reassemble the aspect and
slope categories following the error. The reason for the error is unknown.

Figure 7.12: Interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 1 and 10 metres respectively. The two circular contour lines are shown as line features.

**Test 2:**

Attempted aspect, elevation, and slope reconstruction from two concentric circles of diameter approximately 200 and 23 pixels across. The largest circle was defined as having an elevation of 10 metres and the smallest circle an elevation of 100 metres.

Figure 7.13: Interpolated elevation data calculated when the large and small
concentric circles were defined as having an elevation of 10 and 100 metres respectively. The two circular contour lines are shown as line features.

The presence of the two green patches indicated as having 50 metre elevation provides further evidence of error accumulation and reconstruction failure. There are locations that were coded as having elevation significantly lower than they reasonably should have been; the 10-20 metre sections indicated.

Inside the smaller circle the elevation was reconstructed as being 50-60, 60-70, 70-80, and 80-90, metres. This is an interesting reconstruction failure because the entire area is enclosed by a 100 metre contour line and the reconstruction algorithm has reconstructed a concave cap on top of this "hill"; somewhat like an exploded volcano cap. The little black circle representing the 100 metre contour line can be seen clearly on the zoomed in image in Figure 7.14.

![Figure 7.14: Close up of circular section of interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 10 and 100 metres respectively. The two circular contour lines are shown as black line features.](image)

![Figure 7.15: Interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 10 and 100 metres respectively. The two circular contour lines are shown as line features.](image)
The machine failed to send any aspect data through from the map digitising module to the map display analysis module. As for the previously mentioned aspect and slope reconstructions the aspect variable had to be reconstructed between each successive attempt to get the data sent through.

Figure 7.16: Interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 10 and 100 metres respectively. The two circular contour lines are shown as line features.

This particular slope profile can not reasonably be reconciled with a set of contour lines having circular symmetry.
Figure 7.17: Close up of circular section of interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 10 and 100 metres respectively. The two circular contour lines are shown as line features.

Test 3.

Attempted aspect, elevation, and slope reconstruction from two concentric circles of diameter approximately 200 and 23 pixels across. The largest circle was defined as having an elevation of 50 metres and the smallest circle an elevation of 500 metres.

Figure 7.18: Interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.
This elevation profile differs from the previous reconstructions in a number of ways:
* There is only one high patch located outside the circular contours;
* The directional failure appears to have been supplemented with some other sort of failure of unknown origin. This particular failure has manifested itself by not coding the majority of the section bounded between the large and small circles. The area not coded has approximate circular symmetry; and
* Apart from a few uncoded pixels, the top of the shape has been coded satisfactorily as 450-500 metres; this satisfies the restrictions imposed by the 500 metre contour line.

Figure 7.19: Close up of circular section of interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.
Figure 7.20: Interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.

This figure indicates a discontinuous aspect profile and the northerly aspect is absent. On straight lines, such as those demonstrated, the algorithm has somewhat arbitrarily and unreasonably reconstructed some pixels as having an aspect and other pixels as not having an aspect.

Figure 7.21: Close up of circular section of interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 50 and
500 metres respectively. The two circular contour lines are shown as line features.

Figure 7.22: Interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.

The slope profile for this shape does not make any sense and explanations for its appearance are not readily apparent. Additionally, it does not seem reasonable to code a pixel as having a slope when the same pixel has no elevation category associated
with it. This can be easily verified by comparing the elevation and slope reconstruction pictures in Figures 7.18 to 7.23.

Figure 7.23: Close up of circular section of interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.

Test 4:

Attempted aspect, elevation, and slope reconstruction from two concentric circles of diameter approximately 200 and 23 pixels across. The largest circle was defined as having an elevation of 100 metres and the smallest circle an elevation of 1000 metres.

Figure 7.24: Interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.
In this reconstruction two isolated patches of high ground appear and the top of the "hill" has been reconstructed as being concave. At the circumference of the second circle a narrow band of pixels is coded as having 50-100 metre elevation. The yellow (300 to 400 metre) red (450-500 metre) banding effect demonstrated is curious and would probably require a detailed and time consuming examination of the reconstruction code to explain.

Figure 7.25: Close up of circular section of interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.
Figure 7.26: Interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.

It is curious that the aspect data is "almost right", however it lacks any tangible meaning as it does not corresponded with the slope or elevation data. In this reconstruction, pixels were coded as having an aspect but no slope and this is demonstrated in the Figures 7.26 to 7.29.

Figure 7.27: Close up of circular section of interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.
Figure 7.28: Interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.

Figure 7.29: Close up of circular section of interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.
General Analysis Of The Reconstruction Test.

Figures 7.9 through to 7.29 demonstrate directional instability inherent within the E-RMS software package's reconstruction algorithm. Since the digitised contours have "exact" circular symmetry it is apparent that the slope and elevation data which the software calculates depends on the assignment of north in the database. These reconstructions also show that the same contours will be reconstructed differently depending on the location and orientation of the contours within the database. A failure of the magnitude demonstrated suggests that the reconstruction algorithm has never been tested, is entirely consistent with the demonstrated coding failures in the NPWS own database, and is exactly as expected when using a "least squares" approach to reconstruct a landsurface from contours.

Any reconstruction algorithm can be tricked and it is to an extent true that all reconstruction algorithms have some directional instability associated with them. However, to have failed comprehensively a shape that was simple is not a result that one would aim for. Whilst it is not possible to reconstruct objectively from incomplete information, any good reconstruction algorithm would aim to reconstruct consistently. An algorithm that fails a symmetric test, such as the one given, is not reconstructing consistently. The failure to reasonably reconstruct slope and aspect, from elevation information, may be related to problems associated with assigning slope and aspect attributes to grid cells.

The reconstructions produced by this algorithm do not all look similar. Initially it may be presumed that this is because of the non-linearity of trigonometric functions. Without rigidly defining what is meant by "best" and "optimal" it is impossible to select a "best" or "optimal" reconstruction algorithm. Different reconstruction algorithms will exhibit variations in performance depending on the reconstruction being attempted. All the information provided about the "least squares plane" in the software manual is reproduced in the description and picture for Figure 7.1. Without knowing exactly what this reconstruction algorithm does it is not really possible to
say anything more enlightening than what has already been said. Because the algorithm has failed so comprehensively and is attempting to reconstruct by an inappropriate method, any attempt to effect a partial correction is not worth pursuing. A more suitable reconstruction algorithm is required.

**An Adequate Notice Of Data Quality In This Case Would Have To Warn A User That The Data Was Misleading And Wrong.**

All the figures from the E-RMS tutorial database were screen captured directly from screen images produced by the E-RMS Tutorial database at the coordinates shown. Figures 7.3 to 7.7 inclusive serve to demonstrate that the reconstruction algorithm will give absurd results when the data are entered by the makers of the software. The pictures also serve to demonstrate that in the years since this package has been launched, been commercially available, used in universities, government and semi-government authorities this data has not been subjected to sufficient competent scrutiny. If Australian Standard 3960 for product quality was adhered to then situations such as this could not arise. AS 3960 requires that the designer's or user's confidence that reliability and maintainability requirements can be met depends on adequate knowledge of data failures. An adequate notice of data quality in this case would have to warn a user that the data was misleading and wrong.

**Estimating Fire Intensity Using Data Derived from the E-RMS land surface reconstruction algorithm**

- E-RMS Adding A Vector As A Scalar.
- The Fire Intensity Model Described In The E-RMS Software Manual.
- Australian Standards For Quality System Development.
- The E-RMS Fire Intensity Model As A Quality System.

The E-RMS software manual contains two major examples for which the defective land surface reconstruction algorithm is used for fire management. The first occurs in tutorial exercise three and the second as a model developed in the predictive modelling module.

**E-RMS Tutorial Exercise Three: Determining Areas For Which Fire Potential Is High.**

In tutorial exercise three areas of high fire potential are derived by Boolean overlaying vegetation, slope, aspect and fire history. The vegetation categories selected as exhibiting high fire potential are open forest, woodland, woodland/cleared, scrub, lantana, and river terrace. In the second overlay 5 slope categories were selected, 6-10 degrees, 10-20 degrees, 20-40 degrees, 40-70 degrees, over 70 degrees. Four aspect categories; 180-225 degrees, 225-270 degrees, 270-315 degrees, 315-360 degrees were selected. A NOT transformation was applied to the fire history variable. All locations that had not burnt from 1980 until 1988 were selected. The feature so created is named "HIGH FIRE POTENTIAL" and is then saved as a "USER - DEFINED FEATURE."
E-RMS Adding A Vector As A Scalar.

In the above overlay operation the 5 categories of slope over 6 degrees are selected and combined with the four categories of aspect from 180 to 360 degrees. It is assumed that these entities can be added (in the sense of combing) as scalar quantities. This is not true, any addition operation must account for the directionality inherent within the slope and aspect categories; the addition must be vectorial, however, E-RMS does not have the ability to add vector quantities.

The Fire Intensity Model Described In The E-RMS Software Manual.

All predictive models built with the predictive modelling module are decision tree models. The fire intensity decision tree model described in the software manual is reproduced below:

1. IF Vegetation is Rainforest THEN fire potential is LOW.
ELSE 2. IF Vegetation is Woodland or Grassland THEN
2(a) IF slope is less than 3 degrees THEN fire potential is LOW.
ELSE 2(b) IF Slope is 3 - 10 degrees THEN Fire Potential is MODERATE.
ELSE 2(c) IF Slope is greater than 10 degrees then fire potential is high.
ELSE 3. IF vegetation is Open Forest or Heath THEN
3(a) IF Slope is less than 3 degrees THEN Fire Potential is MODERATE
ELSE 3(b) IF Slope is greater than 3 degrees THEN fire Potential is HIGH.

Any such decision tree model consists of a collection statements. When the instruction is given for the model to "run", the collection of statements is applied to every grid cell in the database. For example, the above model would examine the attribute string for each grid cell in the database and determine that the fire potential was low if rainforest vegetation was present. A modelled variable in E-RMS resembles all other variables in that it contains a number of categories. For example, the categories of the fire potential model are low, moderate, and high. Once such a model has been run each pixel in the database is coded as having categories of the model present or absent. In the above model each grid cell would be coded as having a low, moderate, or high, fire potential. A modelled variable can be displayed and analysed using the map display analysis module and can "be used as a branching variable in further models" (E-RMS Software Manual).

Australian Standards For Quality System Development.

Australian Standards 3901--1987, 3902--1987, 3903--1987, and 3904--1987 deal with quality systems in general. These standards for quality systems point out that in the case where data analysis is required the properties of the data should be firmly established. It is essential that the sophistication of the analysis should never be more than that permissible from the actual information present.

The E-RMS Fire Intensity Model As A Quality System.

The maximum possible change in height that can occur over 1 grid cell and still be classified by the reconstruction algorithm as being less than 3 degrees is not known,
however, because a 10 metre contour interval is the best that could have been used to construct the database, it is known that the minimum potential uncertainty in the first statable category of slope is at least 5.7 degrees ($\tan^{-1}(10/100)$). Given the insensitivity of this database an attempt to run a fire intensity model is clearly overly ambitious. To run a fire intensity model that uses categories of slope that the reconstruction algorithm can not distinguish between, relies on a reconstruction that will not detect a cliff of 274 metres, adds vector quantities as scalars and will code a rise of 100 metres over a horizontally projected distance of 100 metres as having a "slope" of 3 to 6 degrees is to invite most undesirable consequences. It is important to make the point that running a fire modelling program off a database as insensitive and inaccurate as this may be no more reliable than drawing conclusions from a random map.4

1 A diameter of 200 pixels corresponds to 2000 metres at 1:10,000 scale.

2 A diameter of 23 pixels corresponds to 230 metres at 1:10,000 scale.

3 Directional instability arises from the fact that it is impossible to distribute a finite collection of symbols throughout a Euclidean Space having dimension greater than 1 in such a way that the rate at which symbols are encountered is constant for all trajectories through the space.

4 A decision as to whether or not a map is significantly better than a random map may be based on a correlation coefficient between what the map claims is present at particular locations and what is actually there.
Chapter Eight: General Discussion & Recommendations.

- Outline of Chapter
- Verifying Important Observations
- Addressing Issues Raised Concerning E-RMS

Outline of Chapter

This Chapter contains a summary of problems identified concerning the failings of the E-RMS geographic information system package and presents recommendations for corrective action. Since this package is used for the management of National Parks in New South Wales and underpins some of the States' most important environmental databases it is necessary to proceed cautiously with corrective action.

Verifying Important Observations

The first step that should be taken before any corrective action is undertaken is for the University of Western Sydney Hawkesbury to ascertain the factual accuracy of the information contained within this document, for two reasons. Firstly, any reasonable assessment of the value of this work would necessarily address, at least in part, the analytical correctness of the claims contained in this document. Secondly, there is the potential to do considerable damage to existing E-RMS environmental databases if my claims of analytical failings are incorrect and action is taken on the basis of incorrect information. A summary of some of the most important observations contained within this document follows.

Observations 1 to 3 are concerned with analysing interactions between biological organisms and coded properties, using databases that specify the distribution of properties at geographical locations as described in Chapter 2.

1. In the case where the property \( p_i \) is fatal for a particular organism and the absence of the property \( p_i \) at a location element implies the absence of property \( p_k \), it is impossible, from the database, to say anything about the potential or actual interaction between the given organism and the property \( p_k \).

2. It follows immediately from the observation that some organisms require properties that are fatal for other organisms that, in order to write an algorithm for examining interactions (potential or actual) between a biological organism and coded properties, it is necessary to have knowledge of both the organisms physiology and the distribution of fatal properties. It is not generally possible to use a stochastic process operating over the set of location elements to test a hypotheses regarding the effect of a property \( p_k \) on an organism and therefore one has to know how the particular organism will respond to coded properties prior to writing the algorithm for determining interactions between the organism and coded properties.

3. The comprehensive analysis contained within Chapter 2 of the thesis establishes that it is impossible to write a general algorithm for answering well
posed questions concerning interactions between biological organisms and coded properties.

Observations 4 to 7 are concerned with the location coordinates required by E-RMS databases and the precision specified for the E-RMS Apsley Macleay tutorial database.

4. The AMG coordinates stated for the E-RMS tutorial database do not exist.

5. The E-RMS Database Specification Module does not permit the entry of a grid zone and therefore it is not possible to enter AMG coordinates into the E-RMS software package.

6. As a consequence of the inability to enter correct AMG coordinates into E-RMS databases E-RMS can not be correctly extended for national data coverages using AMG coordinates.

7. The accuracy required to report information at the level of precision specified for the E-RMS tutorial database will not be achieved in practice. The reasons for this claim are as follows:
   
   o The claim that it is possible to locate the projection of a location into the correct pixel for all locations described by the E-RMS Apsley Macleay tutorial database requires that it be possible to state the horizontal projection of a location with a certainty of better than 1 part in 1000 along the diagonal of the database;

   o The problems of stating location dealt with in chapter 4 of the thesis;

   o The errors inherent in hand digitising data from topographic maps described in chapter 4 of the thesis;

   o The observation that the AMG coordinates stated for E-RMS databases do not exist; and

   o The fact that this software package can not readjust maps to compensate for deviations in magnetic and grid north.

Observations 8 to 17 arise from examination of incorrect mathematical analysis methods adopted in the E-RMS software manual; as described in chapter 6 of this thesis.

8. E-RMS can not be used as a stand alone GIS and this is because the E-RMS package has no provision, within the software, to state the origin of source documents, or to describe the procedures used to aggregate or classify the data.

9. It is incorrect to speak, as the E-RMS software manual does, of a species of tree occurring "at random in relation to elevation." The distribution of an entity may occur at random over a domain; an entity can not correctly be said to be distributed at random with respect to some other entity.

10. Ultimately to have success doing any sort of inductive statistical analysis it is essential to have some method of knowing the distribution of fatal attributes
for an organism, applying an appropriate filtration and examining the family or families of equivalence classes defined by the coded properties. The E-RMS software does not have any provision for doing this and the need to do so is not discussed.

11. The discussion contained in Chapter 6 of the thesis establishes that Correlation Analysis is used incorrectly in the E-RMS Software.

12. In the E-RMS software manual a distinction is made between entities described as being "absolute probabilities of occurrence" and "relative likelihoods of occurrence". Chapter 6 establishes that these notions of probability can not be reconciled with that of stochastic regularity or accepted definitions of probability.

13. The E-RMS Software manual contains an incorrect application of Bayes' Theorem based on the introduced notions of an "absolute probability" or "relative likelihood" of occurrence.

14. The software manual mistakenly attempts to apply Bayes' Theorem to a process that is stochastically closed.

15. The application of Bayes Theorem described in the E-RMS software manual mistakenly assumes that organisms are distributed at random over the sampling domain.

16. In the software manual the application of Bayes' Theorem requires the false assumption that the mean of a population approaches that of a sample.

17. The statement made in the E-RMS software manual concerning the construction of "user defined variables" (tutorial exercise 3) that: "User defined categories within this variable can then be used like the categories of any other variable in subsequent analysis." is mistaken. This statement is false because logical operators are used in the construction of user defined variables and it is essential that the restrictions imposed by the logical operators used in the construction of the variable not be contradicted by instructions used in subsequent analysis.

Observations 18 to 33 address the failings of the E-RMS land surface reconstruction algorithm and the concurrent problems of fire intensity modelling using E-RMS databases. These claims arise as a result of the analysis in Chapter 7.

18. From the description of the E-RMS land surface reconstruction algorithm provided in the E-RMS software manual it appears that the reconstruction algorithm falsely assumes the information content of each trajectory at a given location is identical.

19. From the description of the E-RMS land surface reconstruction algorithm provided in the E-RMS software manual it is apparent that the algorithm attempts to apply some sort of a "least squares" averaging process to estimate slope, aspect and elevation from contours.

20. The application of a simple arithmetic average, as executed by the E-RMS land surface reconstruction algorithm, is mistaken. This is because a least
squares approach assumes that the error or variability in measuring the
dependent variable is minimal in comparison to the independent variable and
the assumption of linearity can not be sustained under circumstances where it
is required to reconstruct a land surface from contour lines.

21. The E-RMS land surface reconstruction algorithm specifies the slope and
aspect of a grid cell. Such a specification can not be achieved uniquely and the
E-RMS manual never makes it clear precisely what is meant by such a
description. Furthermore, it is falsely assumed that such a specification can be
achieved uniquely and that slope and aspect quantities can be combined as
scalar quantities.

22. The E-RMS tutorial database codes for reconstructed angles that can not be
measured using the information available to the reconstruction algorithm.

23. The minimum stated category of slope for the E-RMS demonstration database
is 0 to 1 degrees. To code categories of slope as being 0 to 1 degree, 1 to 3
degrees, and 3 to 6 degrees at 100 metre resolution and commercially supply
the package without a specific caveat is to imply that the reconstruction has
distinguished between these categories at the 100 metre level. From the
discussion in chapter 7 of the thesis it is clear that the categories of slope 0-1,
1-3 and 3-6 degrees can not be differentiated.

24. In order to interpolate categories of slope 0 to 1, 1 to 3, and 3 to 6 degrees
from a 10 metre contour interval, over a projected distance of 100 metres, it is
necessary to assign 3 classes under circumstances where it is possible to
measure only 1.

25. When using a 10 metre contour interval it is possible to differentiate a slope of
0 to 1 degree from a slope of 1 to 3 degrees only when sampling at intervals of
573 metres; this is an interpolation over half a kilometre. Coding a slope of 0
to 1 degree under such circumstances involves a linear interpolation beyond
the original information by a factor of more than 5. This can not be done in a
mountain range such as that "described by" the E-RMS tutorial database. The
variance of the terrain is too great to reasonably assign "the slope of a grid
cell" to one of the 3 non-distinguishable classes between 0 and 6 degrees.

26. The E-RMS land surface reconstruction algorithm failed to detect a single cliff
in 520,000 of hectares mountainous territory (Figure 7.3).

27. When adopting a linear reconstruction (a straight line as implied by coding a
constant angle over a fixed distance) over 100 metres the minimum sized cliff
that can be detected allowing concessions that are both computationally and
theoretically impossible is, as stated in chapter 7, approximately 275 metres
(almost 1000 feet) and therefore it is not surprising that no cliffs have been
detected.

28. The E-RMS land surface reconstruction algorithm produces contradictory
results. Specific examples of which are shown in chapter 7. These examples
include:

- Incompatible slope elevation and aspect data occurring along a 500
  metre front (Figure 7.5);
A 100 metre by 100 metre representative pixel determined to have a "slope" of 3 to 6 degrees when it is apparent that the elevation change coded across the pixel is 100 metres (Figure 7.6); and

- Square kilometres of land coded as having a slope of zero to one degree and various categories of aspect (Figure 7.7).

29. Figures 7.9 to 7.29 demonstrate that the slope and elevation data calculated by the software depends on the designated direction of north and that the same contours will be reconstructed differently depending on their location and orientation within the database.

30. A failure of the magnitude demonstrated suggests that the reconstruction algorithm has never been tested, is entirely consistent with the demonstrated coding failures in the NPWS own database, and is exactly as expected when using a "least squares" approach to reconstruct a land surface from contours.

31. Australian Standard 3960 requires that the designer's or user's confidence that reliability and maintainability requirements can be met depends on adequate knowledge of data failures. An adequate notice of data quality in this case would have to warn a user that the data was misleading and wrong.

32. E-RMS adds Vector quantities as Scalars. A specific example of this occurs in tutorial exercise three where areas having "high fire potential" are derived by Boolean overlaying vegetation, slope, aspect and fire history.

33. Running a fire modelling program off a database as insensitive and inaccurate as the E-RMS Apsley Macleay database may be no more reliable than drawing conclusions from a random map and relying on a collection of categories that, in general, do not exist.

**Addressing Issues Raised Concerning E-RMS**

GIS construction (the most important stage) is almost always subject to deadlines and accuracy can not be accomplished with a rushed job. Quality may be sacrificed at the very time when it is most important. If people who do not know exactly what they are doing build and or manipulate GIS databases then it is inevitable that false conclusions will be reached and important observations will be obscured. For GIS, accuracy at completion depends on accuracy at commencement and adherence to standards throughout the entire production and user process. Failure to consider the usefulness and extractable nature of information in any GIS database will inevitably lead to problems that may endanger the phenomenon that the GIS is designed to protect. Quite apart from the fact that incorrect data can be of no possible benefit to environmental managers, the inclusion of data obtained from a GIS system that is reporting incorrect data in an Environmental Impact Statement amounts to the tendering of false and misleading evidence.

It is necessary that the problems associated with the E-RMS GIS project/package be comprehensively addressed by people competent to do so. Given the magnitude of the problems outlined above it is essential to recommend that, as required by the Australian Standards for product quality, all developers of experimental models or data of any nature publish errors associated with the data. Unfortunately, as has been
demonstrated, this professional obligation is not necessarily being fulfilled. Information regarding the origin of the data and methods of aggregation used in the compilation of the data is essential if the data is to be subjected to further analysis. To proceed along the path in which the development of the E-RMS package has proceeded is to ignore one of the guiding principles of communication theory; that of the signal to noise ratio (Von-Neumann; 1961): the size of the uncontrollable fluctuations of the mechanism that constitute the noise compared to the significant signals that express the way the entity is behaving. At present the project is promoting substantial map misuse and errors of map misuse are not without consequences (Beard; 1989). I recommend that an enquiry be conducted to investigate:

- The technical feasibility of each project for which the E-RMS package has been used;
- The full extent and consequences of the failings inherent within the package; and
- The suitability of the E-RMS GIS package for the purposes for which it is sold.

Australian Standard AS 3898-1991 Information Processing - Users documentation and cover information for consumer software packages defines a consumer software package as a software product designed and sold to carry out identified functions. The Software and its associated documentation are sold and packaged for sale as a unit. Item 6.3.1 of this standard states that:

"The Purpose of the software and functions it performs shall be described. The extent of its functions shall be clarified, and if necessary, any related functions that it does not perform shall be identified. Where the software consists of more than one program, the purpose and function(s) of each component program shall be described."

To comply with this standard users of the package would have to be informed that:

- E-RMS databases are defined on coordinates that do not exist;
- The land surface reconstruction algorithm produces mistakes of a serious and not fully known magnitude;
- The "statistical analysis" procedures adopted for use in the E-RMS package are based on assumptions that have been shown to be false; and
- It is impossible to write a general algorithm for analysing interactions between physical properties recorded in the database and biological organisms.

---

1 A decision as to whether or not a map is significantly better than a random map may be based on a correlation coefficient between what the map claims is present at particular locations and what is actually there.

2 The evidence is misleading by virtue of the fact that it is assumed that the data is correct.
List of Cited References.

References Chapter 1


- AURISA 91 and AURISA 92. Published in 1991 and 1992 respectively by The Australian Urban and Regional Information Systems Association; Canberra.

- Australian Standard: 3960-1990. Published by The Standards Association of Australia.

- Burrough PA. Five Reasons Why Geographical Information Systems Are Not Being Used Effectively for Land Resource Assessment. Auto Carto 8; volume 2; 1987

- The Macquarie Dictionary. Published by The Macquarie Library Pty Ltd; Macquarie University; Australia; 1991.


References Chapter 2


• Rotman JJ. The Theory of Groups. An Introduction. Published by Allyn and Bacon; Boston; 1965.


• Stevens SS. On the Theory of Scales of Measurement. Science; Volume 103; Number 2684; June 1946.

References Chapter 3


References Chapter 4


References Chapter 5


• Cimon N, Coe PK, and Quigley TM. A Regression Technique for estimating the time required to digitise maps manually. International Journal of Geographic Information Systems; 1990; vol 4; #1; pp 51-54.


• Minutes of the First Meeting of the working party to deal with standard topographic mapping specifications for all scales. 5-7/9/1982 (CMA Bathurst, 1982)

• Taib AK and Trinder JC. LIS/GIS Data Acquisition by Photogrammetry. AURISA 91.
• Tralli DM. A Sense of Where You Are. The Sciences. Published by the New York Academy of Sciences; May/June 1993.


References Chapter 6


• Goode HH. Deferred Decision Theory. In Recent Developments in Information and Decision Processing. Edited by RE Machol & P Gray. Published by Macmillian; 1962; New York.


• Kenkel JL. Introductory Statistics for Management and Economics. Published by PWS Publishers; Massachusetts; 1984.


References Chapter 7

• Anon; 1982. The story of the Sydney Harbour Bridge. Publication released on the Fiftieth Anniversary of the Opening of The Bridge.


• Hill RO. Elementary Linear Algebra. Published by Academic Press; Orlando, Florida; 1986.


• Resnikoff H. The Illusion of Reality. Published by Springer-Verlag; New York; 1989.


References Chapter 8

• Australian Standard: 3898-1991. Published by The Standards Association of Australia.


List of Figures and Tables

Figures & Tables Chapter 2

List Of Figures For Chapter 2.

• Figure 2.1: The map \( f:W \rightarrow \Theta \) takes locations in the real world and maps them to particular location elements addressable to the database.

• Figure 2.2: The characteristic function \( g: P \rightarrow \Omega \) that constructs \( \Omega \), the finite set of all distinct property codings, from \( P \), the finite set of available properties.

• Figure 2.3: The injective function \( \psi: \Theta \rightarrow \Omega \) that maps every member of the finite set of location elements to a member of \( \Omega \), the set of all distinct property codings.
- Figure 2.4: The injective functions $\bullet: \oplus \rightarrow \alpha$ and $\bullet: \ominus \rightarrow \beta$ that map every member of the finite set of location elements $\oplus$, to a member of $\alpha$ and $\beta$.

List Of Tables For Chapter 2.
- Table 2.1: The scales of measurement (Birkhoff in Stevens; 1946).

**Figures & Tables Chapter 3**

List Of Figures For Chapter 3.
- Figure 3.1. The description of location elements within an E-RMS database. (E-RMS software manual).
- Figure 3.2: Digitised information is stored in coverages (E-RMS Software Manual).
- Figure 3.3: Map showing areas burnt since 1980 and distribution of rainforest, within the area described by the tutorial database.

List Of Tables For Chapter 3.
- Table 3.1: Entries in E-RMS Tutorial Database.

**Figures & Tables Chapter 4**

List of Tables For Chapter 4.
- Table 4.1: Coordinates for "The Lion" geodetic station.
- Table 4.2: AMG Easting and Northing Coordinates Stated for E-RMS Tutorial Database.

**Figures & Tables Chapter 5**

List of Tables For Chapter 5.
- Table 5.1: Minimum mapping unit for maps of various scales.
- Table 5.2: Entries in E-RMS Tutorial Database (E-RMS software manual).

**Figures & Tables Chapter 6**

List of Figures For Chapter 6.
- Figure 6.1: Ticks denote the recorded presence of an item of interest and crosses denote its recorded absence at surveyed sites scattered through out the area described by the database.

**Figures & Tables Chapter 7**

List of Figures For Chapter 7.
- Figure 7.1: Reproduction of picture describing E-RMS reconstruction algorithm. A "least squares" plane is fitted to the elevation data for the eight points demonstrated. The plane is then used to calculate the elevation, slope and aspect at the centre of the grid cell (E-RMS Software Manual).
- Figure 7.2: Sampling contours in different directions.
• Figure 7.3: Locations the tutorial database has coded as having a slope over 70 degrees (cliffs).

• Figure 7.4: Locations the tutorial database has coded as having 40-70 degree slope.

• Figure 7.5: Close up of E-RMS tutorial database slope reconstruction and elevation.

• Figure 7.6: Close up of E-RMS tutorial database slope reconstruction and elevation.

• Figure 7.7: E-RMS tutorial database containing flat land with aspect.

• Figure 7.8: This figure contains circular contours digitised from the circles inscribed on the A3 sheet of paper used to carry out the symmetric tests on the E-RMS land surface reconstruction algorithm.

• Figure 7.9: Interpolated elevation data. Calculated when the large and small concentric circles were defined as having an elevation of 1 and 10 metres respectively. Figure 7.10: Close up of circular section of interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 1 and 10 metres respectively. The two circular contour lines are shown as line features.

• Figure 7.10: Close up of circular section of interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 1 and 10 metres respectively. The two circular contour lines are shown as line features.

• Figure 7.11: Interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 1 and 10 metres respectively. The two circular contour lines are shown as line features.

• Figure 7.12: Interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 1 and 10 metres respectively. The two circular contour lines are shown as line features.

• Figure 7.13: Interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 10 and 100 metres respectively. The two circular contour lines are shown as line features.

• Figure 7.14: Close up of circular section of interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 10 and 100 metres respectively. The two circular contour lines are shown as black line features.

• Figure 7.15: Interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 10 and 100 metres respectively. The two circular contour lines are shown as line features.

• Figure 7.16: Interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 10 and 100 metres respectively. The two circular contour lines are shown as line features.
- Figure 7.17: Close up of circular section of interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 10 and 100 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.18: Interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.19: Close up of circular section of interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.20: Interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.21: Close up of circular section of interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.22: Interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.23: Close up of circular section of interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 50 and 500 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.24: Interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.25: Close up of circular section of interpolated elevation data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.26: Interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.27: Close up of circular section of interpolated aspect data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.

- Figure 7.28: Interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.
• Figure 7.29: Close up of circular section of interpolated slope data calculated when the large and small concentric circles were defined as having an elevation of 100 and 1000 metres respectively. The two circular contour lines are shown as line features.

List of Tables For Chapter 7.

• Table 1: Some values of n, with their corresponding change in height over 100 metres and corresponding angle.
• Table 2: Entries in Database Specification Module for UWSH Campus.

Meta Data Record.

<!--META NAME="Author" CONTENT="Bryan Hall">
<TITLE>Bryan Hall Master Thesis Introduction</TITLE>
<META NAME="DC.Title" CONTENT="Environmental Mapping Systems - Locationally Linked Databases">
<META NAME="DC.Creator" CONTENT="Bryan Hall">
<META NAME="DC.Type" CONTENT="text">
<META NAME="DC.Date" CONTENT="1994">
<META NAME="DC.Format" CONTENT="text/html">
<META NAME="DC.Coverage" CONTENT="Australia">
<META NAME="DC.Description" CONTENT="Substantial resources have been devoted to developing computer based data systems, because it is believed that these data systems will lead to greater understanding of geographical systems and the possibility of real time mathematical modeling of real or hypothetical scenarios. ... It is unreasonable to draw inferences from the displays created by computer software packages without first having a rigorous understanding of the errors inherent in the system and the importance of precision and accuracy to the outcome of the questions for which answers are sought."

<META NAME="DC.Relation" CONTENT="University of Western Sydney Masters Thesis - Submitted 1994">
<META NAME="DC.Source" CONTENT="...">
<META NAME="DC.Subject" CONTENT="GIS, Geographical Information Systems, Spatial Databases, Maps, Mapping, Statistics, Mathematical Analysis">
<META NAME="DC.Publisher" CONTENT="Published Electronically on au.riversinfo.org by Bryan Hall">
<META NAME="DC.Rights" CONTENT="Copyright (©) Bryan Hall">
<META NAME="DC.Identifier" CONTENT="A Review Of The Environmental Resource Mapping System and A Proof That It Is Impossible To Write A General Algorithm For Analysing Interactions Between Organisms Distributed At Locations Described By A Locationally Linked Database and Physical Properties Recorded Within The Database. 1994"/>
<META NAME="DC.Language" CONTENT="en"/>
A paper summarising aspects of the information contained within this thesis was subjected to independent review by three impartial referees and published in the Australian Journal of Soil and Water Conservation (Volume 1 number 9 February 1996).