Infometric and statistical diagnostics to support an intelligent approach to Interpolation

Dr Claire Jarvis; Dr Neil Stuart; Mr William Cooper

Department of Geography, University of Edinburgh

chj@geo.ed.ac.uk

Abstract. Application specialists who use GIS commonly have high-level knowledge of their wider task, and low-level knowledge of the specific system commands as given in help systems or user manuals. However these users may not necessarily have the intermediate knowledge that experts in GI Science have gained from working with GI systems over several years.

Focusing on interpolation functions, this paper explores the design of an 'intelligent' module that sits between task and GIS. To help users gain the necessary knowledge to complete their task and minimise the possibility of methodological error, we demonstrate a method that combines knowledge from multiple sources. We observe that *both* infometric (or cognitive) knowledge and statistical knowledge are required to find a solution that jointly meets the requirements of a particular user and data set. Our approach to providing this mix of knowledge is to construct a network of rules that assist the user to select an appropriate interpolation method according to the task-related knowledge (or "purpose") of the user and the characteristics of the data. The network triggers exploratory diagnostics that are run on the data sets when a rule requires them to be evaluated.

Following analyses of the data set in relation to the intended purpose, the user is advised which interpolation method might be and should not be considered for the data set. Any parameters required to interpolate the particular data set (e.g. distance decay parameter for Inverse Distance Weighting) are also supplied through in-built optimisation routines. The rationale of the decision process may also be examined, so the 'intelligent interpolator' also acts as a learning tool.

1. INTRODUCTION

1.1 'Intelligent' GIS

Facilitating the appropriate and efficient use of GIS by creating more 'intelligent' GIS to support decision makers has long been identified as a priority for basic research within the environmental modelling community (e.g. Burrough, 1992; Densham & Goodchild, 1989; Fischer & Nijkamp, 1992). This goal is also congruent with the desire to provide 'easy-to-use' spatial analysis functions expressed during the late 1980s. Anselin (1989, p14-15) for example argued that "*With the vast power of a user friendly GIS increasingly in the hands of the non specialist, the danger that the wrong kind of spatial statistics will become the accepted practice is great*". Ten years on, it is useful to reflect on the advances towards 'intelligence' in GIS, to question whether the original intentions remain desirable and, if so, what research remains to be carried out.

The last decade has seen several methods developed as part of research into artificial intelligence becoming used by researchers into GIScience. Followers of the GeoComputation series will be familiar with the introduction of expert systems (e.g. Leung & Leung, 1993), neural networks (e.g. Fischer & Reismann, 1999; Rigol *et al*., 2001), fuzzy logic (e.g. Stefanakis *et al*., 1999), artificial life and cellular automata (e.g. Câmara *et al*., 1996), genetic algorithms (e.g. Murnion & Carver, 1996) and more recently genetic programming and autonomous agents (e.g. MacGill *et al*., 1999; Westervelt & Hopkins, 1999). Broader establishment of some of these techniques in geography is marked by books by Openshaw (1997) and Hewitson (1994) writing from quantitative human geography and physical geography perspectives respectively. In general, we would argue that most of this research has viewed AI methods as *alternatives* to existing statistical tools. AI techniques have been shown as viable alternatives to methods such as maximum likelihood classification (e.g. Jarvis, 1996), parametric regression (e.g. Cheesman & Petch, 1999) and specific process models (e.g. Dawson & Wilby, 2001). It is much less common however to see AI methods being used to encapsulate the broader domain knowledge of a GIScientist who uses this knowledge to select tools from an existing, viable, collection although exceptions may be found (e.g. Morse, 1987; Zhu, 1996; Zhu & Healey, 1992). This leads us to argue that we presently have artificial intelligence *in* GIS, but not artificially intelligent GIS, because of the low level at which AI methods are being deployed within the GIS toolbox rather than as a wrapper around the tools.

1.2 Incorporating 'intelligence' within GIS

Application specialists who use GIS commonly have high level knowledge of their wider task, and low-level knowledge of the specific system commands typically provided by help systems or user manuals. However

these users may not necessarily have the *intermediate* knowledge (Bhavnani & John, 2000) that experts in GI Science have gained from working with GI systems over several years. This intermediate knowledge can be vital for ensuring that appropriate analysis is conducted (e.g. Anselin, 1989). Indeed, it has been suggested that one of the main goals identified for spatial DSS over the last decade is to make GIS tools available to users with different levels of expertise (Cowen & Shirley, 1991).

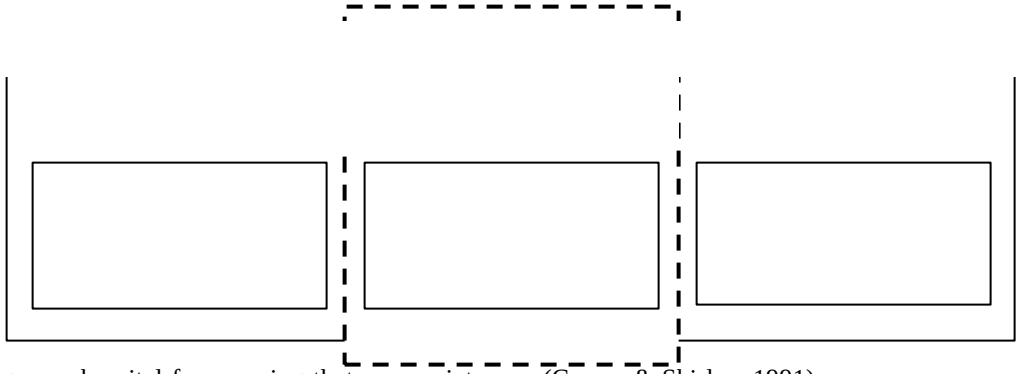


Figure 1: Knowledge and reasoning required ('Intermediate knowledge' concept after Bhavnani and John (2000))

Intermediate knowledge is acquired from 'rules of thumb' formalised and refined through the experience of processing real data sets on a case-by-case basis; this is the intelligence required of an expert GIS user that an AI module would be designed to provide. An example use of intermediate knowledge is when an expert GIScientist chooses a specific method of analysis. This can involve mediating between the desirability of using different techniques and the observed characteristics of the data. For instance, an expert uses intermediate knowledge when considering whether a data set encoded using a particular data model is suitable for undertaking a viewshed analysis or whether one should first transform the data to a different encoding, bearing in mind inaccuracies that might occur from this preparatory process.

When designing methods that capture and use intermediate knowledge, several issues need to be addressed. These include:

- When intermediate knowledge is being used to assist in selecting appropriate methods of analysis, how should one balance between knowledge that refers to the 'general best practice' with specific knowledge obtained from a diagnostic analysis of the particular data set in use (cognitive versus statistical knowledge)?
- Is it preferable to select methods according the wider purpose and domain of the analysis, or should choice be based mainly on the results of quantitative analysis?
- Where should intermediate knowledge be stored and accessed? For example, should this knowledge be task specific, associated with specific GIS functions, 'types' of analysis or stored as meta-data?
- To what extent should a user be aware of the internal decision support processes. Should this be hidden, or at what level of detail and at what stages in the analysis should there be interaction with the AI module?

1.2.1 Should methods be selected according to purpose and domain, or the characteristics of the data?

With the emergence of faster computers, a shift may be identified in more recent years towards the use of process-intensive data manipulation techniques in GIS, rather than knowledge-based methods. We argue however that this computationally extensive approach, advocated for example by Burrough (1992), fails to tap into higher processes of cognition that an expert would employ. In regard to Figure 1, this relates both to the expertise in the application domain and to specialist 'intermediate' style knowledge.

While arguments exist for the use of inductive, data focused methods, these are least strong where a body of established, theoretical knowledge exists or there is a base of empirical knowledge based on experimental evidence. As Openshaw and Alvanides (1999) observe in regard to spatial analysis methods, '*The ultimate aim is to develop an intelligent partnership between user and machine, a relationship which currently lacks balance.*' Andrienko and Andrienko (1997) too make the important point, in the context of intelligent data visualisation, that even where data-model approaches may be able to take into account the *characteristics* of the data they fail to consider whether the given interpretation meets the *objectives* of the particular task.

1.2.2 Should intermediate knowledge be associated with GIS functions, or specified as meta-data?

With the intention of providing a measure of intelligence to GIS, 'appropriate' methods for analysing particular data sets have been encapsulated within their accompanying meta-data structures (e.g. Stefanakis *et al*., 1999; Vckovski & Bucher, 1996). The argument has been that implementing a particular pre-defined model of the field under consideration both saves analysis time and maintains consistency of use. Vckovski and Bucher for example associated interpolation methods with data sets in the from of meta-data, suggesting that their "*VDS* [Virtual data sets] *can serve for many applications without need for conversions and transformations.*"

We argue that a simplistic binding of 'appropriate method' to data is fraught with problems, for two reasons. Firstly, data are regularly aggregated, or partitioned, altering their characteristics considerably in a manner that has the potential to render such encapsulated methods inappropriate. Since the same data set can be reused for many different purposes, and the appropriateness of different spatial analysis methods varies according to purpose, metadata needs to be re-evaluated for each occasion. Encoding intermediate intelligence as meta-data leaves little scope for targeting a method according to purpose and domain.

Secondly, early object oriented research discovered considerable difficulties when attempting to streamline ontologies for geographical objects. For example, while one person may define a hydrological ontology based on the watershed, another may conceptualise hydrological thinking in terms of drainage basins. Rather, as Fonseca and Egenhofer (1999) note, for improved inter-operability object profiles need to be constructed dynamically such that multiple ontologies can be mapped to system classes. In the case of intermediate knowledge, ontological complexity similarly exists when encoding methods of spatial analysis as meta-data, as research expressing spatial queries in natural language demonstrates (e.g. Shariff *et al*., 1998; Wang, 2000).

1.2.3 Should a user be aware the decision making process, or should this be hidden?

Rather than *hiding* the complexity of GIS methods from the user, as has previously been suggested of intelligent GIS, we advocate a shift towards *supporting* the client to use GIS software. As with later decision support systems, it treats the *user* as the final decision maker, acknowledging that superior thinking and intuition are not, as yet, realistic components of a GIS. A philosophical shift within the design of proprietary GIS has slowly followed where lower level functionality and algorithms are now becoming more accessible, if not yet entirely transparent (Sondheim *et al*., 1999).

Pragmatically, this 'supportive' philosophy allows the development of a number of function-related modules that are not bound to a particular GI system but, rather, can be used in support of many. This module-by-module approach also reflects the fact that artificial intelligence works best where the domain of the application can be tightly specified, in highly domain specific contexts.

Exploring the means by which the knowledge of GIScience specialists (intermediate knowledge) can be captured and used to assist applied users of GIS forms the rationale for this paper. The requirement to have a confined domain for AI methods encapsulating knowledge suggests that the example task should be focused upon a particular class of GIS functions. The problem of choosing appropriate methods for interpolating point-based observations of environmental variables to create continuous representations of environmental gradients of the required accuracy for modelling can be particularly time consuming and depends greatly on the data set. For many applied users, interpolation is a preliminary task, not the sole purpose of their analysis using GIS. Given this situation, we explore the advances necessary to assist such users through the development of an 'intelligent' module that sits loosely between task and tool, using interpolation as an example class of function. That is, we express intermediate knowledge in relation to the functionality of a GIS rather than as part of a data model. This approach assumes that a user is more able to specify, in broad terms, what they wish the GIS to do to achieve their purpose. The properties of the actual data set may mediate the final selection of a method, and the user can be assisted by querying the GIS to explore and to understand this. In order to re-create the flexibility of a human expert, the module draws initially and influentially on the cognitive knowledge of the user and a theoretical rule base, but balances this with hidden, low-level and supportive statistical analyses (Figure 2).

Figure 2. Elements of knowledge: from cognitive to statistical

2. THE TASK OF INTERPOLATION

To focus our analysis, we examine the common scenario of a GIS user who wishes to construct continuous surfaces from scattered point observations. The fact that there is no universally 'best' interpolation method poses a potential problem, despite a number of useful overviews of different methods (e.g. Burrough & McDonnell, 1998; Lam, 1983; Mitás & Mitásovà, 1999). This situation arises since the 'best' method varies according to the purpose envisaged of the interpolated surface and the unique characteristics of a particular data set. Moreover, the largely separate development of kriging and splining techniques has resulted in a dispersed and poorly connected literature base.

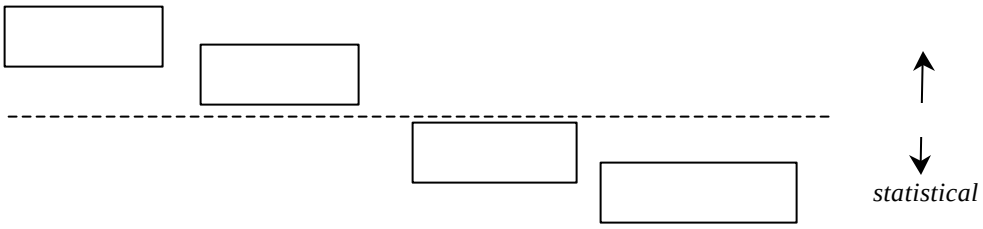
Table 1: Interpolation with 'intelligence': previous research

Reference	Rule-led	Statistics-led	Domain	Interpolation methods	Status
Bucher (1998)		Yes			Design
Bucher & Vcovski (1995),				Kriging family	
(Maslyn, 1987)	Yes		Mining/ geology		Implementation
The thesis? Ref ...	Yes		Geological contouring	Trend surface analysis, kriging family, smoothing splines, triangulation, distance weighted average	Design
Dimitrakopoulos (Dimitrakopoulos, 1993)	Yes		Mining/ geology	Kriging family	Design

Expert users know that poor results arise for example where ordinary kriging is applied with highly skewed or bimodal data or when a variogram cannot be modelled realistically. Similarly, the use of high order trend surfaces in x and y is undesirable where the trend can more simply be eliminated using collateral data such as elevation. Choosing an appropriate interpolator is important, both for maintaining visual realism and as part of a strategy to control error. Inaccuracies arising through the use of a less appropriate interpolator may subsequently propagate throughout subsequent modelling applications (e.g. Burrough, 1992; Heuvelink, 1998; Jarvis & Stuart, 2001c).

Incorporating intelligence within interpolation has attracted some limited attention over the past fifteen years (Table 1). The earliest work (Dutton-Marion, 1988; Maslyn, 1987) incorporated a rule–based rationale for the choice of method. Necessarily therefore, the domain of interest required focus, in these cases on geological applications (e.g. Dutton-Marion, 1988; Maslyn, 1987) or a sub-class of interpolation functions (e.g. Dimitrakopoulos, 1993). However, the use of rule-based methods alone places arguably too high a burden of interaction and data exploration upon the user. Within Table 1, a shift may be identified in more recent years towards the use of intensive statistics rather than rule-based diagnostic methods to support the choice of interpolator. For example, an 'Extended' Exploratory Data Analysis (EEDA)' method is presented by Bucher (1998) as a standardised procedure to derive the implicit information in the data, where the statistics that incorporate the data characteristics relevant to the selection of the correct interpolation method are formalised into a structured rule-set.

We propose a flexible method of assisted interpolation



which combines the findings of previous researchers, acknowledging that a simple rule based approach (Dimitrakopoulos, 1993; Maslyn, 1987) is computationally efficient for selecting methods within

a specified domain, that exploratory analysis of the data is an essential pre-requisite to making an informed choice (Bucher, 1998) and that the ability to meet a user's accuracy requirements is a further factor that often influences the choice of method (Heuvelink *et al*., 1989). Our approach is one in which users are guided through the choice process, allowing them to express their knowledge of the data and any expected relationships while being helped to answer questions of more complex spatial nature through a facility to interrogate their underlying data only where necessary.

3. MODULE DESIGN

In the design of the 'intelligent' module, a number of our strategies are similar to the principles of scenario-based software design originally intended to assist with human-computer interactions (Carroll, 2000). Some of the properties of scenario-based design relevant here include how scenarios help one to understand and formalise an activity (in this case interpolation), and the creation of an initial, usable tool whilst allowing a process of reflection and learning than can improve the design.

In our case, we evolve a set of initially limited interpolation scenarios where the term 'scenario' captures both the construction of multiple 'views' of a point data interpolation problem domain, and a limited group of methods known to have been applied to this type of data (Section 3.1).

3.1 Interpolation scenario

The most effective work in developing 'intelligent' spatial analysis to date has focused on common purposes such as land evaluation or line generalisation for which there are a well-defined set of choices (e.g. Zhu, 1996). We shall define our scenario as the context of providing continuous estimates of primary environmental data sets, as we believe this topic typifies an area of increasing usage of GIS by a wide range of users whose main specialism is other than GIScience, but who nevertheless require to carry out appropriate and accurate interpolation.

Many models in ecology, agriculture and entomology are primarily driven by meteorological variables such as maximum and minimum temperature, potential evapo-transpiration or rainfall. Such data at the daily temporal scale required are found only at point sources. For an understanding of how the processes being modelled apply throughout the landscape, for example as a precursor to modelling dispersion and movement, interpolation of the input data at a variety of time steps is fundamental. Given the large numbers of users of interpolated temperature data in particular, the intelligent module is being designed with this as the initial focus application.

3.2 Interpolation methods

In the examples we present, the user is assisted in the choice between partial thin plate splines (Hutchinson, 1991), simple, ordinary and universal kriging (Deutsch & Journel, 1992), trend surface analysis that incorporates linear regression and automatic inverse distance weighting as potential interpolators. Encoding a very wide variety of interpolation algorithms was not considered feasible in this first prototype module, necessitating the choice of methods firstly of relevance to the particular chosen scenario but also of strategic importance. The selection of techniques followed a literature review made in another context (Jarvis & Stuart, 2001b) which showed them to be pertinent to the interpolation of temperature data at multiple spatial and temporal scales and locations, and more widely useful within environmental modelling applications (e.g. Hutchinson, 1995; Oliver & Webster, 1990).

3.3 Module overview

To help users gain the necessary knowledge to complete their task and minimise the possibility of methodological error, we demonstrate an adaptive knowledge base that combines knowledge from multiple sources. Observing that both infometric, or cognitive knowledge (Section 3.4.1) and statistical knowledge (Section 3.4.2) are required in finding a solution that meets the requirements of a particular user and data set, our approach merges simple, task related, requests of the user, a rule base and automatic diagnostic data analysis. In this case, the cognitive knowledge has been extracted from the literature and verified by experienced users on the subject of interpolation. This knowledge is supplemented and made specific by requesting data from the user regarding the particular application task. The network triggers exploratory diagnostics *when required of the rule base*. That is, some rules are evaluated using results from subservient statistical diagnostics. The module therefore becomes more than a standard expert system that contains a static knowledge base and inference engine (Durkin, 1984), since the evidence upon which rules are evaluated is updated dynamically.

3.4 Module diagnostics

3.4.1 Infometrics

We use the term infometrics to include both information about the reasoning process and qualitative knowledge about associations in the data (e.g. temperature and elevation are linearly related), both of which are used to structure the knowledge base. An overview of the ideal components of knowledge within an 'intelligent' module may be found within Figure 3.

- **Expectations and uses of the interpolated data**

When considering how the rules should be structured, we place a strong emphasis within the hierarchy on the responses from the user regarding their use for the data. The 'acceptability' of an interpolated surface can differ markedly, depending on whether the surface is intended for use in further modelling, or the purposes of visualisation, as demonstrated by Declercq (1996). Additionally the scientists using the intelligent module are likely to have expectations for their interpolated surface, and potentially an understanding of information that may be used to guide the process of interpolation. For example, prior knowledge of whether a surface is expected to be smooth or rough, whether there is trend in the data or available corroborative data (e.g. linear relationship between elevation and temperature) will often influence which interpolation methods are first investigated. A user may also have preferences as to whether the values of the interpolated surface should always remain within those of the original data, or occasionally exceed them. In order to assist the users to understand the questions asked, we provide visual support relating to the issues raised (Figure 7). Overall, the proportion of questions asked of the user compared to the number of rules evaluated using statistical metrics is intended to be low, but the user's responses will strongly influence the order in which the knowledge-based diagnosis and subsequent statistical analyses will be carried out.

- **Technical rules regarding interpolation methods**

Given the large volume of literature covering both the theory and application of interpolation methods, the expertise to build the inference engine for this project was taken from books, drawings and previous research in addition to human expertise. The use of the literature base is important since expertise is scattered worldwide, and often fragmented in coverage. Few researcher theoreticians for example regularly use both spline and kriging methods, despite their acknowledged similarities (Hutchinson, 1993).

As Figure 3 shows, this body of knowledge may be divided firstly into rules regarding the general characteristics of interpolators (Rule type 1), and secondly the specific assumptions of the interpolators that need to be met by the data set (Rule type 2). For example, splines are generally considered to smooth data (Rule type 1), and the expert user of partial thin plate splines would typically need to establish that the trace diagnostic (Hutchinson & Gessler, 1994) showed that the fit of the data to the spline surface was statistically valid (Rule type 2). Rules falling into these two categories form the bulk of the knowledge within the intelligent module. In general, type 1 rules are evaluated through direct questions to the user, while type 2 rules are evaluated 'behind the scenes'.

For completeness, Figure 3 also refers to a third main group of rules, the case-based/experience based group. Establishing a methodology to encompass the development of these rules is an area for future research.

Figure 3: Infometric diagnostics: components within the knowledge network, by proportion of rules

3.4.2 Statistics

Because of the inherent assumptions behind the many interpolation algorithms, each method performs differently under different data set conditions. As Burrough (1986) noted "*It is unwise to throws one's data into the first available interpolation technique without carefully considering how the results will be affected by the assumptions inherent in the method*". Regardless of which exploratory data technique that is used, "*the time taken to explore, understand, and describe the data set should be amply regarded*" (Isaaks & Srivastava, 1989). Often, a user is unaware of the implicit characteristics of the input data being offered for interpolation.

Subservient to the infometrics, but critical to success, statistical methods trigger only when required by the network to assist with the characterisation of the data set. The results of the statistics can inform the choice of the most suitable interpolation method, and may also be used to carry out the necessary pre-processing of a data set prior to its interpolation.

The set of statistical methods used for diagnosing the appropriateness of otherwise of a data set for certain methods of interpolation have been chosen by analysing the work of experienced GIS users. For example, it is well known that different interpolation methods will perform better or worse depending on the spatial association measurable within the data. As Bucher (1998) identifies, the automatic exploring of the data set for this characteristic is a paramount consideration. Additionally, testing initially for stationarity, and subsequently for spatial association is a principle that goes beyond good geostatistical practice. The identification and where necessary extraction of 'trend' from a data set has also been shown to be a basic practice that significantly improves the accuracy of many interpolation methods (Jarvis & Stuart, 2001b).

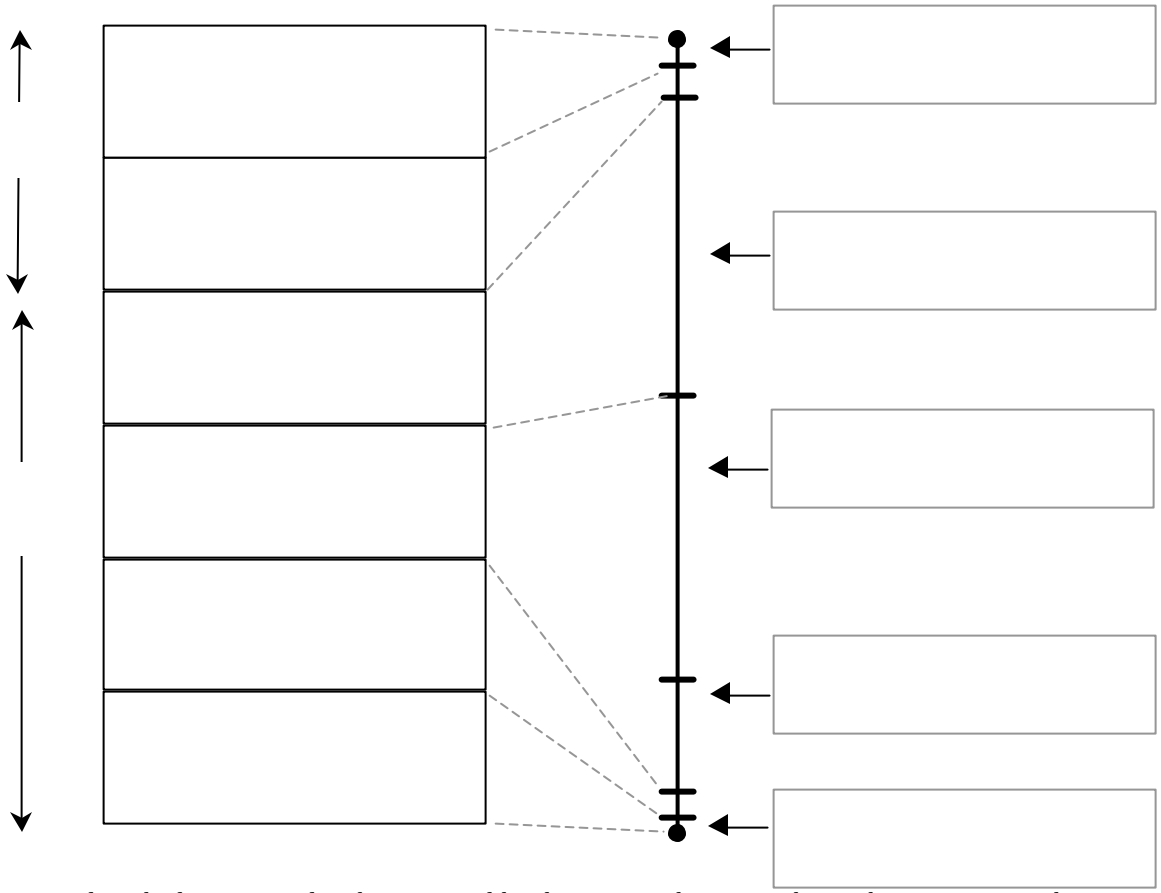


Table 2: Statistical diagnoses fired by rules within knowledge net

Method Evaluated	Associated Statistical Diagnoses Conducted
General	Trend
Kriging	Spatial correlation, normality of data, anisotropy (See Dimitrakopoulos (1993) and (Bucher, 1998) for examples)

Partial thin plate splines	Spatial correlation, linearity of trend, order of derivative of spline, normality of data

Rather than attempt to enumerate or rank all feasible alternatives to the interpolation problem, following the diagnosis the user is advised of the two interpolation methods most likely to satisfy their requirements. Additionally, where an interpolation method is clearly unsatisfactory for the task, this is identified. As Cameron & Abel (1997) note, enumerating and ranking all alternatives to a decision making process is impractical, but providing tangible boundaries to the set of likely solutions provides what may be termed a 'satisficing' result. Once the main type of method has been diagnosed, the system can also assist the user in setting any parameters (e.g. decay parameter for IDW, variogram model and parameters) that may be required for the particular interpolation method (Section 3.5).

3.5 Parameter setting

In addition to the initial phase which the most 'appropriate' methods are diagnosed, further statistics and potentially data preparation will be required to ensure that the subsequent interpolation process is as painless as possible for the user. Not only the choice of method, but also its parameter settings, can be critical if one is to avoid misleading results (e.g. Hodgson, 1993). The majority of GIS leave the selection of parameters entirely to the user.

Table 3: Parameters required of the user for interpolation tasks by proprietary GIS and statistical tools

Method	Associated Parameters Required by Other GIS/Statistical Tools
General	Trend element (e.g. regression parameter, order of xy trend)
Kriging	Variogram model type, variogram parameters, anisotropy, no. of neighbourhood points
Inverse distance weighting	No. of neighbourhood points, decay parameter

3.6 Teaching tool

Additionally, the rationale for the decision process is provided back to the user so that the 'intelligent interpolator' also acts as a learning tool. This should allow users to develop their own knowledge of interpolation for use in future applications.

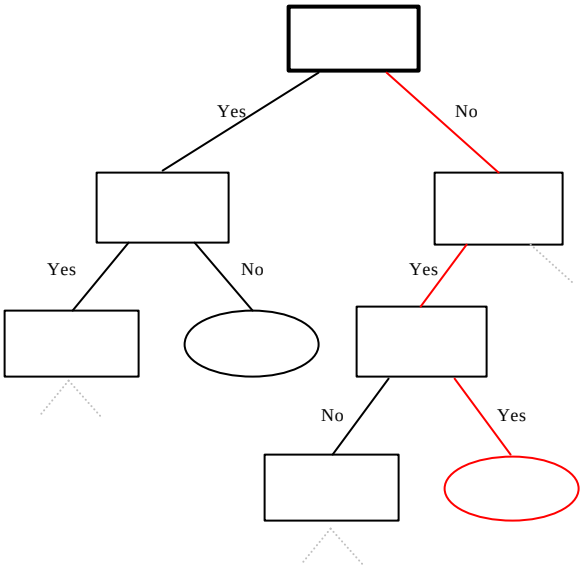
The reporting of rules that have been fired has two roles. Initially it provides a means by which an expert may verify the actions of the intelligent module, and subsequently it allows the inexperienced user to learn the questions that they should be posing and evaluating to make an appropriate choice of method.

4. IMPLEMENTATION OF A PROTOTYPE INTELLIGENT MODULE

We considered that a stand-alone module could better serve users using a variety of the many GIS systems (Arc-Info, Erdas GIS, Arc View, Info-Map, Spans, Genamap, etc.) and statistical packages (e.g. Minitab, SPSS, S-Plus) currently commercially available. This was because:

- Each package has a different subset of interpolation algorithms;
- Each package has different scripting or interfacing capabilities such that a closely coupled module would only serve a subset of users unless re-implemented multiple times;
- There is general lack of heuristic language capabilities in commercial GIS packages, forcing at least some measure of coupling for the implementation process.

To this point, we have considered broadly what is to be produced, and the types of knowledge that will be incorporated. This section reports on the implementation of a prototype 'intelligent interpolation' module, and focuses particularly upon the management of cognitive knowledge and the communication of the diagnostic process to the user.



4.1 Software environment

A combination of Java and the Jess knowledge based system were used to implement the prototype software (Figure 4). *Java* ã from Sun is a platform independent

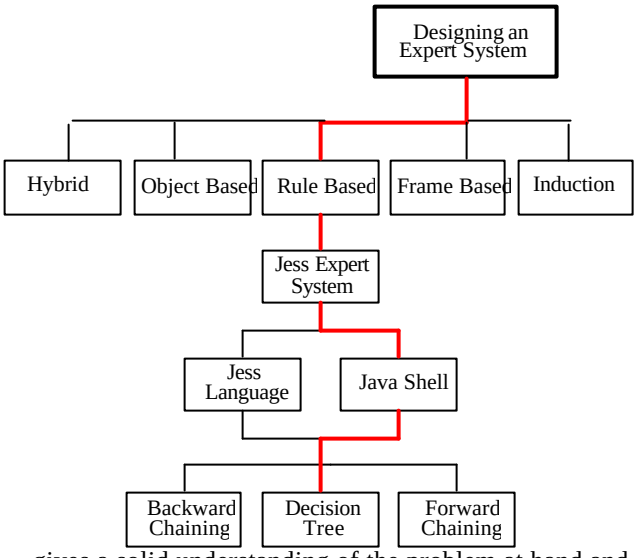
object orientated language, and *Jess ã* from E.J Freidman Hill (Sandia Corporation) is an expert system shell written in Java and based on the widely established *Clips* expert system. Java was chosen based on its suitability for use across different operating systems. Further the Java language was suitable to use for development of an explanation facility, for calling external programs (interpolation algorithms) and for making a graphical user interface (Section 4.3). Jess was chosen for its many features such as being able to accommodate a rule based approach, being able to code the knowledge easily, and having an inference engine.

Figure 4: Selecting suitable software for expressing and controlling knowledge

4.2 Knowledge management

4.2.1 Knowledge Acquisition

We term the approach that we used for knowledge acquisition a 'Phase Teaching' technique, a concept that works on the basis of phased refining of the knowledge gained from interviewing experts and research. The phase teaching method is composed of two phases, the first being a broad and shallow examination and gathering of the intelligent (expert) knowledge to construct a fast prototype. At this stage, broad background reading on interpolation literature



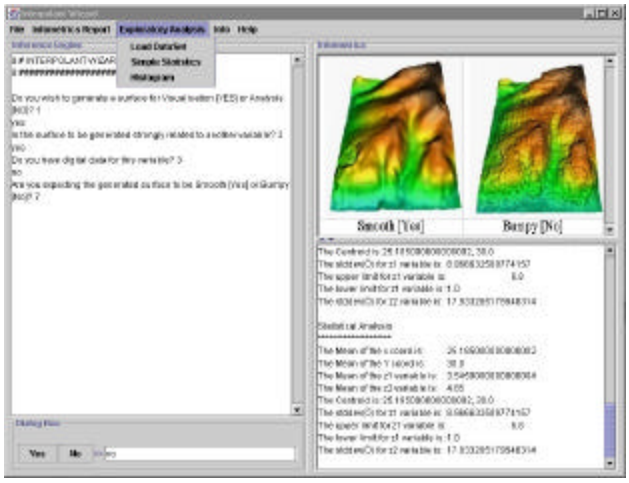
gives a solid understanding of the problem at hand and helps in clarifying the different terminology used in the domain. This is then transferred to the rule base of the intelligent system and a fast prototype developed. The second phase is hands on interaction with the prototype and refinement of the rule base and system, following more detailed reading and further 'in-depth' interviews with experts. Decision trees were used for this knowledge refinement process, based on their ability to express knowledge in a formalism that is often easier to interpret by experts and ordinary users (e.g. Janikow, 1998; Lagacherie & Holmes, 1997; Skidmore *et al*., 1996).

4.2.2 Structuring knowledge

In considering a control strategy by which the knowledge is structured and interpreted, a review of the various tools for building intelligent (expert) systems was necessary (Figure 4). The majority of these are expert system shells, high level programming languages and mixed programming environments. A rule-based system was preferred for this research as it best matches the methodology of deciding an interpolation method, by capturing "the global problem solving approach used by the expert" (Durkin, 1994, p627).

Figure 5. Binary decision tree

As mentioned in Section 4.2.1, decision trees were used for knowledge elucidation and refinement. A decision tree structure, in this case a binary structure,



was also used as a mechanism for knowledge control. This choice was made since decision trees represent a logical reasoning system, which derive sound conclusions from formal declarative knowledge. Additionally, the shortest route to making a decision is always taken. For example, if an intelligent system is represented by many 'rules', then in following a path down the tree, rules that are not relevant to the case in hand are effectively bypassed. Finally, the resulting models are easy to understand, as they can be directly expressed as a set of IF … THEN rules.

A binary decision tree is one where each node of the tree has only two transition branches and are typically used to implement knowledge based on a sequence of yes / no questions (Figure 5). Each decision node has associated with it a question and an answer node an answer (Giarratano, 1998). Inside the nodes, some decision occurs that transfer control via any of its branches to other nodes or leaves. In this type of structure, the inference process starts at the top of the decision tree i.e. the root and follows one and only one of the sub nodes, either a [yes] or [no] branch. The

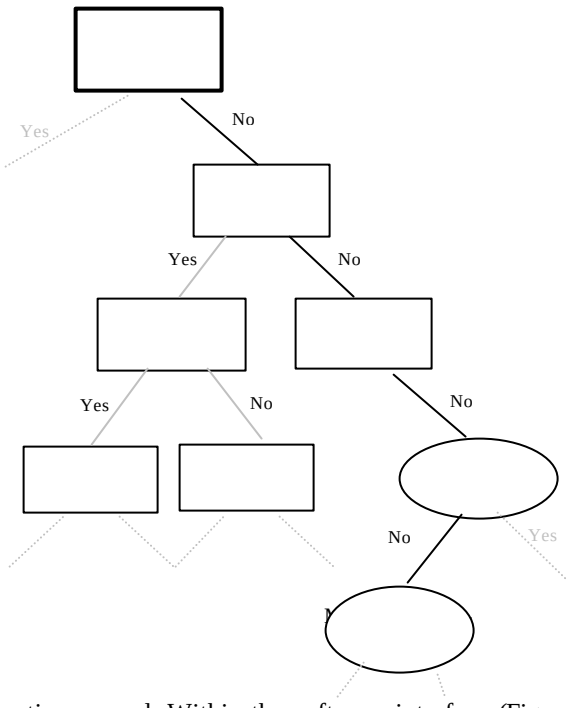
heuristics dictate that one problem leads to consequences that in turn lead to other problems until a conclusion is arrived at an answer node. This allowed us to emphasise the purpose for which the interpolated surface is intended, and the user's knowledge of their domain, by simply placing these rules near the root of the tree.

For example, in Figure 6, the root question is 'Do you wish to interpolate a surface for the purpose of visualisation?' Questions relating to the expected nature of the surface follow, and finally the analytical rules or statistical tests that run directly from the data. In the figure, user-driven rules have been placed in rectangular boxes, and analytical rules/tests that run directly from the data have been placed within oval units.

Figure 6. Example decision nodes for interpolation

4.3 User interaction

Given the growing diversity of user populations and application domains of GIS, different levels of question support and feedback are provided by the module as the diagnostic process progress. Yet, as Egenhofer (1997) notes, a serious disadvantage of textual questioning implicated in Figure 6 '*is that it forces users to translate a spatial image they may have in their minds about the situation they are interested in, into a non-spatial language'*. Multi-modal communication, the use of complementary media to communicate the same idea in parallel cognitive 'channels', was therefore adopted to support users to understand more fully the spatial context of the



questions posed. Within the software interface (Figure 7), the interplay between the left hand text questioning panel and right hand statistical and visual panels follow this concept. The upper area is used to display visualisations that support the user in understanding the rules asked in the left hand pane, on a rule by rule basis.

Figure 7. Prototype user interface

In the example of Figure 7, the images are bound statically with the appropriate rule, and represent abstract notions of the degree of spatial continuity. Other images are used to convey what is meant by 'exact' interpolation, the averaging or extrapolating tendencies of different techniques and the typical shapes of different variogram models. Following ideas from the literature on exploratory data analysis in GIS and geostatistics (e.g. Gunnink & Burrough, 1996; Pannatier, 1996), a further set of dynamic 'exploratory' images are constructed by the software. These are used for example to compare the actual distribution of the user's data with a normal distribution, provide experimental variograms in different directions to assist with the exploration of isotropy, and to allow the verification of modelled variograms against experimental variograms in the case of kriging-related questions. The images are intended to assist the user in responding to particular questions as (or if) they arise, rather than to explore their data interactively. For the less experienced user, these images may be the most important first step in considering the many facets both of spatial data and also the rationale for choosing a particular interpolator.

The lower pane is used to report the results of the statistical tests that the knowledge base carries out and which the user is otherwise unaware of. These results support users with more expertise, and for the most advanced user might form the basis for their judgement in their own right. This pane also reports, at the end, a bounded solution set and suggested parameters (e.g. variogram model, distance decay parameter, order of trend) that the user might apply using stand-alone packages. On request, this pane may subsequently be used to report which rules have been fired and why. This provides information on how decisions are reached by interpolation experts from which the less experienced user may learn.

5. EVALUATION OF THE PROTOTYPE

An initial verification of the intelligent interpolator has been performed to assess the module. While verification has been defined as "*... the process of ensuring that the intelligent system conforms to specifications, and that its knowledge base is consistent and complete within itself*" (Gonzalez & Barr, 2000), we view the process rather more broadly. We considered whether the structure and content of the rules conform to those expected by local experts, an important process since automatic techniques to check if a knowledge base is accurate and complete are in practice not very useful. This evaluation initially

involved checks and refinements of the individual rules, their semantics and ordering by the local experts from whom knowledge was gleaned. Secondly, by using both the decision-tree presentation and through 'hands-on' experience with the prototype, independent experts identified whether the rules and structures implemented were consistent with their experience and minor differences of opinion were resolved. The modular implementation of the rules also enabled those with knowledge of individual domains of interpolation to assess sub-components of the knowledge base and structure. Finally, the prototype was assessed informally by GIS users with little experience with interpolation to determine the level of usability of graphical and semantic aspects of the software that we considered were crucial to its future uptake.

The evaluation of an intelligent module should incorporate both processes of verification and validation. Validation can be regarded as "*the process of ensuring that the output of the intelligent system is equivalent to those of human experts when given the same inputs*" (Gonzalez & Barr, 2000). We identify two main phases within this validation process, the creation of test data sets and an evaluation of the system response to these. The module will be validated using the interpolation of monthly and daily temperature data over Britain, a domain and geographical context familiar to the authors for which exhaustive trials have yielded a good base of empirical, quantitative knowledge on the relative accuracy and visual representativeness that can be achieved using different interpolation techniques (Jarvis & Stuart, 2001a; Jarvis & Stuart, 2001b). Use of Turing tests, where experts validate all or part of the results, was deemed unsuitable for this prototype module since they impose excessive time restraints and responsibility on experts. Rather, we followed Isaaks and Srivastava (1989) in proposing a full comparisons of accuracy of results from each potential interpolation method, assessed using both univariate (e.g. bias, range) statistics and bivariate (e.g. correlation coefficient, RMS error) methods. Jack-knife cross-validation methods were used to generate these relative comparisons of the numerical accuracy of the methods (Efron & Gong, 1983). Visual evaluation of the results was also valuable (e.g. Declercq, 1996).

6. DISCUSSION

The decision to implement the intelligent interpolation module in a combination of Java and Jess languages decreased the time taken to develop a system which can choose an appropriate interpolation method for a given data set. The use of Java also provided a means for distributed the module over the Internet for use by multiple users. The use of a decision-tree method both for knowledge capture and structuring is expected to ease the future expansion of the system to more diverse domains with the development of additional object oriented classes in Java. The modularity of rules, the separation of control from knowledge and the ability provided by Jess to check for consistency will be beneficial when the module begins to be used for further techniques and in other domains. Drawing a decision tree was also found to be a good way to present the information back to the domain expert for verification.

The combination of infometric *and* statistical diagnostics within this knowledge-based module provides a flexibility that goes beyond a typically brittle standard rule base, allowing the capacity for rules to be applied beyond interpolation in one specific application domain. Rather, they are better able to cope with conflict, with different data characteristics and the different needs of diverse users. Meeting diverse application needs is a particular problem in more traditional knowledge–based systems. We have adopted a scenario-based approach when developing the prototype module, both to ensure that the rules are not over-stretched and also to make the system easier to verify.

Rather than *hiding* the complexity of the interpolation methods from the user, as has previously been suggested of intelligent GIS, we advocate a philosophical shift towards *supporting* the client to use GIS software. The rational behind this shift is to instil the important scientific principles of 'know your data', and to support the wisdom of 'parameterising your methods'. Extending the provision of information from statistical and non-interactive visual reports to incorporate interactive ESDA style multiple 'views' of the data as analysed should assist communication with the more experienced user in particular, and enhance their ability to provide an informed response.

The intelligent module, intended to sit between 'tool' and 'task', may be valuable for users of multiple GI systems, therefore extending its potential scope to a wider range of interpolation tasks than would be achievable if a closely coupled design were to have been adopted. Loose coupling also affords other benefits. For example, the wisdom of automatically fitting variogram models meets with some caution within the geostatistical literature (e.g. Deutsch & Journel, 1998; Webster & Oliver, 1990), and the de-coupled design of the module also allows the possibility of interactive comparisons within Variowin (Pannatier, 1996) or similar tools. We also acknowledge that where GI tools are used as components of near-real-time systems, more integrated, focused solutions, that include autonomous or semi-autonomous decision making for specific applications, may be needed (Williams, 1995).

Currently, the intelligent module draws on the first category of knowledge, the theoretical characteristics of different interpolators. The challenge is how to encode application-specific information, which is both qualitative and quantitative. Applied approaches

reported in the literature may be biased by availability of software or the manner in which methods were implemented, and the information reported may be considered 'incomplete' if too complex a matching process is designed.

In summary, while there is no lack of books and papers on the subject of interpolation, there are few consolidated accounts of how and when to apply the guidelines they present. Moreover, the number of statistical analyses that should be undertaken if an interpolation problem is to be approached well is high, and commonly involves multiple software packages. The intelligent interpolator module is an approach to address these problems through the application of infometric and statistical diagnostics that are combined within a knowledge-based reasoning tool. While many examples of artificial intelligence methods applied within particular GIScience applications may now be found (Artificially intelligence in GIScience), we argue that it is rarer to find artificial intelligence used to support new users to perform tasks and avoid pitfalls with basic GIS functionality (Artificial intelligent GIS). The module described here is a prototype solution than can evolve to meet in part this need of the many new users of GIS.

REFERENCES

Andrienko, G.M. and Andrienko, N.V. (1997) "Intelligent cartographic visualization for supporting data exploration in the IRIS system", *Programming and Computer Software*, vol. 23, pp. 268-281.

Anselin, L. (1989) *What is special about spatial data?* NCGIA Technical Report 89/4, National Centre for Geographical Information and Analysis: Santa Barbara

Bhavnani, S.K. and John, B.E. (2000) "The strategic use of complex computer systems", *Human-Computer Interaction*, vol. 15, pp. 107-137.

Bucher, F. (1998) "Using extended exploratory data analysis for the selection of an appropriate interpolation model". *Geographic Information Research: Trans-Atlantic Perspectives*, (ed. by M. Cragglia), pp 391-403, Taylor & Francis: .

Bucher, F. and Vckovski, A. (1995) "Improving the Selection of Appropriate Spatial Interpolation Methods", *Lecture Notes in Computer Science*, vol. 988, pp. 351-364.

Burrough, P.A. (1992) "Development of intelligent geographical information systems", *International Journal of Geographical Information Systems*, vol. 6, pp. 1-11.

Burrough, P.A. and McDonnell, R.A. (1998) *Principles of Geographical Information Systems*, Oxford University Press: Oxford.

Câmara, A.S., Ferreira, F. and Castro, P. (1996) "Spatial simulation modelling". *Spatial analytical perspectives on GIS*, (ed. by M. Fischer, S. H.J. & D. Unwin), pp 201-212, Taylor & Francis: London.

Cameron, M.A. and Abel, D.J. (1997) "A problem model for spatial decision support systems". M.-J. Kraak, M. Molenaar & E.M. Fendel (Eds.), *Proceedings of the Seventh International Symposium on Spatial Data Handling*. Taylor & Francis, Delft: 89-99

Carroll, J.M. (2000) *Making Use: Scenario-Based Design of Human-Computer Interactions*, The MIT Press: Cambridge, M.A.

Cheesman, J. and Petch, J. (1999) "Interpolation of severely non-linear spatial systems with missing data: using kriging and neural networks to model precipitation in upland areas". *Geographic Information Research: Trans-Atlantic perspectives*, (ed. by Massimo & Cragglia), pp 175-188: .

Cowen, D.J. and Shirley, W.L. (1991) "Integrated planning information systems". *Geographical Information Systems*, (ed. by D. Maguire, M.F. Goodchild & D. Rhind), pp 297-310, Longman: London.

Dawson, C.W. and Wilby, R.L. (2001) "Hydrological modelling using artificial neural networks", *Progress in Physical Geography*, vol. 25, pp.

Declercq, F.A.N. (1996) "Interpolation methods for scattered sample data: accuracy, spatial patterns, processing time", *Cartography and Geographical Information Systems*, vol. 23, pp. 128-144.

Densham, P.J. and Goodchild, M.F. (1989) "Spatial decision support systems: a research agenda", *Proceedings of GSI/LIS'89*. ACSM/ASPRS/AAG, Virginia: 707-716

Deutsch, C.V. and Journel, A.G. (1992) *GSLIB Geostatistical Software Library and User's Guide*, Oxford University Press: Oxford.

Deutsch, C.V. and Journel, A.G. (1998) *GSLIB Geostatistical Software Library and User's Guide*, Oxord University Press: Oxford.

Dimitrakopoulos, R. (1993) "Artificially intelligent geostatistics: a framework for accommodating qualitative knowledge-information", *Mathematical Geology*, vol. 25, pp. 261-279.

Durkin, J. (1984) *Expert systems: Design and Development*, Prentice-Hall: New Jersey.

Durkin, J. (1994) *Expert systems: design and development*, Macmillan: New York.

Dutton-Marion, K.E. (1988) *Principles of interpolation procedures in the display and analysis of spatial data: a comparative analysis of conceptual and*

computer contouring. PhD Thesis, University of Calgary: Calgary

Efron, B. and Gong, G. (1983) "A leisurely look at the bootstrap, the jacknife, and cross-validation", *American Statistician*, vol. 37, pp. 36-48.

Egenhofer, M.J. (1997) "Multi-modal spatial querying". M.-J. Kraak, M. Molenaar & E.M. Fendel (Eds.), *Proceedings of the Seventh International Symposium on Spatial Data Handling*. Taylor & Francis, Delft: 785-799

Fischer, M. and Reismann, M. (1999) "Parameter estimation in neural spatial interaction modelling by a derivative free globalization method", *Geocomputation'99*, Virginia: CD-ROM

Fischer, M.M. and Nijkamp, P. (1992) "Geographic information systems and spatial analysis", *Annals of Regional Science*, vol. 26, pp. 3-17.

Fonseca, F.T. and Egenhofer, M.J. (1999) "Ontology-driven geographic information systems". C.B. Medeiros (Ed.), *7th ACM Symposium on Advances in Geographic Information Systems*, Kansas City: http://www.spatial.maine.edu/~max/ACMGIS99.pdf

Giarratano, J.C. (1998) *Expert systems: principles and programming*, PWS: Boston.

Gonzalez, A. and Barr, V. (2000) "Validation and Verification of Intelligent Systems- What are they and how are they Different", *Journal of Experimental and Theoretical Artificial Intelligence*, vol. , pp. 407-420.

Gunnink, J.L. and Burrough, P.A. (1996) "Interactive spatial analysis of soil attribute patterns using exploratory data analysis (EDA) and GIS". *Spatial analytical perspectives on GIS*, (ed. by M. Fischer, S. H.J. & D. Unwin), pp 87-110, Taylor & Francis: London.

Heuvelink, G.B.M. (1998) *Error propagation in environmental modelling*, Taylor and Francis: London.

Heuvelink, G.B.M., Burrough, P.A. and Stein, A. (1989) "Propagation of errors in spatial modelling wth GIS", *International Journal of Geographical Information Systems*, vol. 3, pp. 303-322.

Hewitson, B.C. and Crane, R.G. (1994) *Neural Nets: Applications in Geography*, Kluwer Academic Publishers: Dordrecht.

Hodgson, M.E. (1993) "Sensitivity of spatial interpolation models to parameter variation", *ASPRS-ACSM Annual Convention1992*. ACSM, Alberquerque

Hutchinson, M.F. (1991) "The application of thin plate smoothing splines to continent-wide data assimilation". *Data Assimilation Systems, BMRC Research Report No. 27*, (ed. by J.D. Jasper), pp , Bureau of Meteorology: Melbourne.

Hutchinson, M.F. (1993) "On thin plate splines and kriging". *Computing and Science in Statistics 25*, (ed. by M.E. Tarter & M.D. Lock), pp 55-62, Interface Foundation of North America: Berkley.

Hutchinson, M.F. (1995) "Interpolating mean rainfall using thin plate splines", *International Journal of Geographical Information Systems*, vol. 9, pp. 385-404.

Hutchinson, M.F. and Gessler, P.E. (1994) "Splines - More Than Just a Smooth Interpolator", *Geoderma*, vol. 62, pp. 45-67.

Isaaks, E.H. and Srivastava, R.M. (1989) *An Introduction to Applied Geostatistics*, Oxford University Press: New York.

Janikow, C.Z. (1998) "Fuzzy decision trees: issues and methods", *IEEE Transactions on Systems, Man and Cybernetics Part B. - Cybernetics*, vol. 28, pp. 1-14.

Jarvis, C.H., Stuart, N. (1996) "The sensitivity of a neural network for classifying remotely sensed imagery", *Computers and Geosciences*, vol. 22, pp. 959-967.

Jarvis, C.H. and Stuart, N. (2001a) "A comparison between strategies for interpolating maximum and minimum daily air temperatures: a. The selection of 'guiding' topographic and land cover variables", *Journal of Applied Meteorology*, vol. 40, pp. 1060-1074.

Jarvis, C.H. and Stuart, N. (2001b) "A comparison between strategies for interpolating maximum and minimum daily air temperatures: b. The interaction between number of guiding variables and the type of interpolation method", *Journal of Applied Meteorology*, vol. 40, pp. 1075-1084.

Jarvis, C.H. and Stuart, N. (2001c) "Uncertainties in modelling with time series data: estimating the risk of crop pests throughout the year", *Transactions in GIS*, vol. 5, pp. In Press.

Lagacherie, P. and Holmes, S. (1997) "Addressing geographical data errors in a classification tree for soil unit prediction", *International Journal of Geographical Information Science*, vol. 11, pp. 183-198.

Lam, N.S.-M. (1983) "Spatial Interpolation Methods: A Review", *The American Cartographer*, vol. 10, pp. 129-149.

Leung, Y. and Leung, K.S. (1993) "An intelligent expert-system shell for knowledge-based geographical information systems. 1. The tools",

International Journal of Geographical Information Systems, vol. 7, pp. 189-199.

MacGill, J., Openshaw, S. and Turton, I. (1999) "Web-based multi-agent spatial analysis tools", *GeoComputation'99*, Virginia: CD-ROM

Maslyn, R.M. (1987) "Gridding advisor: An expert system for selecting gridding algorithms", *Geobyte*, vol. 2, pp. 42-43.

Mitás, L. and Mitásovà, H. (1999) "Spatial interpolation". *Geographical Information Systems*, (ed. by P.A. Longley, M.F. Goodchild, D.J. Maguire & D.W. Rhind), pp 481-492, Wiley: New York.

Morse, B. (1987) "Expert interface to a geographical information system", *AutoCarto-8*: 535-541

Murnion, S. and Carver, S. (1996) "Automatic analysis rule generation using genetic algorithms", *GeoComputation'96*, Leeds: http://www.ashville.demon.co.uk/gc1996/

Oliver, M.A. and Webster, R. (1990) "Kriging: a method of interpolation for geographical information systems", *International Journal of Geographical Information Systems*, vol. 4, pp. 313-332.

Openshaw, S. and Alvanides, S. (1999) "Applying geocomputation to the analysis of spatial distributions". *Geographical Information Systems - Principles and Technical Issues*, (ed. by P.A. Longley, M.F. Goodchild, D.J. Maguire & D.W. Rhind), pp 267-282, Wiley: New York.

Openshaw, S. and Openshaw, C. (1997) *Artificial Intelligence in Geography*, Wiley: Chichester.

Pannatier, Y. (1996) *VARIOWIN: Software for spatial data analysis in 2D*, Springer-Verlag: New York.

Rigol, J.P., Jarvis, C.H. and Stuart, N. (2001) "Artificial neural networks as a tool for spatial interpolation", *International Journal of Geographical Information Science*, vol. 15, pp. 323-343.

Shariff, A.R.B.M., Egenhofer, M.J. and Mark, D.M. (1998) "Natural-language spatial relations between linear and areal objects: the topology and metric of English-language terms", *International Journal of Geographical Information Science*, vol. 12, pp. 215-245.

Skidmore, A.K., Gauld, A. and Walker, P. (1996) "Classification of kangeroo habitat distribution using three GIS models", *International Journal of Geographical Information Systems*, vol. 10, pp. 441-454.

Sondheim, M., Gardels, K. and Buehler, K. (1999) "GIS interoperability". *Geographical Information Systems - Principles and Technical Issues*, (ed. by P.A. Longley, M.F. Goodchild, D.J. Maguire & D.W. Rhind), pp 347-358, Wiley: New York.

Stefanakis, E., Vazirgiannis, M. and Sellis, T. (1999) "Incorporating fuzzy set methodologies in a DBMS repository for the application domain of GIS", *1999*, vol. 13, pp. 657-675.

Vckovski, A. and Bucher, F. (1996) "Virtual Data Sets - Smart Data for Environmental Applications", *3rd international Conference on GIS and Environmental Modelling*. National Centre for Geographical Data and Analysis, Santa-Fe, New Mexico: CD-ROM

Wang, F.J. (2000) "A fuzzy grammar and possibility theory-based natural language user interface for spatial queries", *Fuzzy Sets and Systems*, vol. 113, pp. 147-159.

Webster, R. and Oliver, M.A. (1990) *Statistical methods in soil and land resource survey*, Oxford University Press.

Westervelt, D.J. and Hopkins, L.D. (1999) "Modelling mobile individuals in dynamic landscapes", *International Journal of Geographical Information Science*, vol. 13, pp. 191-209.

Williams, G.J. (1995) "Templates for spatial reasoning in responsive geographical information systems", *International Journal of Geographical Information Systems*, vol. 9, pp. 117-131.

Zhu, X. (1996) *A knowledge-based approach to the design and implementation of spatial decision support systems*. PhD Thesis, University: Edinburgh, 274 pp

Zhu, X. and Healey, R. (1992) "Towards intelligent spatial decision support: integrating expert systems and GIS", *GIS/LIS'92*, San José: 877-886