Multi-Objective Optimisation on Transportation Networks

Dan Costelloe, Peter Mooney and Adam Winstanley.
Department of Computer Science,
National University of Ireland Maynooth (NUIM),
Co. Kildare. Ireland.
Telephone: +365 1 123456 Email:
{danc, pmooney, adamw}@cs.may.ie

Abstract

A multi-objective optimization scheme for transportation planning is described. Multi-criteria analysis provides a major advantage in its ability to take into account a range of different, often unrelated criteria, even if these criteria cannot be directly related to quantitative outcome measures. The approach described is specifically addressed to public transportation networks but it is also applicable to other types of physical networks including computer networks. Research into developing an evolutionary approach to public transportation network optimisation, by use of a carefully chosen fitness function, is outlined.

1. Introduction

Transportation analysis within a GIS (Geographic Information System) environment has become common practice in many application areas. Transport, by its very nature, lends itself to a multi-disciplinary study. An every increasing need for complex path algorithms and path computation has developed from the rapid emergence of GIS systems such as intelligent vehicle systems. Optimal route planning is made complicated by the existence of factors such as multiple modes, planned arrival and departure schedules, multiple fare structures, dynamic changes to the network. Route or journey planning is the systematic search through a transportation network to find an *optimal* journey specification. This specification is nearly always required to satisfy some initial set of constraints (times, road types, costs). The set of constraints is better described as a preference for particular routes and departure/arrival times and desired departure and destination locations. Constraints may be placed on variables or criterion that are easy to quantify, for example departure and destination time. However, other criteria are more difficult to quantify. Examples include preference for certain modes or transport and preference for particular road types.

Optimal journey specifications can now be defined as a journey specification exhibiting minimal values for all variables (criteria, objectives) considered. However, humans are seldom capable of discovering these optimal solutions unless the network search space is relatively small. It is very often the case that comprehensive searches are too expensive in terms of information gathering and retrieval and search time. To avoid the costs involved in the searching process, in terms of effort and time, humans will only attempt to find *any* satisfactory journey specification. Behavioral scientists define the term *satisficing* (Nijkamp and Van Deft 1971) for this type of human behaviour in regard to information searching and decision-making in large search spaces. Humans are risk aversive in selecting alternative journeys when a journey specification that satisfies certain minimal, weak, conditions and criteria has been found.

We propose a series of computer-aided techniques to assist travellers in searching for and planning more efficient journeys on a transportation network. A given (transportation route finding) problem may have a set of solutions, some good, some not so good. Within this set of solutions (if it exists), there also exists a subset of optimal (best) solutions. Depending on the problem instance, there may be one optimal solution or a group of them. Using these techniques we can find the best solution or set of best solutions corresponding to journey constraints and optimisation criteria. Optimisation may now be redefined as the task of finding the(se) best solution(s).

2. Criteria and Objectives

Early routing models typically used the standard linear programming techniques to optimise an objective function consisting of a single criterion or a weighted combination of multiple criteria. This type of approach to multi-objective routing does not allow a complete analysis of trade-offs between the various criteria when some weighted combination is optimised. The weighted combination approach does not in any way guarantee that all non-dominated paths will be discovered. It is those non-dominated paths that describe journey specifications exhibiting minimal values for criteria such that they cannot be bettered by other journey specifications on all criteria. A simple example provides motivation for these claims. Suppose that a journey from A to B is described by a 3-D vector AB with elements [time required, financial cost, distance] and that for a particular network model (with A and B defined) this vector is [100, 10, 200]. Another vector AB* (also describing a journey from A to B) has elements [120, 5, 210]. If one compares the vectors AB and AB* on a weighted combination or sum of elements it is easy to determine the *minimal* journey specification. However, this naïve analysis yields a decision on optimality that is not taken with respect to the entire decision variable space. As the vector dimensions grow larger to incorporate more criteria and objectives this analysis becomes more unpredictable and unstable.

When more than one optimisation criteria are involved, aggregate-sum approaches are often applied to condense the multiple objectives into one to make the optimisation process easier. For some problems, however, this approach may not be feasible, as trade-offs may exist between criteria: an increase in one may result in a decrease in another (some or all of the time), depending on the values of the other criteria. It is in cases like this that a technique called Multi-objective Optimisation may be employed. For a survey of Evolutionary Multi-objective Optimisation techniques, see (Coello 1999). Optimisation techniques are used to achieve the ‘best’ (or as close to the ‘best’ as possible) solution(s) to a given problem. The ‘best’ solution may be the one that takes the least amount of time to compute, costs the least financially or achieves the highest score according to some evaluation scheme. Many optimisation techniques involve navigating the *search space* for an optimal solution. Some problems have very large search spaces, meaning that simple, brute-force searches are too complex in terms of the time it would take (or the necessary resources) for the search to complete. It is for this reason that approaches such as Tabu Search (Glover 1990), Simulated Annealing (Metropolis and Rosenbluth 1958), and Genetic Algorithms (Holland 1975) (among others) have evolved.

3. Multi-Objective Optimisation.

In multi-objective optimisation problems, each objective may be represented as a vector entry, with the vector itself representing a solution, for example:

$$X\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

represents a single solution to an n-objective optimisation problem. Given two vectors (*X* and *Y)*, we say that $X <_p Y$ (*X is partially less than Y* or *X dominates Y*) if:

$$X <_p Y \Leftrightarrow \forall i(x_i \leq y_i) \wedge \exists i(x_i < y_i)$$

Solutions which are non-dominated (the *Pareto-optimal* set) can be considered as better overall solutions than those in the whole set as they have no other solutions that are better than them on all criteria. While this process doesn't necessarily identify any one outstanding solution, it does narrow down the search space to a set of solutions, which may be more easily navigated.

Multi-modal, multi-criteria optimised route planning is not yet well studied in the literature. Methods for solving single objective planning problems have been studied extensively for the past 40 years. However, almost every important real world problem involves more than one objective. Multiple objective optimisation problems are similar to single optimisation problems except that they have a stack (vector) of objectives (criteria) to optimise rather than just a single one. (Costelloe et al. 2000) provides a solution methodology for multi-objective optimisation of routes on a static network model with at least three objectives.

3.1 Producing the Pareto-optimal set

Our implementation produces a set of candidate solutions *C* from a graph *G* (model of the public transportation network) and extracts the Pareto-optimal set *P* from *C*. Each member of *C* has an associated path description vector of the form:

$$C_{ab}\begin{bmatrix} time \\ \cos t \\ changes \end{bmatrix}$$

Each candidate represents a journey from *a* to *b* with its associated time, cost and number of modal changes. Of course other choices of criteria are allowed depending on the situation. *P* is then obtained by searching the set *C* for non-dominated solutions. The solutions in *P* are considered better than those that are not in *P*, because their path description vectors cannot be bettered on all criteria. Once *P* has been constructed, the decision-maker must then choose which P_{ij} to use. With this approach the path description vector may be extended to handle more entries (objectives) to analyse tradeoffs between these entries.

3.2 Computing the Pareto Optimal Set.

The first step in computing the Pareto Optimal Set is to produce the set of candidate solutions C. Each node in the graph model G = (V,E,S) is either a intermediate stopping point or major station (intersection) on the public transport network. The set S represents the set of

routes operating on the entire network. Nodes can then be partitioned into departure destination pairs. Then for every departure destination pair, several different paths must be computed. These are deemed the candidate solutions for the optimal path(s) between the departure and destination nodes. Individual candidate solutions are obtained by running several different route finding algorithms each with different strategies for finding the shortest, cheapest etc path between the pair of nodes in question. Examples of the route finding algorithms implemented are A*, Dijkstra's Algorithm and Bellman-Ford-Moore. Each of the algorithms implemented optimise on one criteria. The problem in hand (as described above) has three criteria. To deal with this, each algorithm is run three times optimising on a different variable over each separate run. After the route finding algorithms have terminated a set C_{ab} exists containing paths and path description vectors for paths between nodes a and b. Each element in the path description vector is the cumulative value of the corresponding criteria over the entire path from node a to node b. The Pareto Optimal approach is then applied (as described in section 3.1) to construct P_{ab}, the Pareto Optimal Set of journey specifications between the nodes a and b. The path description vectors of candidate solutions are compared with each other rather than with some predefined global optimum path description vector. This is a consequence of the fact that it is not intuitive to define a global optimum vector for paths between any two nodes.

However, this approach has worked very well for static network structures. Static network structures are networks that do not change over time or any changes that occur are separated by a long period of time. This type of static model is mathematically sound but hardly a realistic model of a public transportation network. Transportation networks are inherently dynamic. Changes in route patterns, traffic congestion, road/street availability can change dramatically in a short space of time. Such dynamic changes often render previous estimates of shortest paths or optimal journey specifications incorrect. After a dynamic change, a public transportation information system must quickly update information on shortest/optimal path specification, connectivity structures etc in order to provide up-to-date information to queries. To capture this dynamic behaviour we adopt approaches from the field of dynamic graph algorithms.

4. Dynamic Graph Algorithms

Graph theory and graph data structures and algorithms are inextricably linked with models of any type of physical network. It is well known that computing shortest paths and connectivity relations over a network is the most important task in many network and transportation analyses. Transportation networks possess different levels of congestion, road availability and throughput during different periods of the day. Therefore it is unrealistic to precompute all shortest paths and connectivity at the start of the day and use these to answer queries over the remainder of the day. Shortest paths and other network properties must be updated in real-time that is as soon as a dynamic change occurs on the transportation network. Figure 1 below details the process of updating after a dynamic change. Queries regarding shortest paths or connectivity relations at time t = T are answered regarding the network model in its most up-to-date state.

Dynamic graph data structures and associated algorithms (Eppstein 1998) provide a robust model with which to effectively model the dynamic nature of a public transport network. When a dynamic change fundamentally changes some property or characteristic of the public transport network it is inefficient to re-compute this and related properties from scratch each time. Dynamic information updating (edge congestion, node availability) has a tremendous effect on the planned optimised route causing it, in most cases, to deteriorate. Dynamic algo-

rithms have been shown to be remarkably better than static shortest path algorithms for solving dynamic shortest path problems (Frigioni and Nanni. 1998).

		Dynamic Change	Dynamic Change	QUERY	Dynamic Change
TIME	t = 0	t = t1	t = t2	t = t3	t = t4
ACTION	Build G = (V,E,S)	UPDATE	UPDATE		UPDATE
ACTION	Compute Optimal Paths	UPDATE	UPDATE	Return Results	UPDATE

Figure 1: A timeline table of process of dynamic update to a graph model G = (V,E,S) when queries are mixed with update requirements.

5. Graph Mutation

Evolutionary computation described by (Holland 1975) takes a set of candidate solutions to a problem and, using techniques borrowed from natural selection and evolution, evolves these solutions towards 'fitter' states. In our case 'fitter' states means better solutions in terms of journey specifications. This process has been shown to be efficient at gaining near-optimal solutions to hard problems in polynomial time. The fitness of a set of solutions is evaluated by implementing some decision making process to quantify the relative merit of this evolved set of solutions over the old set of solutions. The evolved set of solutions are deemed 'fitter' if and only if they are in some way quantifiably better that the old set of solutions.

This aspect of our evolutionary approach involves making random edge-insertions and edge deletions to the original graph *G*, producing *G'*. In the context of a public transportation network this mimics the addition of a route to some previously unused street or road or alternatively the removal of a route link between two nodes. This can be viewed as a random form of route service renewal. All choices regarding edge removal or addition are made randomly. In essence we initiate an unbiased scheme to make alterations to the current graph model of the transportation network. After each change is made a new graph is created. This new graph (a solution) must be evaluated for its fitness to be evolved further. This fitness evaluation requires the Pareto Optimal set of optimal journey specifications to be recomputed. The Pareto Optimal set of the evolved graph and that of the original graph are compared. Based on this comparison conclusions may be drawn on the relative efficiency of the new graph model over the original.

Formally, the insertion or deletion of edges causes the current graph to be mutated into some different graph *G'*. The resultant candidate set of this graph, *C'* can be used to produce another Pareto-optimal set *P'*. If it is the case that the cardinality of the set *P'* is less than that of *P* (for a given journey between two nodes) then we know that the random insertion or deletion has produced a better solution than was previously identified. The following example illustrates this (2-d vectors are used for simplicity – modal changes have been omitted):

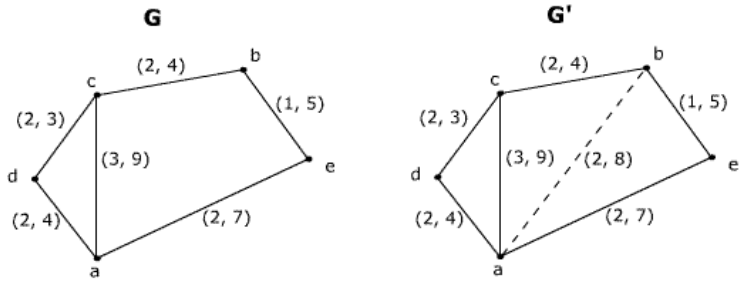

Figure 2: Depicting an original graph model G = (V,E) (with two optimisation criteria). G' is a graph with an edge inserted.

For journeys between node a and b, the following candidate set is produced from the graph G:

Path	Time	Cost
a->d->c->b	6	11
a->e->b	3	12
a->c->b	5	13

Extracting the Pareto-optimal set gives:

Path	Time	Cost
a->d->c->b	6	11
a->e->b	3	12

If a random edge-insertion were to produce G', the new candidate set produced is:

Path	Time	Cost
a->d->c->b	6	11
a->e->b	3	12
a->c->b	5	13
a->b	2	8

The Pareto-optimal set now becomes:

Path	Time	Cost
a->b	2	8

Members of the Pareto-optimal set for the original graph G are no longer present; it is still possible to use these routes however as the decision maker may choose an alternative route based on other criteria that are not covered by this approach.

The application domain of the graph mutation scheme has, at present, only been applied to public transportation networks in an effort to evolve more efficient network and route structures. Public transportation networks may be viewed as dynamic masks that overlay an existing network (i.e. a road and street network). To this end, this mask may be evolved and changed in an effort to improve the configuration of the network and route structure of the mask without any alteration to the underlying physical network. After a series of evolutions the public transportation network mask (by inspecting the Pareto Optimal Set for journey specifications between nodes) should admit 'better' or more 'fitter' solutions (in terms of journeys and multi-criteria optimisation) than any of its predecessors. This mutant graph (and mask) may then be strongly considered as a replacement for the existing mask. More efficient and better-planned network structures allow the public transportation system operating on this network to operate more effectively. For example there may be a better distribution of routes or a more efficient distribution of these routes causing certain routes to avoid areas of high congestion for example. Also, if the structure of the network is improved route-finding algorithms may find it easier to compute optimal path specifications. The approach of graph mutation is restricted to providing a more efficient and effective structure for journey planning i.e. in regard to network design. It is not necessarily a constituent part of multi-objective analysis.

6. Conclusions and Further Work

The methodology described is designed for implementation as part of the design for Internet based public transportation information systems. Such systems provide infrequent visitors and users of public transportation with a set of optimal journey specifications to choose from during the journey planning process. This set of optimal journey specifications provides valuable guidance during the process of navigating oneself in an unfamiliar urban or suburban environment on a given transportation system. The multi-objective approach to decision making is not in any way confined to the domain of transportation analysis and optimisation. In fact any problem requiring the optimisation of solutions to problems defined in terms of a number of (conflicting or independent) criteria when it is neither possible nor sensible to combine all criteria in some form of aggregation of the criteria.

In this paper we have provided an evolutionary computation framework for the solution of the problem of multi-objective optimisation on a transportation network. This framework has been implemented for optimisation problems involving three criteria. However, more criteria (provided that they can be quantified in some way) may be added without any major changes to the theory. The Pareto Optimal processing stage is $O(n^2)$ in computational complexity and remains so despite the addition of further criteria to the problem. This is a result of the definition of vector domination found in section 3. Preliminary results also found that no particular route finding algorithm was dominant in finding solutions that turned out to be members of the Pareto Optimal set of solutions. Dijkstra's algorithm for example will always find the shortest path (on one cost metric) over a graph structure (Cormen 1999). However there are two reasons why it does not hold a majority on the number of solutions it provides to the Pareto Optimal set. Firstly, while the algorithm will optimise on one criteria or metric the other criteria in the problem are cumulatively gathered and specified in the path description vector. It is then the vector itself (and the tradeoffs between criteria) that determine the solutions suitability for inclusion into the Pareto Optimal set. Secondly, Dijkstra's algorithm, as

well as the other algorithms, performs well on easily quantifiable objectives such as path length, overall path time. However, for criteria that are difficult to quantity i.e. level of convenience of a route or number of modal changes the algorithms perform poorly. This is due to the decision making involved in optimising such quantities. To try to optimise the total number of modal changes one needs to perform some form of look ahead in an attempt to predict possible interchange and connection points further 'downstream' of the current node.

The issue of optimisation on a transportation network provides a fertile ground for further research. Our proposal is novel in that it investigates ways of identifying inefficient network structures and deals with multi-objective problems on a dynamic public transportation network. On real-world networks such as these, many of the objectives may be loosely formulated. Examples include, convenience of a route specification, favouritism towards particular routes, road types etc. There has been little research work documented on a crossover approach to optimal design of transportation networks.

References

Coello, C. A. C., "An Updated Survey of Evolutionary Multiobjective Optimization Techniques: State of the Art and Future Trends", Proceedings of the Congress on Evolutionary Computation, Vol. 1, 3-13, IEEE Press, 6-9 July 1999.

Cormen T.H. , *Introduction to Algorithms*. MIT Press, Massachusetts, USA. 1999

Costelloe D, Mooney P and Winstanley A., Multi-Objective Optimisation and Dynamic Routing Algorithms on Public Transportation Networks. *Proc. of GIScience 2001,* Savannah, Georgia USA.

Eppstein D., Dynamic Graph Algorithms. in *Algorithms and Theory of Computation Handbook,* CRC Press, Florida. USA. 1998.

Frigioni D, Nanni U. *Experimental Analysis of Dynamic Algorithms for the Single Source Shortest Path Problem*. Technical Report, University of Rome, "La Sapienza", Italy. 1998.

Glover, F., Tabu Search: A Tutorial, *Interfaces*, Vol. 20 (1), 74-94, July-August 1990.

Holland J.H., *Adaptation in Natural and Artificial Systems,* University of Michigan Press, Ann Arbor, MI 1975.

Metropolis, N and Rosenbluth K.., Equations of State Calculations by Fast Computing Machines, *J. Chem. Phys*., 21, 1087- 1092, 1958.

Nijkamp P and Van Deft A., *Multi-Criteria and Regional Decision Making,* Martinus Nijhoff Science Division, Leiden 1971.

An empirical study to assess the accuracy of simple aerial interpolation methods

Paul Brindley[1], Stephen Wise[2], Peter Fryers[3], Ravi Maheswaran[3] & Robert Haining[4]
[1] University of Sheffield
Sheffield Centre for Geographical Information and Spatial Analysis (SCGISA)
Sheffield, S10 2TN
E-mail: P.Brindley@Sheffield.ac.uk

[2] University of Sheffield
Department of Geography
Sheffield, S10 2TN
E-mail: S.Wise@Sheffield.ac.uk

[3] University of Sheffield
School of Health and Related Research
Sheffield, S1 4DA
E-mail: P.R.Fryers@Sheffield.ac.uk
R.Maheswaran@Sheffield.ac.uk

[4] University of Cambridge
Department of Geography
Cambridge, CB2 3EN
E-mail: RPH26@Cam.ac.uk

Extended abstract

The development of Geographic Information Systems (GIS) has been assisted by the recent expansion in the availability of digital data. However, this has also highlighted a number of problems including technical issues associated with incompatible zonal systems. The issue arises when data values for one zoning system are required for another, but the boundaries of the datasets do not align. The well-documented solution is areal interpolation, whereby data from existing source zones are re-estimated for the required target units (see Goodchild and Lam, 1980; Flowerdew and Openshaw, 1987; Flowerdew et al*, 1991; or Fisher and Langford, 1995). However, there are numerous methods of areal interpolation each with their own assumptions which directly affect the estimates made.*

With respect to the definition of accuracy there is some ambiguity within the literature. Commonly, accuracy refers to the degree of conformity of a measured or calculated value in comparison to its real value. However, Buttenfield and Beard (1994) stipulate that accuracy differs conceptually from error by measuring discrepancy from a model*, while error measures discrepancy from the* truth*. Using this latter terminology, although the actual pollution concentrations for each ED are not known, the investigation seeks to examine any differences that may exist between differing models for the estimated levels of pollution.*

When assessing the accuracy of differing areal interpolation techniques, the majority of studies have concentrated on the size and shape of source zones (Fisher and Langford, 1995; Cookings et al*, 1997; Okabe and Sadahiro, 1997; Sadahiro, 1999a; Sadahiro, 1999b; and Sadahiro, 2000), the location of representative points (Sadahiro, 2000), or the number of source boundaries intersected by the target boundary (Sadahiro, 1999b). Few investigations have examined the properties of the data values themselves or the effect on accuracy of*

differing population distributions within the target zones. This paper seeks to readdress this with assessment of the level of error that is associated with simplistic methods of interpolation (point-in-polygon and areal weighted) within a study area of the city of Sheffield. Parameters including the areal size of target zones, the underlying population distribution and the steepness of gradient for the source unit attribute data within each target zone, will be examined.

This discussion of the accuracy of simple areal interpolation techniques has developed from within the framework of a project undertaken within Sheffield on the relationship between outdoor air pollution and respiratory and cardiovascular disease. Using this project as context the paper seeks to interpolate gridded pollution data as supplied by the Environmental Protection Service, Sheffield City Council, to enumeration district (ED) boundary data for the city of Sheffield. The pollutants investigated include: carbon monoxide – CO, nitrogen oxide – NO_x, Particulate matter - PM_{10} and Sulphur dioxide – SO_2. EDs are the smallest spatial unit for aggregated data within British Census data and contain approximately one hundred and fifty households (Rhind, 1991). The intention is then to assess the relationship between air pollution and health data on admissions, symptoms and mortality for respiratory and cardiovascular disease, and the interaction with socioeconomic deprivation and age. Thus, environmental information which is frequently collected in a diverse number of zonal systems (in this case as grid data) will be compared to census derived socio-economic data.

'Intelligent' methods of areal interpolation use ancillary information to improve the estimation of the target-zone densities (Flowerdew and Green, 1989 and Flowerdew et al*, 1991). In the context of this investigation, pollution estimates for EDs are required to reflect the underlying population distribution within the ED so that people's exposure levels can be measured. Thus, the better represented the population distribution is, the greater the chance that this portrays a truer picture of the population's exposure.*

Relatively new accurate commercial point databases incorporating population distribution are now available which could possibly enhance areal interpolation estimates required for an underlying population. Based upon Ordnance Survey's CodePoint data (which provides a precise geographical location for each 'postcode unit') PostPoint Professional includes the unit postcode location with an accuracy of one metre and also contains the number of domestic and non-domestic delivery points. Point-in-polygon aerial interpolation at the PostPoint locations, weighted by the number of domestic properties sets the benchmark ('gold-standard') by which the other methods of aerial interpolation were examined. Thus, the effect of only having the population weighted centers (included within the 1991 United Kingdom census boundaries) for each ED (point-in-polygon aerial interpolation) or no additional information (aerial weighted technique) was evaluated. The data utilized is shown within figure one and analysis was performed using ArcView – the results are illustrated within figure two. Quintile plots and log scatter-plots were used to compare the absolute differences between the differing methods and the 'gold-standard' for each ED.

The simplistic areal interpolation methods gave more accurate estimates if smaller target areas were present or if the source data's variability within the target zone was more constant. Less error was associated with the method if the underlying population distributions were concentrated within just one or two clusters rather than being either dispersed or concentrated in numerous distinct clusters. Areal weighting consistently gave estimates closer to those obtained by the 'gold-standard'. However, this statement should be placed within the context of the study in which the majority of the coverage is urbanized, thus, reflecting a more evenly distributed population than more rural areas.

It was shown that although differing areal interpolation techniques might generally give similar overall results, substantial differences are apparent upon closer examination. No

individual factor had a dominant influence on the accuracy of areal interpolation using either areal weighting or point-in-polygon techniques. However, the results demonstrate that the steepness of gradient for the source unit attribute data within each target zone is an important parameter influencing the accuracy of the method, and as such, should not be overlooked.

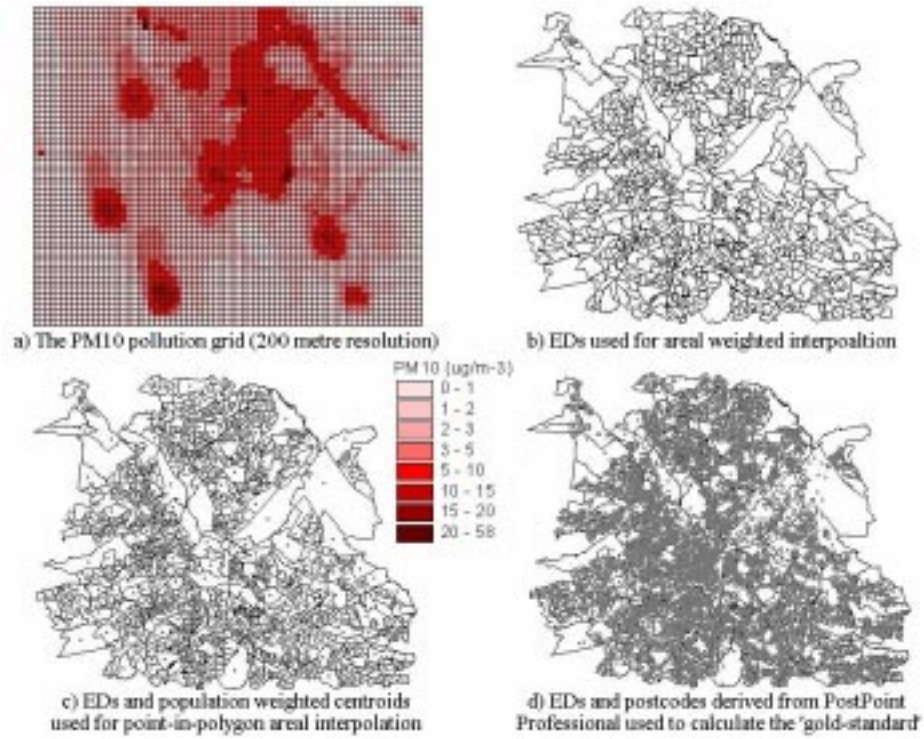

Figure 1. The data utilised within the investigation

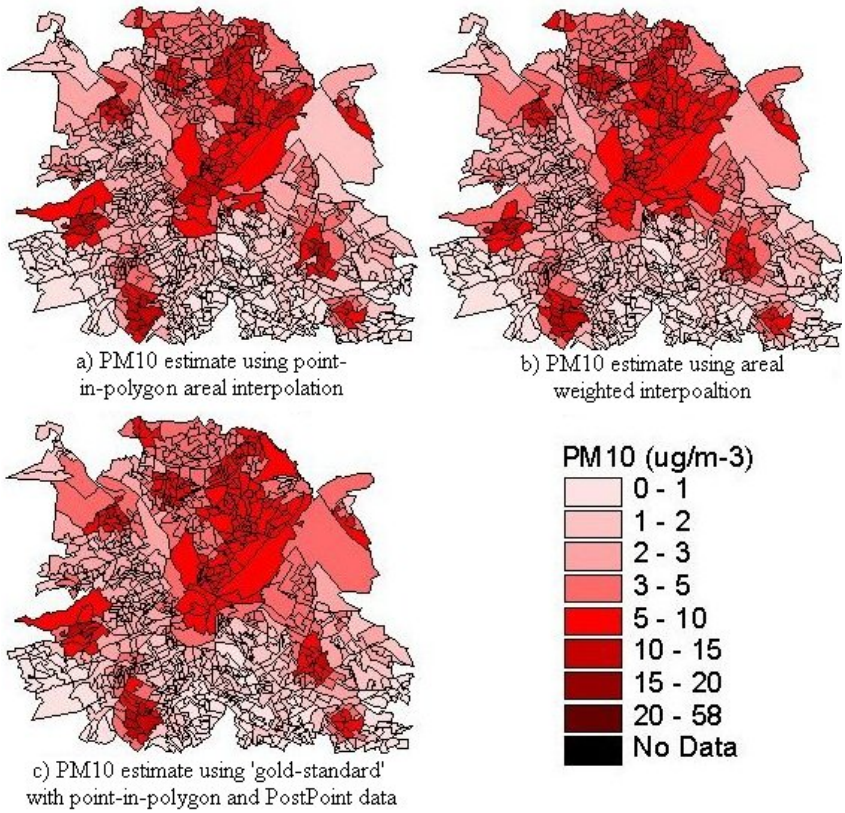

Figure 2. Average PM10 pollution values at ED level, derived from various simplistic areal interpolation techniques

References

Buttenfield, B. and Beard, M. 1994. 'Graphical and geographical components of data quality' in H. Hearnshaw and D. Unwin (eds) Visualizations in Geographical Information Systems, Wiley: Chichester.

Cookings, S., Fisher, P.F. and Langford, M. 1997. Parameterization and visulaisation of the errors in areal interpolation. *Geographical Analysis* **29** (4): 314:328.

Fisher, P.F. and Langford, M.1995. Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A* **27** (2): 211-224.

Flowerdew, R. and Green, M. 1989. Statistical methods for inference between incompatible zonal systems. *North West Regional Research Laboratory*: research report no. **1**, The University of Lancaster.

Flowerdew, R., Green, M. and Kehris, E. 1991. Using areal interpolation methods in geographic information systems. *Papers in Regional Science: the Journal of the RSAI* **70** (3): 303-315.

Flowerdew, R. and Openshaw, S. 1987. A review of the problems of transferring data from one set of areal units to another incompatible set. *Northern Regional Research Laboratory*: research report no. **0**.

Goodchild, M.F. and Lam, N. N-S. 1980. Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing* **1**: 297-312.

Okabe, A. and Sadahiro, Y. 1997. Variation in count data transferred from a set of irregular zones to a set of regular zones through the point-in-polygon method. *International Journal of Geographical Information Science* **11** (1): 93-106.

Rhind, D.W. 1991. 'Counting the people: the role of GIS' in D.J. Maguire, M.F. Goodchild and D.W. Rhind (eds) *Geographical Information Systems, volume 2: Principles and Applications*, Longman: New York.

Sadahiro, Y. 1999a. Accuracy of areal interpolation: a comparison of alternative methods. *Journal of Geographical Systems* **1** (4): 323-346.

Sadahiro, Y. 1999b. Accuracy of count data transferred through the areal weighting interpolated method. *Center for Spatial Information Science and Department of Urban Engineering*, CSIS discussion paper no. **5**, University of Tokyo.

Sadahiro, Y. 2000. Accuracy of count data estimated by the point-in-poly method. *Geographical Analysis* **32** (1): 64-89.

Archaeological Predictive Modelling in Cultural Resource Management

Toma• Podobnikar, Tatjana Veljanovski, Zoran Stanèiè and Krištof Oštir
Scientific Research Centre of the Slovenian Academy of Sciences and Arts
Gosposka 13, 1000 Ljubljana, Slovenia,
tel.: +386 1 470 64 95, fax: +386 1 425 77 95
Tomaz@zrc-sazu.si, TatjanaV@zrc-sazu.si, Zoran@zrc-sazu.si,
Kristof@zrc-sazu.si

Abstract

Predictive models are becoming increasingly often used in archaeological cultural resource management. Beside this, extremely successful and productive application, predictive models can be and are used as an effective tool in archaeological site location explanation. The main objective of this presentation is to discuss and present some aspects of practical applications of predictive modelling. The presentation will start with theoretical introduction to predictive modelling and will be followed with some methodological issues. Special emphasis will be paid to the presentation of several case studies. The first set of case studies will focus on the application of multiple overlays of spatial information layers for modelling potential of Bronze Age settlement sites location and barrows. Further more some results of the multivariate statistics for the analysis of Roman settlement patterns will be presented. Finally, we will demonstrate how site locations were predicted in Slovenian highway constructions project. Presentation will be concluded with some general remarks and practical suggestions for future work.

1. Background

Doing prediction is accepted procedure in every bit of situations where an endeavour is to look into some continuation (yet unknown) of some actions or some behaviour. Recalling on only a few, prediction is being supposed in weather forecast, however, the similar line can be find in risk analyses, animal behavioural studies in biology, etc. Following various themes ("subjects") being under predicting aim in different fields, the predicting techniques differ as well. It might be said, that in a GIS sense, prediction becomes restricted from comprehensive prediction in that at the end it yields *spatial* predictive information. In addition, some of these techniques have been successfully used for several archaeological implications.

Human knowledge of the past is neither perfectly certain nor perfectly uncertain. Predictive modelling should therefore be described as an attempt to draw recognized behaviour further, an extension, of an otherwise unobservable phenomenon. One way of dealing with limited certainty is to think in terms of probabilities. Another way that becomes used also in the field of archaeology and even more in cultural resource management is to think in terms of potential. Speaking thus about a map of archaeological potential in a given landscape, as a result of prediction, we shall be well aware that in turn it could play a role of valuable output, particularly important not only for archaeologists but for environmental planning i.e. landscape managers as well.

Within archaeology the study of settlement patterns is important when debating some degree of knowledge as regards the landscape and past human occupation of the area. Through this

point we wish to introduce the expression of archaeological predictive modelling. Predictive modelling is a technique, used to predict site locations in a region, on the basis of observed patterns, or, on assumptions about human behaviour (Kohler and Parker 1986, Judge and Sebastian 1988). Archaeological predictive models are essentially based on the fundamental assumption that our knowledge of known archaeological sites allows us to establish which factors influenced their location in the landscape and to use this data in empirical testing. Prediction is merely the elucidation of settlement "rules" in a form which allow us to map locations which conform to the "conditions" predicted by the model for settlement. To achieve this we analyse the relationship between the natural and social environment and archaeological site location. What should not be neglect is unpleasant fact that we can never be certain how far from entirety is available archaeological data sample.

Although, predictive "modelling" is something archaeologists always keep in their mind when trying to locate the sites, the early beginnings of predictive modelling as known today, i.e. performed with GIS, go back to late 1970s, the time when geographical information systems were introduced into archaeology. Predictive models have a fairly long tradition in American archaeology, where they have been intensively used for cultural resource management. Generally, the legal basis underwriting archaeological fieldwork is different in European countries and allows archaeologists to carry out work on privately owned land with relative ease. As a consequence of this situation archaeological predictive models were not utilised in Europe until the later 1980s. Today, we may speak about two branches of predictive models: those developing related to the pragmatic aspect of use (i.e. mapping sensitive areas within a landscape in order to protect current archaeological heritage), and the academic ones (where an attempt to reconstruct past societies lead towards models which would increase our understanding of settlement on the basis of locational determinants).

Therefore, main distinction between the two approaches is evident proneness of academic approach to establish the determinants which influenced settlement behaviour in the past. Following this, a given archaeological site can be understood as a reflection of certain decisions and represents the final choice within a certain cultural system. The selection of, for example, a dwelling-place in a given landscape is then a complex reflection of such decisions which can also be influenced by personal preference, in accordance with the potential offered by the natural environment. The incorporation of a human component within locational modelling acknowledges the role of subjective judgement, but at the same time models are becoming more cognitive. Despite this, the analysis of the social environment might include those variables which are understood as descriptive of man's cultural, religious, ideological or economical relationship with the landscape, as they can tell more, but are often omitted from modelling procedure as they are more difficult to be obtained and then being reliably transformed into a GIS layer.

The scope of this paper is not to discuss in detail theoretical approaches to predictive modelling. However, two approaches can be defined: inductive and deductive. The truth is that both approaches overlap in practise. With the inductive approach, one starts with the basic archaeological data and the model is developed, based on the correlation between known archaeological sites and physical landscape. The deductive approach starts with a priori theoretical knowledge and a try to deduce relevant conclusions related to the logic of settlement patterns and land use in the past – upon those the model is then developed.

The technological problems when applying any of the above theoretical approaches can also be briefly mentioned. Generally speaking, one can apply a number of alternative methods of procedure. The Boolean overlay of variables for which it has been proved that has in some

way influenced location patterns may be utilised along with multivariate statistical techniques (linear regression, logistic regression, discriminant analyses), or some other decision support methods (i.e. Damster Shafer theory - belief method). Although multivariate statistical approach (particularly linear regression) is very powerful and gives a detailed insight into the relationship between individual variables analysed, it can only be used when the number of sites available for analysis is large.

Very important, and often questionable, step in developing predictive models is testing of its accuracy and evaluation of its predictive power. It is important to emphasize this particular problem when archaeological predictive models are subjected to testing. More often than not, archaeological sample representing sites or other past features in a region is not unlimitedly large, and often we are forced to include all the available data into the model developing process. This is then resulting in that no independent data to test the model is left.

2. Predictive modelling through the case studies

The aim of this section is to bring closer some results and experiences with predictive modelling work. The first set of case studies will focus on the application of multiple overlays of spatial information layers for modelling potential of Bronze Age settlement sites location. Further more some results of the multivariate statistics for the analysis of Roman settlement patterns are shown. Both of them were performed to test the behaviour of two different methods, on two different data samples, and are therefore considered as "academic applications of predictive modelling". Then, we will demonstrate how site locations were predicted in Slovenian highway constructions project, and in turn we would demonstrate how this approach might be considered as a useful practical application in archaeological resource management.

3. "Academic" applications on the island of Braè, Croatia

Our applications on the island of Braè relates to work carried out in the Central Adriatic where an international team of archaeologist, historians, geographers and other specialists from Croatia, Canada, Britain and Slovenia have been studying the Central Dalmatian islands more than 15 years. The Adriatic Island Project has carried out work within a transact of islands running from the mainland to the small island of Palagru•a in the centre of Adriatic and incorporating islands Braè, Šolta, Hvar, Vis and a number of smaller islands. Currently, a synthesis of the project which includes all of the islands is in progress.

During the field campaign in 1994 all the interesting areas were surveyed using an extensive survey sampling strategy. During the fieldwork a total of more than 600 sites were recorded to the Archaeological relational database (Stanèiè et al. 1999).

Next, a natural environment database was produced. Thematic data of soils, geology and contour lines from different data sources were acquired. From contour lines, high accuracy digital elevation model (DEM) was produced. The alternative, vegetation data had been collected by satellite images using a LANDSAT Thematic Mapper. Combining natural database data, some more sophisticated models including natural variables were produced as for example vegetation index, erosion model etc.

3.1 Predicting Bronze Age hillfort locations

Main objective was to test how predictive modelling can be used on small number of sites. The model was developed on the area measuring around 120 km^2, what is approximately a

quarter of Braè surface. The selected area is well overlapping also with one out of three physiographic regions on the island of Braè. What is further important is, that here almost half of Bronze Age barrows and hillforts were located, and the area is therefore considered as a centre of Bronze Age activities on the island of Braè.

As working with very small data set (8 hillforts), great effort for sites prediction was aimed to the use of innovative variables. Variables that might have influenced hillfort locations should describe natural and social environment. Following this, we analysed a relationship between hillfort locations and the following variables: relief (elevation, slope, and ridge/drainage index, rim index and relief below index), cumulative viewshed index, distance from the coast, distance between the hillforts, cumulative distance from barrows and some others (Stanèiè and Kvamme 1999) for which we believed could be important. All the correlations were examined by various quantitative and statistical approaches.

On the basis of the established investigations about the relationship a threshold value on each variable was defined. This threshold values could be defined in different ways, but here was employed simple criteria, such that all the known sites would be captured by the threshold. Finally, a predictive model was made as a sum of those for which was proved to have influenced site locations (figure 1).

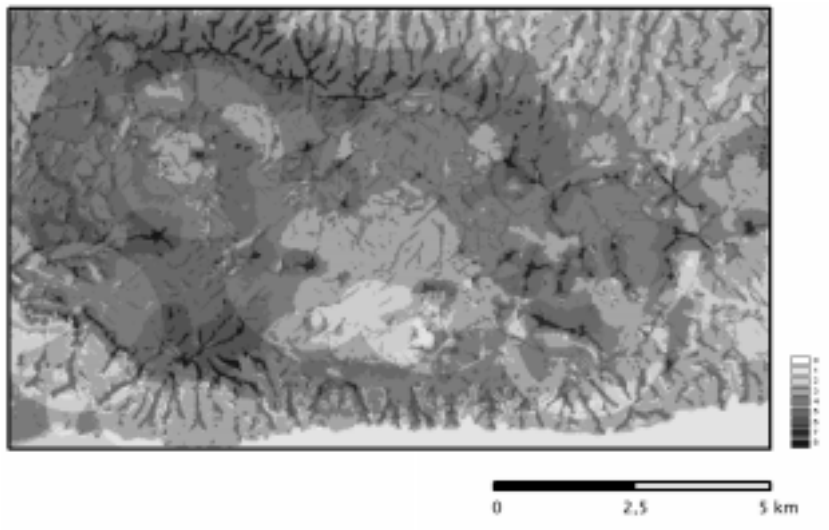

Figure 1: Archaeological predictive model for the Bronze Age hillfort locations on the island of Braè.

When analysing the final result we wish to evaluate the technique of predictive modelling used also through its transparency. Following this, the predictive model of Bronze Age hillfort locations is simple but powerful. Due to its simplicity and having good transparency, we may say the Boolean overlay methodology used is suitable for working even with smaller data samples. Increasing capabilities of GIS, on the other hand, are providing extensive insights into past systems of land occupation, and therefore allow archaeologists to test hypotheses and to make interpretations beyond.

3.2 Predicting Roman settlement sites

Second case study aimed to examine the potential of multivariate statistical techniques in the development of predictive models. Special emphasis was paid to the locational analysis of Roman settlement sites dated from the 2nd century BC to the 2nd century AD. A total of 29

sites out of approximately 90 Roman sites that were actually recorded could have been interpreted as settlements (Stanèiè et al. 1999). Beside the number of these Roman settlements is not large (concerning multivariate statistic requirements) and the specifics of its distribution over all the island of Braè, we decided that the model should be developed for the island as a whole.

An important step in analysing settlement patterns is to show that the characteristic of site locations differs significantly from general locations across the landscape. The first step in variable selection was based both on previous experience and empirical measurement of data. Therefore the initial set incorporated eight variables: elevation, measure of surface slope, measure of surface aspect, quality of soils, land use, erosion, distance from the coast and distance from Sennone limestone, which is viewed as an extension of surface geology (for more details see Stanèiè and Veljanovski 2000). Finally, four variables were recognised as possible predictors for Roman settlement locations on the island of Braè: aspect (south-west facing slopes), distance from Sennone Limestone (close to thin limestone zone), elevation (avoidance of elevated areas) and slope (less steep places). Promising variables were employed to the linear regression analyses.

Due to the fairly large number of sites, it was hoped that regression analysis might be applied. Regression solutions and the resulting functions relating site potential on 29 observations were not very promising in any combination we made (the adjusted R-square, interpreted also as a part of explained variance, did not exceed 48% in any of these cases).

In general, linear regression-based predictive model makes it possible to predict the potential that a site possesses at a given location when substantial care handling data is introduced. Despite the results was not as good as anticipated, it should be emphasised that this technique is good to provide a more detail insight into the importance of individual variables and their contribution to the settlement pattern. Regarding our case study, it was confirmed that the thin layer of Braè "marble" really affected locational decision-making on the island. Many queries are operating even today as worldwide known Braè's marble is trenching there. The other three important factors for Roman settlement seems to be aspect (south-west facing terrain), elevation and after that slope.

When the results (figure 2) were analysed in more detail it was realised that the reason for poor performance may have been in the archaeological data. What we have interpreted as a homogenous type of site called Roman settlements could have been in fact a mixture of several types of Roman settlement. Therefore it was decided to carry out more refined analyses of all Roman settlement sites used in the analyses in order to provide ideas on possible clusters of sites. The K-Means Cluster Analysis was employed and 3 distinguish clusters were identified, representing proneness to three different strategies of control over natural resources. We denoted them as: seaport type of Roman sites, quarry type of Roman sites and agriculture type (see Stanèiè and Veljanovski 2000).

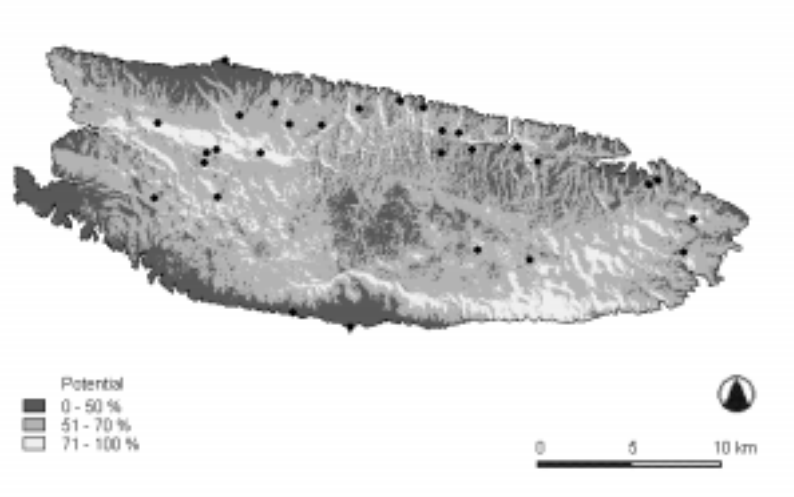

Figure 2: Archaeological predictive model for the Roman settlement sites on the island of Braè.

Resulting predictive model derived from the linear regression was not satisfactory mainly because the transparency was quite poor. Beside the lack of statistical significance, deficiency of the model is also that the insight into the impact of single variables observed in predictive model is somehow mist. However, the linear regression technique, applied to archaeological data, remains promising and powerful as a method to explore the relative relationships between variables affecting site location.

4. Practical application in archaeological resource management in Pomurje, Slovenia

It is important to emphasize that within the context of the Pomurje study, our goal could not be oriented towards the development of a model that would aid our understanding of the known distribution of the sites. Our aim was clearly designed - to develop a consistent model, which would assist in the process of highway planning (for more the details see Stanèiè et al. 2001).

The Slovenian national plan seeks to complete and link highways in a network across the country. Prolongation of the building programme affects the public purse detrimentally. Concern over costs led to an invitation to design a methodology for archaeological site prediction for a test section of a motorway corridor in the Pomurje region (along the Mura River), in northeastern Slovenia. The corridor measured approximately 11.2 km in length and archaeological prospection had already been carried out in the area. The results from archaeological fieldwork could therefore also be used for evaluation of the predictive methodology.

Instead of repeating all over background on predictive modelling methodology, we prefer to explain briefly what guided us in our choice of the method we after all implemented. Considerable problems often occur when the sample of sites analysed is small, and are particularly obvious when multivariate statistics are used (Stanèiè and Veljanovski 2000). Another issue is that most of the variables used are, naturally, based on the topographic characteristics of the area. Surface topography plays a special role in predictive models as a consequence of its association to past land use. As the Prekmurje case study area is almost flat, it would be difficult, and misleading, to extract any rules, or predict potential, solely on the basis of the relief-based distribution of the sites.

4.1 Variables for the model and methodology

With regard to the situation described above and the objective of the case study, complex methods based on multivariate statistics failed to suit our data. Therefore it was decided to apply a simple traditional method of Boolean overlay to create a predictive model of site location in the region. As deficient number of variables was a potential pitfall when performing analysis, we additionally investigated the potential of remotely sensed data to the case study.

A high resolution and accurate DEM is required for modelling settlement distribution in flat areas (Stanèiè and Oštir in press) and a DEM with a cell size of 25 by 25 metres, with a height accuracy of 3 m and a positional accuracy of 10 m was generated using radar images and radar interferometry. In addition to geological structure data and the soils data, a remotely sensed land use layer was created. Hydrology is also relevant. Given the case study goals (looking at current archaeological heritage), artificial ditches were not excluded from the hydrology vector layer, and these were considered "streams". Following this the distance to the nearest stream and distance to the main river within the region were calculated. Using these two layers, our aim was to enable an evaluation of the importance of the Mura River as a possible communication route.

Given that our study area is a large plain, differences observed in variables such as elevation, slope and aspect could never play significant roles. Hence we were primarily concerned the variables that could provide a description of the lowland itself and this we felt was likely to be obtained from remotely sensed data.

Besides land use and a vegetation index, original information relating to surface soil type was attained through the generation of mineral delineation layers. These indices are extensively used in mineral exploitation and vegetation analyses, essentially because they can indicate minute differences between various rock types and vegetation classes. In many cases, judiciously chosen indices can highlight and enhance differences which cannot be observed in the original colour bands displays from satellite imagery. Therefore we produced three layers suitable for GIS analyses representing Clay Minerals, Ferrous Minerals and Ferric Minerals (Iron Oxide). Actually, these indices represent the presence of minerals on the surface. In calculating these indices we hoped to obtain evidence for significant variation in the lowlands which was not be anticipated from the geomorphic variables.

Exploration and statistical analyses were carried out for five archaeological groups: prehistoric settlements (9), Roman and Early Medieval settlements (11), barrows (29), undated barrows (18) and isolated finds (mainly prehistoric axes, 15). Univariate statistical tests (Chi-square and the Student *t*-test) were used to test the correlation between site locations and each of the variables. The results of univariate statistical tests indicate that different site types in the study area tend to occur in different environmental settings; however, from a statistical point of view, the correlation was quite tentative. Nevertheless, all the variables that were initially examined for significance were later included in the model, if they fulfilled "threshold" conditions. As two scale types were involved, two criteria were introduced: a) for continuous variables, the interval between minimum and maximum values for sites was defined, and the circumscribed area considered as having potential, b) for nominal variables the area covered by classes associated with sites was treated as having potential. Twelve binary layers were prepared for each archaeological sample.

4.2 Weightening of the model

The initial model was designed as a simple sum of binary layers. Cells that fulfilled the entry condition for a particular variable were assigned a value of 1 and the rest a value of 0. As a result, we attained five models for different site types, with a potential ranging from 0 to 12. Evidently, the score was a combined measure of the potential of the cell. For the purposes of overall compliance, the predictive model for the region was then obtained by merging the five model outputs (for the modelling process see figure 3).

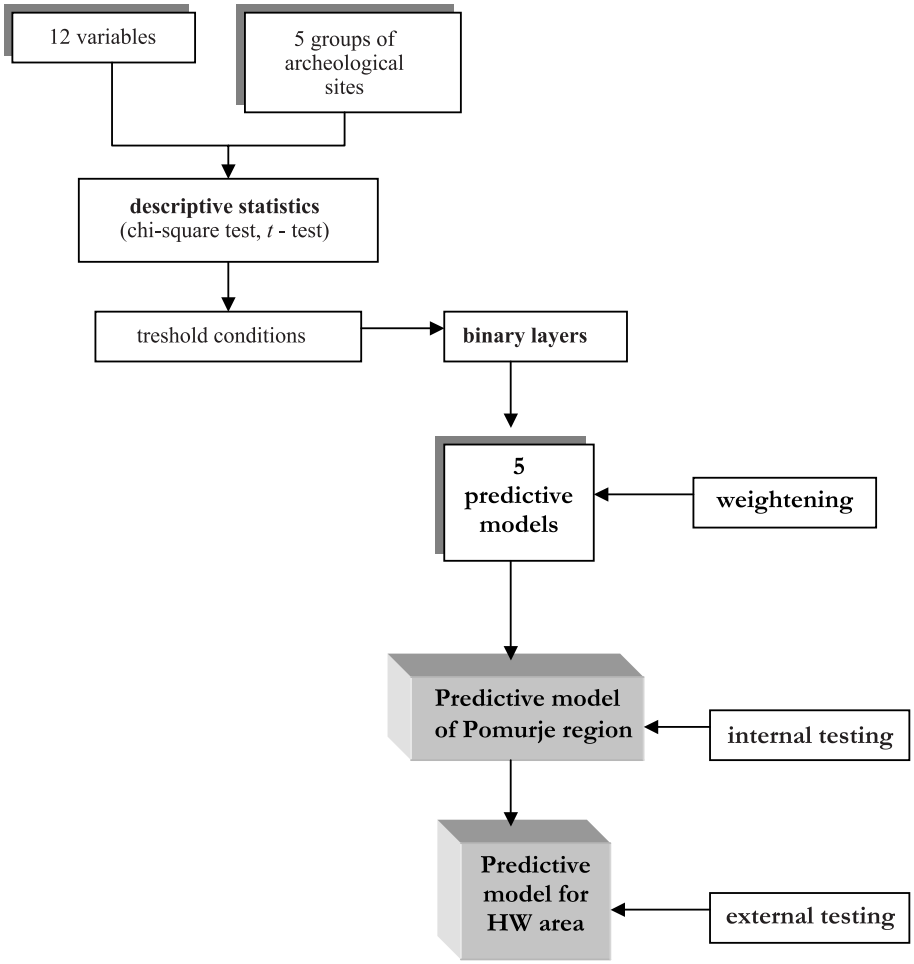

Figure 3: Steps in development of predictive model for the Prekmurje region.

Often when modelling archaeological potential the weights are applied, as there are rarely situations to claim how all the variables plays equally important. The weights are supposed to represent the relative relationship of locational factors within the past settlement system. As opposed to weights usually obtained through complex statistical examination, we believed that the weights most appropriate for a model of this kind should be based simply on the "sum of potential". Criteria were based on the assumption that a good predictor isolates highly specific and, therefore, a close or discrete signal in comparison to the values of the background. If a variable interpreted as a potential indicant also covered a large part of the study area then a smaller weight was assigned to that variable. Consequently, weights were applied in proportion to the observed power of the variable. As three mineral indices and one vegetation index seemed to be highly correlated we decided to give these additional weights. Each of their powers was diminished by a quarter and ultimately they represented – in their combined form – a single variable.

4.3 Validity of weighted model

The validity of the resulting weighted model was tested by measuring the accuracy of its predicted potentials using two different samples. The first was a sample of all archaeological sites that were used for the development of the model. This type of testing is known as internal testing and was carried out as a control of the process. Its results indicate that the model's prediction of potential was correct by about 81%, if "high potential" incorporates areas with potentials ranging from 70 to 100 (figure 4). The remaining 19% of sites fell within the limits of "medium potential" (40 to 70); no sites fell into the "low potential" range.

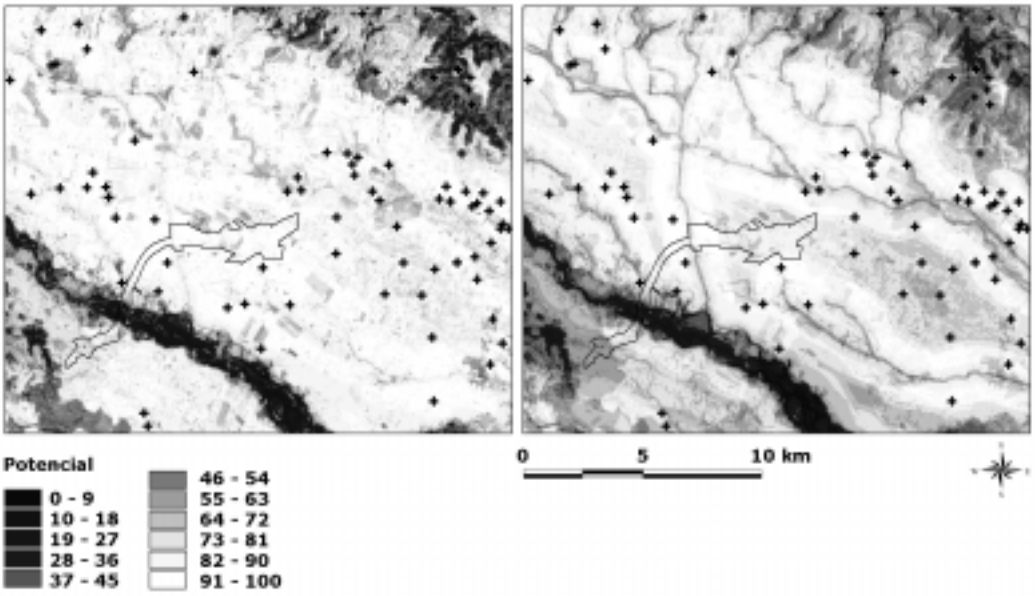

Figure 4: Archaeological predictive model for the Pomurje region: unweighted model (left) and weighted model (right) with the distribution of the sites upon which the model was developed.

The data used for secondary testing were obtained through archaeological survey in the area of the proposed highway corridor. The second subset of archaeological data was therefore an actual independent testing sample as the sample was gained from recent archaeological survey. This test indicates that the model's predictions are accurate. Although it is difficult to ascertain from the final model why particular sites within the region are situated within a specific area or why the potential of another site is lower, the model still distinguishes between locations where sites are most likely to be present and locations where they are not. The model does not, however, allow us to isolate the most important environmental or social determinants of site locations. In the end what we are providing to highway network planners is simply the expected distribution of archaeological resources. Although there are clearly problems in this process the predictive power of such a model should still be useful for highway network planners. Ultimately, the relatively poor performance of the model may simply result from the geographical conditions that are far from suitable for predictive modelling. The manner of determining threshold conditions and the need to generate five smaller predictive models for specific site types may also mitigate against the model's efficient performance.

5. Conclusions

According to the experiences gained through the case studies and work with different predictive techniques and approaches the following may be statable. The strength of predictive modelling is that the methodology itself enables exploration and evaluation of locational factors, what is mainly important in archaeological research. The knowledge of possible site location determinants can be tested and improved, whilst assisting interpretation from archaeological point of view greatly. Despite this, we must be aware that each of predictive modelling techniques has its advantages and disadvantages, and choosing suitable technique shall be assisted by clear purpose of modelling aims and the availability and quality of the data.

The highway case study encourages us that predictive modelling also has great practical potential. What we propose from applicative point is that archaeological predictive models can also support highway planning or other comprehensive interventions in a landscape. Integration of predictive models in highway planning is opportune in the early stages of planning (whilst selecting a highway corridor), here the results of predictive models provide immediate information for damage estimates to archaeological monuments. Secondly, predictive models can also be used during the construction stage, their main objective being integrate archaeological fieldwork methodologies with construction. Thirdly, results from predictive models can be used as a consolidated planning tool to define required fieldwork methodologies. Each of these propositions can play facilitate the efficient use of time, and promote cost efficient practice. Therefore, modelling in GIS environment still has big potentials for wide range of further applications.

References

Judge, W.J. and Sebastian, L. (eds.), 1988. Quantifying the present and predicting the past: Theory, method, and application of archaeological predictive modeling, U.S. Government Printing Office, Washington, D.C.

Kohler, T.A. and Parker, S.C., 1986. P*redictive Models for Archaeological Resource Location*. In: M.B. Schiffer (ed.), Advances in Archaeological Method and Theory 9, Orlando, Academic Press, pp. 397–452.

Stanèiè, Z. and Kvamme, K., 1999. Settlement Pattern Modelling through Boolean Overlays of Social and Environmental Variables. In Barcelo, J.A., Briz, I., Vila, A. (eds.), New Techniques for Old Times, CAA 98, BAR International Series 757: 231-237.

Stanèiè, Z. and Oštir, K., 1999 – in press. Producing Digital Elevation Models with Radar Interferometry. In Computer Applications and Quantitative Methods in Archaeology. Proceedings of the CAA conference, Dublin 1999.

Stanèiè, Z. and Veljanovski, T., 2000. Understanding Roman settlement patterns through multivariate statistics and predictive modelling. In Lock, G. (ed.), Beyond the map – Archaeology and Spatial Technologies. NATO Science Series, IOS Press: 147 –157.

Stanèiè, Z. Veljanovski, T., Oštir, K. and Podobnikar, T., 2001 – in press. Archaeological predictive modelling for highway construction planning. In Stanèiè, Z. and Veljanovski, T. (eds.), Computing Archaeology for Understanding the Past, CAA2000 Proceedings, Archaeopress, BAR International Series.

Stanèiè, Z., Vujnoviæ, N., Èaèe, S., Podobnikar, T. and Burmaz, J., 1999. The archaeological heritage of the island of Braè, Croatia. The Adriatic Project, Volume 2. Archaeopress, BAR International Series 803.

The value of spatial information – Decision-analytical assessment of a quality component

Sytze de Bruin & Arnold Bregt
Centre for Geo-Information, Wageningen University
P.O. Box 47, 6700 AA Wageningen, The Netherlands.
sytze.debruin@staff.girs.wag-ur.nl

Abstract

This paper proposes and illustrates a decision analytical approach to compare the value of alternative spatial data sets. Unlike most other work using probabilistic cost benefit analysis, its focus is on value of control. This is a useful concept when deciding upon the best spatial data set for applications where uncertainty is due to error in the reported data. By choosing from data sets, one can actually control the probability distribution of error. Application of the concept requires probabilistic accuracy measures and a loss function representing the cost of incorrect judgement about some target property. This is illustrated by an assessment of the suitability of two digital elevation models (DEMs) for determining the volume of sand required for building a container port. To demonstrate flexibility of the approach, DEM accuracy assessment was based on both a random and a systematic sample of error data, using design-based estimation and model-based prediction, i.e. geostatistics. Analysis results included the expected loss for each combination of DEM and sampling strategy. These indicated that both DEMs were about equally suitable for the intended use. Operational practicability of the method is highly dependent on the willingness of database producers to give access to sample information similar to e.g. the quick looks provided to potential users of remote sensing imagery.

1. Introduction

Despite advances in understanding components of data quality (e.g. Chrisman, 1995; Guptill and Morrison, 1995), almost no progress has been made in the development of methods to assess fitness for use (Veregin, 1999). This paper proposes a decision analytical approach to compare the value of alternative spatial data sets. Its focus is on a particular stage in a decision making process, i.e. the step where spatial information is used for selecting the best option from an existing set of alternatives (Calkins and Obermeyer, 1991). Our point of departure is the concept of value of information from the field of decision analysis (Raiffa and Schlaifer, 1961; Morgan and Henrion, 1990). Smith and Honeycutt (1987) provide an early example of using this concept in spatial analysis. It has been operationalised in software for the oil and gas exploration industry (e.g. Riis, 1999), where it is used to assist decision making as to whether to invest in additional site investigation (e.g. seismic shooting). Other applications concern, for example, sales management (Casimir, 1999), site investigation for hydrogeological design (James and Freeze, 1993; Forsyth, 1997) and human health risk management (Thompson and Evans, 1997; Lin *et al*., 1999).

Unlike the latter papers, which are mainly concerned with the economics of additional information about the outcome of random events, this one focuses on assessing and comparing the expected value of existing spatial data sets. By choosing an alternative data set, one can

actually control the (prior) probability distribution of error in required data. A relevant concept of data worth in this context is that of expected value of control (EVC), which is the value of being able to control how an uncertainty (e.g. error) resolves. This value is determined within the context of a particular decision, i.e. selecting an alternative (e.g. development option) from a set of available alternatives. The surplus value of a data set then equals the expected profit from using that data, in terms of an increase of the expected value of the decision outcome, or a decrease of expected loss. This approach is briefly explained below and it is illustrated using a simple case study.

2. Decision trees

In decision analysis, value of information is the expected desirability of reducing or eliminating uncertainty in a chance node of a decision tree. For a maximiser of expected value [$E(V)$], it is the difference between the $E(V)$s of acting with and without that information (e.g. Von Winterfeldt and Edwards, 1986; Morgan and Henrion, 1990). For example, consider a raffle with a prize worth •1000 (euro). A ticket costs •10 and the chance of winning is 0.005. Suppose that before buying a ticket someone offers to tell you unambiguously whether you will win or not. Should you ask for this clairvoyant's advice? If so, assume you would believe him and that your decision would fully depend on his tip. How much should you be willing to pay for it? The corresponding decision tree is shown in Fig. 1(a). Decision nodes (squares) indicate a point where a decision is to be made while chance nodes (circles) denote a random move of nature. Note that the upper and lower branches of the tree are alike, except for the sequence of decision and chance nodes. The expected value of the clairvoyant's perfect information (EVPI) equals the difference between the values of the leftmost nodes of the upper and lower branches, i.e. EVPI = •4.95.

Another situation arises if one could actually control the outcome of an uncertain event. For example, suppose that another clairvoyant offers to unmistakably tell you the winning number in the raffle and you are the first person to buy a ticket. Accordingly, the chance of winning given this information equals one [see Fig. 1(b)], resulting in an $E(V)$ of •990. This amount equals the expected value of control (EVC). It is easy to see that if, on the same event, the clairvoyant would only tell you the winning number to be odd (i.e. you are given partial control) his advice would be worth nothing.

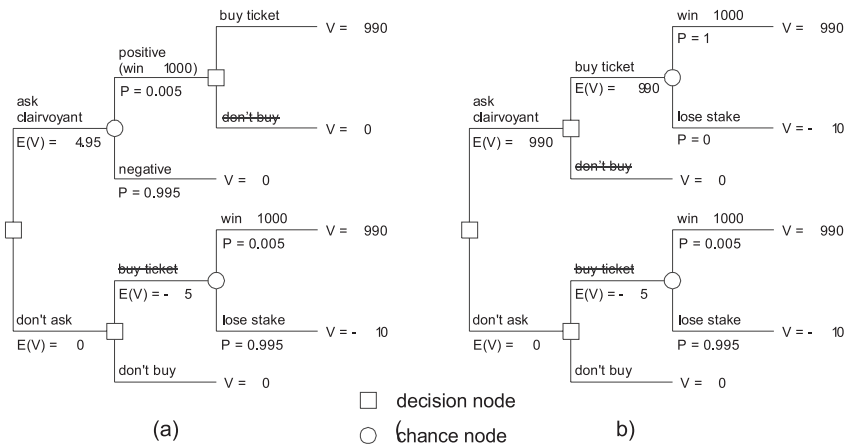

Figure 1. Decision trees illustrating the concepts of value of information (a) and value of control (b). The clairvoyant of (a) knows whether you will win when participating in the raffle. In case (b) he knows the number of the winning ticket.

The above examples concern non-spatial discrete decision problems. Yet, the approach is equally applicable in the spatial domain and may involve continuous decision variables having a continuum of choice alternatives. This is demonstrated in the following case study.

3. Case study: MTC-Valburg

3.1. MTC-Valburg project

The Multimodal Transport Centre (MTC) is a planned junction for the transhipment of containerised freight between different means of international transport (road, rail, and water). It is located near Valburg in the east of the Netherlands. One component of the MTC-Valburg project is the development of a container port. For this purpose, an area of approximately 41.2 ha situated on the river Waal (see Fig. 2) is to be raised to 15 m above NAP (Amsterdam Ordnance Datum). Our case study concerns the volume of sand required for raising the terrain. Computation of this volume calls for up-to-data elevation data, which is to be derived from an existing digital elevation model (DEM). Any incorrectly acquired volume of sand would lead to waste of resources (see below). We examined the fitness of two DEMs with grid spacings of 25m and 2m, which will be referred to as MANMADE and MERDEM respectively. Both DEMs were analysed as point rasters. Elevations in between raster points were obtained by bilinear interpolation.

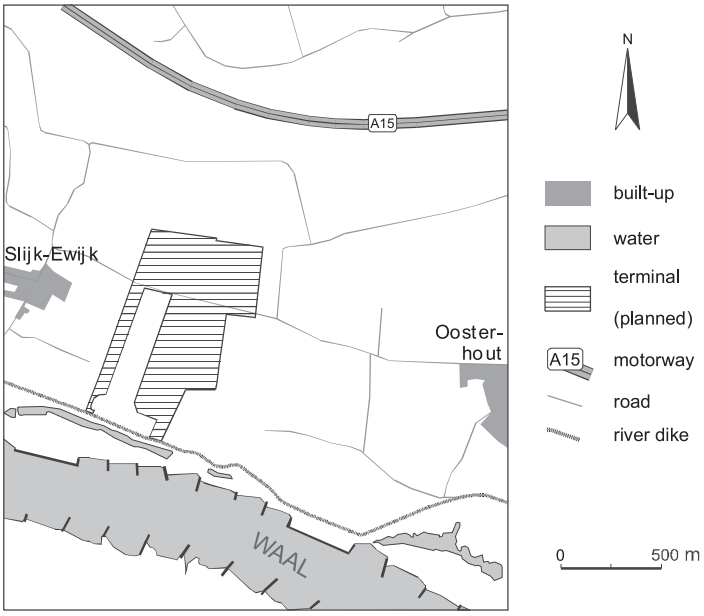

Figure 2. Location of the planned container port on the river Waal.

3.2. Expected loss

For the sake of this case study, it was assumed that DEM inaccuracies are the only source of error in the computed volume of sand required to raise the terrain (e.g. no subsidence). Thus, computations based on a perfect DEM would result in the exact volume and, therefore, zero loss. It was further assumed that resulting elevations being within project specifications, i.e. + 15 m NAP ± 2 cm, are left as is. The price of (locally extracted) sand was fixed at •4.55 m^{-3}. Elevations in excess of +15.02 m NAP or below + 14.98 m NAP would require additional road transport of sand to or from a local extraction site (costs: 2.6 km ´ •0.091 m^{-3} km^{-1}). Zero loss would correspond to the event that the determined volume of sand results in a mean elevation of exactly + 14.98 m NAP. The resulting loss function is given by Eq. (1):

$$L(d,x) = \begin{cases} -0.2366 \times (d-x) & \text{if } d - x \le 0 \\ 4.55 \times (d-x) & \text{if } 0 < d - x \le 16480\,\text{m}^3 \\ 74984 + 0.2366 \times [(d-x) - 16480] & \text{if } d - x > 16480\,\text{m}^3 \end{cases} \quad (1)$$

where d is a decision about the unknown required volume of sand (X), which is bounded by the project extent, the + 14.98 m NAP plane, and the mean DEM-reported elevation with the uncertain mean DEM error over the project area added (or subtracted) at the base. x denotes a realisation of X. The figure 16480 in Eq. (1) corresponds to the product of elevation tolerance (0.04 m) and the project area (4.12 ´ 10^5 m^2). Note that outside the tolerance interval the sand price is not considered. That is because the project should always meet its specifications and because in case of an excess or deficiency, elevation is assumed to be brought to the nearest bound. Any deficit or surplus volume of sand is bought or sold at the same price of • 4.55 m^{-3}.

A rational decision about the required volume of sand is the volume that minimises the statistical expectation of Eq. (1), denoted by $E[L(d, x)]$. This volume and the corresponding loss depend upon the accuracy of the employed DEM within the project area. The difference in $E[L(d, x)]$ resulting from using either MANMADE or MERDEM is a measure of the surplus value of the one DEM with respect to the other. This value (of control) can be determined using an approach similar to that used for the raffle of Fig 1(b), with the difference that the probability space and decision space are now continuous. Computation of $E[L(d, x)]$ thus requires integrating a loss function [Eq. (1)] over a continuous probability distribution function. This is commonly done by discrete approximation (Goovaerts, 1997):

$$\begin{aligned} E[L(d,x)\,|\cdot] &= \int_{-\infty}^{+\infty} L(d,x)\,dF(x|\cdot) \\ &\approx \sum_{k=1}^{K+1} L(d,\bar{x}_k) \times [F(x_k|\cdot) - F(x_{k-1}|\cdot)] \end{aligned} \quad (2)$$

with $F(x \mid \times)$ being a conditional cumulative distribution function of the uncertain X, x_k, $k = 1, ..., K$ are threshold values discretising the range of variation of x-values, and is the mean of the half open interval $(x_{k-1}, x_k]$.

3.3. DEM accuracy

Computation of $E[L(d, x)]$ by Eq. (2) does not require knowledge of absolute values of x. It suffices to assess differences $d - x$. These can be computed from volumetric error, say e, as $d - x = d_u - e$, where d_u is an uncorrected decision with an offset equal to the DEM-derived volume. The cumulative distribution of e can be estimated by comparing the DEMs to a series of reference elevations at sample spots. As this procedure involves averaging over many locations, the distribution of e can be approximated by a normal distribution (according to the central limit theorem) with parameters *mean volumetric error* and *standard deviation*. Note that these parameters should bear on the spatial support of the project area. Therefore, knowledge of e.g. just the RMSE of DEM elevation at point locations is insufficient.

In the following, two scenarios are considered: one (scenario 1) in which the DEM customer has access to a random sample of 50 reference elevations and another (scenario 2) in which he gets a systematic sample of size 87. The 50 sample locations were obtained by simple random sampling within the planned port area. The 87 systematically sampled points were on the nodes of a 200 m square grid covering a larger area. Only 12 of these points were situated within the project area; the other 75 fell outside that area (see Fig. 3). Elevation data were

obtained by Real Time Kinematic (RTK) GPS survey. Any discrepancy between DEM and reference elevation was recorded as DEM error, i.e. errors in the GPS survey were ignored.

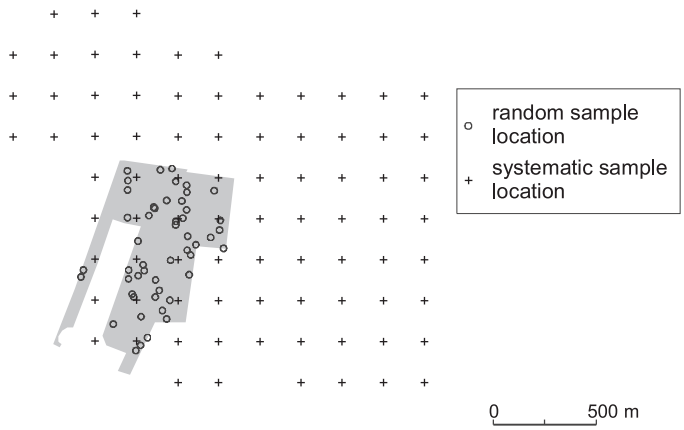

Figure 3. Random (scenario 1) and systematic (scenario 2) GPS-measurement locations.

Approaches for statistical inference from spatial sample data differ fundamentally between those based on classical sampling theory and those based on stochastic models of spatial variation (see e.g. Brus and De Gruijter, 1997). The former are referred to as design-based strategies, and the latter as model-based strategies. Design-based strategies do not depend on any assumptions about the spatial continuity of the attribute of interest. Independence of the sample data is induced by randomisation at the sampling stage. As a consequence, design based strategies provide objective estimates of the spatial mean and variance (or standard deviation) of a quantitative attribute. In case of the simple random sample (scenario 1), the mean and its standard deviation can be estimated simply by:

$$\check{\delta}e = 4.12 \times 10^5 \times \bar{z} = 4.12 \times 10^5 \times \frac{1}{n}\sum_{i=1}^{n} z_i \qquad (3) \qquad [\mathrm{m}^3]$$

and

$$s(\check{\delta}e) = 4.12 \times 10^5 \times s(\bar{z}) = 4.12 \times 10^5 \times \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(z_i - \bar{z})^2} \qquad (4) \qquad [\mathrm{m}^3]$$

where z_i is the difference between DEM-elevation and reference elevation at location i and n is the sample size (n = 50). The figure 4.12 ´ 10^5 in Eqs. (3, 4) equals the area of the planned container port.

Scenario 2 requires a model-based approach. In this case, mean DEM error over the project area and the standard deviation of mean error were predicted using ordinary block kriging. See e.g. Isaaks and Srivastava (1989) or Goovaerts (1997) for the equations. The approach requires making subjective decisions; e.g. about how to model spatial continuity of DEM error. To approximate the shape of the project area, it was discretised into a block composed of 44 points. The actual block kriging was done using Gstat software (Pebesma, 2000), which allows for arbitrarily shaped blocks. The choice between design-based and model-based strategies is extensively discussed by Brus and De Gruijter (1997). Goovaerts (1999) gives a practical overview of some tools for statistical inference with both strategies.

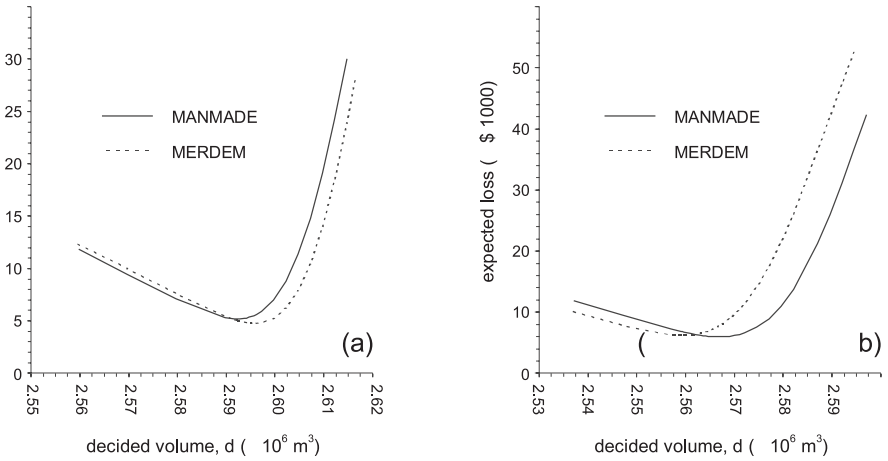

Figure 4. Expected loss due to DEM error in case of design-based (a) and model-based (b) uncertainty assessment.

4. Results

Fig. 4 shows graphs of the expected financial loss associated with decisions based on each of the DEMs for design-based and model-based uncertainty assessment. The optimum (Bayes') decision of each scenario can be read from Fig. 4 as the point where minimum expected loss (i.e. Bayes' loss) is achieved. The graphs [Figs. 4(a, b)] employ common *x*-axes for the curves of MANMADE and MERDEM. This has been achieved by using information on the mean DEM elevation over the whole project area. This information can only be obtained if one has full access to the DEM elevations (not just error at sample sites). It may be questioned whether such info will be available prior to actually buying a data set. Yet, it is not required for computation of expected losses and derived quantities.

Table 1 summarises the Bayes' loss of each situation. Several quantities expressing DEM fitness in the MTC-Valburg context can be derived from the table entries. In case of design-based uncertainty assessment (scenario 1), the expected value of control of MERDEM over MANMADE equals •5183 - •4776 = •407. Thus, MERDEM is expected to be the best buy if it is less than •407 more expensive than MANMADE. In case of model-based assessment (scenario 2) the EVC attains a negative value (- •202), i.e. using MERDEM is expected to result in •202 more loss. In either case, the differences between the expectations are rather small, indicating that both DEMs are about equally suitable for the application at hand. The EVC of a perfect DEM (resulting in zero loss) in excess of MANMADE was estimated at •5183 or •5966, depending on the method of uncertainty assessment. These figures are but a fraction of what it would cost to generate such a DEM.

Assessment method		Digital Elevation Model	
(scenario)		MANMADE	MERDEM
Design-based	(1)	• 5183	• 4776
Model-based	(2)	• 5966	• 6168

5. Discussion

Increasingly, spatial data customers may choose from several data sets providing similar thematic information but at different prices and accuracy levels. The customer is responsible for assessing whether a data set meets the needs of a particular application (fitness for use), but it is the data producer's task to provide the required quality documentation (Veregin, 1999).

This paper introduced the concept of expected value of control as a measure of fitness for use in spatial decision making. Application of the concept requires probabilistic uncertainty measures and a loss function representing the cost of incorrect judgement about some target property. Uncertainty assessment, in this context, concerns comparison of the data sets with reference data as required by the application at hand. As it is infeasible to have an exhaustive reference model, uncertainty estimates need to be inferred from sparse sample data. The problem that arises is how to obtain a sample before actually purchasing the data, because it is at that point that fitness for use assessment really makes sense.

In the remote sensing community it is now common practice to provide potential users of satellite imagery with access to so called quick looks. These sub-sampled images allow examination of cloud cover percentages and comparison of the utility of several products. A similar approach may be set up for other spatial data sets. For example, Fisher (1998) argued that data providers should deliver high precision spot heights as part of their DEM products. Why not give access to a sample of error vectors for evaluation purposes prior to the purchase? Alternatively, data suppliers could provide a sample of their product for the potential client to compare it with a set of collocated reference data.

The case study demonstrated that expected loss and derived quantities could be computed without all the data-detail required for assessments about the ultimate decision variable, i.e. volume of sand. For the latter purpose, a complete data set would be needed. In other words, providing sample data does not automatically imply loss of sales, unless fitness of the offered data set is inferior to that of alternative products.

If sample data is provided by a database producer, the customer may not have control over the sampling design. Consequently, he or she must adapt the inference mechanism for obtaining error estimates to the design available. Although design-based strategies warrant objectivity and validity of the results, they can only be applied if the sample locations are selected by (restricted) randomisation. Model-based approaches are more versatile, but at the price of full dependence on a postulated stochastic model. Several more aspects need to be considered when choosing between model-based and design-based approaches (see Brus and De Guijter, 1997). Our case study included both strategies. It was not intended to demonstrate the consequences of choice, e.g. in terms of performance, but rather to show the possibility of choice. Such flexibility is essential to make fitness for use assessment operationally practicable.

References

Brus, D.J., and De Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). Geoderma, 80, 1-59.

Calkins, H.W., and Obermeyer, N.J., 1991, Taxonomy for surveying the use and value of geographical information. International Journal of Geographical Information Systems, 5, 341-351.

Casimir, R.J., 1999. Strategies for a blind newsboy. Omega, 27, 129-134.

Chrisman, N.R., 1995. Living with error in geographic data: Truth and resposibility. In: Proceedings Ninth Annual Symposium on Geographic Information Systems in Natural Resources Management (Vancouver: GIS World). pp. 12-17.

Fisher, P.F., 1998. Improved modeling of elevation error with geostatistics. GeoInformatica, 2, 215-233.

Forsyth, M.C.I., 1997. The economics of site investigation for groundwater protection: sequential decision making under uncertainty. Journal of Environmental Economics and Management, 34, 1-31.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation (New York: Oxford University Press).

Goovaerts, P., 1999. Geostatistical tools for deriving block-averaged values for environmental attributes. Geographic Information Sciences, 5, 88-96.

Guptill S.C., and Morrison, J.L., 1995. Elements of Spatial Data Quality (New York: Elsevier Science).

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

James, B.R., and Freeze, R.A., 1993. The worth of data in predicting aquitard continuity in hydrogeological design. Water Resources Research, 29, 2049-2065.

Lin, C., Gelman, A., Price, P.N., and Krantz, D.H., 1999. Analysis of local decisions using hierarchical modeling, applied to home radon measurement and remediation. Statistical Science, 14, 305-337.

Morgan, M.G., and Henrion, M., 1990. Uncertainty: A Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis (Cambridge: Cambridge University Press).

Pebesma, E.J., 2000. Gstat user's manual. (Utrecht: Department of Physical Geography, Utrecht University), http://www.geog.uu.nl/gstat/.

Raiffa, H., and Schlaifer, R., 1961. Applied Statistical Decision Theory (Boston: Harvard University).

Riis, T., 1999. Quantifying the value of information. Petroleum Engineer International, June 1999, 48-50. Also available at:

http://www.caesarsystems.com/Technica/Value_of_Info/valueof.htm.

Smith, G.R., and Honeycutt, D.M., 1987. Geographic data uncertainty, decision making and the value of information. In: GIS '87, Proceedings 2nd International Conference, San Francisco, October 26-30, 1987 (Falls Church, VA: ASPRS/ACSMUSA), pp. 300-312.

Thompson, K.M., and Evans, J.S., 1997. The value of improved national exposure information for perchloroethylene (perc): a case study for dry cleaners. Risk Analysis, 17, 253-271.

Veregin, H., 1999. Data quality parameters. In: Geographical Information Systems - Volume 1, Principles and Technical Issues, 2nd Edition, edited by P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind (New York: John Wiley & Sons), pp. 177-189.

Von Winterfeldt, D., and Edwards, W. 1986. Decision Analysis and Behavioral Research (Cambridge: Cambridge University Press).

Finding Analogous Structures in Cartographic Data

Diarmuid O'Donoghue, Adam Winstanley
Department of Computer Science,
National University of Ireland, Maynooth
Co. Kildare, Ireland
{diarmuid.odonoghue; adam.winstanley}@may.ie

Abstract

In this paper we describe the application of analogical structure matching techniques to the domain of Geographic Information Systems (GIS). Automatic categorisation of large-scale topographic data into roads, buildings etc. *can be based on isolated objects. We describe how identifying analogous clusters of objects can categorise ambiguous polygons by introducing context into the categorisation process. We describe a number of cartographic classification tasks that can be performed by analogical structure matching, resulting in a useful classification tool that operates in a cognitively plausible manner.*

1. Introduction

Manually recorded large-scale (1:1250 and 1:2500) topographic data consists primarily of boundary definitions, where lines combine to form polygons enclosing areas of land. But these "line drawings" are of limited usefulness to both cartographers and the general public. Classifying individual polygons into feature types such as buildings, roads, made-land, or water *etc.* vastly increases the usefulness to these data in Geographic Information Systems (GIS), but is expensive to perform manually on terabytes of cartographic data.

Automatic classification of geometric topographical data into object types (and/or feature codes) can be partially accomplished through isolated-shape recognition [Keyes & Winstanley, 2000], by focusing on parameters of individual object such as total area, boundary length, and shape. Performance can be improved by extending the classification mechanism with contextual information. This improves the accuracy of automatic classification, because we can frequently resolve ambiguous data by examining its context to provide evidence for category membership (thus informally, for example, we can say that a square on a map is more likely to depict a house if it is adjacent to a road).

We describe a method of matching clusters of topographical objects against a known prototype cluster by identifying *analogous structures*, and in this way we automatically infer the identify of unclassified polygons. An analogy is a comparison between a well known *source* and a problem *target.* The source acts as a predictor for the target, because the source supports inference about that target. The ability of analogical comparisons to support inference is the prime reason for our use of the analogy process to perform topographical classification. Our categorisation technique involves reasoning with collections of objects, and thus the reliability of our category assignment process improves by incorporating neighbourhood data in the classification.

A variety of cognitive studies have been carried out to ascertain the exact nature of the analogy process. A widely studied example [Duncker, 1945] concerns the problem of treating a

patient suffering from an inoperable tumour. One set of subjects trying to solve this problem are given the source domain of a country ruled by an evil dictator and whose fortress can only be reached by sending troops down multiple roads to simultaneously converge on the fortress thereby overwhelming it. Subjects that do not receive the “fortress” information have a 10% solution rate, while 80% of subjects that receive the fortress domain give the required convergence solution [Gick and Holyoak, 1980]. Thus, analogies have a profound effect on people’s ability to generate inferences and solve problems. Inferring the category of unclassified topographic objects is just one of many domains that may be solved using analogical comparison.

Classifying unknown topographical objects through structure matching requires two complimentary tasks. The central, or core, activity concerns generating the largest possible mapping between the problem data the some pre-stored prototype. Since Gentner [1983] identified that analogies are built on an implicit parallelism between two information structures, computational modelling of the analogy process has been the focus of much work. This work has largely focused on creating more efficient algorithms for identifying the largest structure mapping - including [Falkenhainer, Forbus and Gentner, 1989; Keane, and Brayshaw 1988; Veale, O’Donoghue and Keane, 1999; Salvucci and Anderson, 2001].

The second activity revolves around determining the boundaries between this problem data and “irrelevant” background information. This requires identifying clusters of information that can be (temporarily) isolated from the remainder of the map data, in order to allow the core mapping process to proceed. Category prototypes play a significant role in boundary identification, with the efficiency of the matching process being reliant on domain selection. This retrieval activity is a necessary precursor to the matching process and has received comparatively little attention - see [Forbus, Gentner and Law, 1994; Plate, 1998; Crean and O’Donoghue, 2001].

2. Geometric Analogies

In this paper we focus on comparisons between sets of geometric objects, broadly similar to those comparisons found in IQ tests [Evans, 1968; Bohan and O’Donoghue, 2000]. These have the structure A is-to B as C is-to an unknown D. (If a square within a circle (A) changes to a striped square within a circle(B), what do we do with a triangle within a square (C)?). Solving geometric analogies is founded upon identifying a matching on the information structures between A and C. It is only by comparing the relationships between objects (and *not* comparing similar objects themselves) that we see the square (from A) and the triangle (from C) play the same role in Figure 1. It is the information structure and the relationships between objects that dictate that the square and the triangle are matching objects in this problem. Thus, the solution will involve a small striped triangle. Were we to compare similar objects, we would align the two squares and we would end up with a large striped square - which is clearly incorrect! Algorithms to perform the structure-matching task are more complex, and are based on analyses of and comparisons between information structures - rather than on the content of this information itself.

Figure 1.: A Simple Geometric Analogy problem

Computational modelling of the geometric analogy process can be achieved in two steps. Firstly, (find and) represent the known information about both problem domains using predicate calculus assertions. Secondly, find the largest isomorphism between the two sets of data [Gentner, 1983; Veale, O'Donoghue and Keane, 1999]. For the problem in Figure 1, domain A might be represented as "`contains (circle, square)`" while C might be "`contains (square, triangle)`". Aligning these predicate structures will identify the *mapping* indicated by Figure 2. Solving the above geometric analogy is built upon identifying this inter-domain mapping (between A and C).

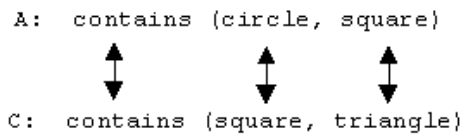

Figure 2 : A predicate mapping identified on structural similarity

Next we identify the attribute transformation [Bohan and O'Donoghue, 2000] that occurs in the source domain, changing between parts A and B above and denoted (A › B). For this we need only note the attribute alterations that must be applied. In the previous example we note that the square changes from plain to striped. (Later, this shall correspond to identifying the required classification for an unclassified topographic object). We apply the attribute transformation (A›B; read, A changes-to B) to the equivalent objects in C. (So, the square maps onto the triangle, therefore if the square becomes striped then so must the triangle.) The predicate information does not change in this problem, as it does not for topographical classification problems. While the general process of analogy is far more complex, we shall not analyse it in further detail here.

2.1 Domain Representation

Before the analogy process can be applied to the problem described, we must first represent the problem information in a suitable manner. This involves quite an amount of analysis of the underlying topographic data, and shall be discussed later. We identify two relationships between topographic objects.

i) `adjacent (a, b)` indicates that two polygons share a common boundary.

ii) `point-adjacent (a, b)` indicates that the adjacent two polygons meet only at a single point.

Generally these two sets of information are mutually exclusive, although this may not be the case, for example, when two objects touch at multiple locations. It is these two simple relations that form the basis of our structure matching process which in turn drives the category assignment procedure. A typical topographic domain that consists of say 5 topographic objects may contain, approximately, five of the first predicate and two examples of the second predicate. The analogy process takes this problem information and applies the analogy process to this information.

2.2 Analogical Comparisons in Topographic data

Applying analogy to topographic data is even simpler than the geometric analogies previously described. First let us consider the problem of categorizing a single unclassified polygon contained within a cluster of otherwise categorized polygons. Let us also assume that the correct source domain has been selected for use with the given target problem. Finding the

correct classification for the unclassified polygon requires two simple steps.

i) Identify the largest possible structure matching between the problem and the solution template.

ii) Find the object that maps with the unclassified polygon, and apply the class of that poly gon to the unclassified polygon.

As with the analogy process itself, the new classification is derived by a process of pattern completion applied to the inter-domain mapping. So, if the unclassified object matches with an “unmade land” polygon, then the unclassified polygon also assumes the classification of “unmade land”. Pattern completion itself is a relatively straightforward algorithm, but its simplicity belies the fact that it must be only applied after the pre-requisite processes have been carried out.

Of course the usefulness of this categorization technique is completely reliant on the applicability of the identified source domain. This has two implications. First, great care must be taken in constructing the store of candidate source domains to ensure that the inferences mandated by each comparison are valid. Secondly, we need a reliable technique to retrieve the most appropriate domain from the stock of candidate source domains. Retrieval is currently initiated by identification of a target problem containing a single unclassified polygon. We use an attribute based retrieval scheme to select the most appropriate source domain. This retrieves the most similar source domain that contains not just polygons of the required types, but also in the required configuration. However, in this paper we focus on the use of polygon attributes to effect retrieval. This causes retrieval of a similar source domain with the same contents, which differing by the classification of a single polygon. For a description on structure based retrieval see [Crean and O’Donoghue, 2001].

3. System Architecture

Having described the central theory of category assignment by structure matching, we now describe a software realisation of this system. This technique requires that the underlying geographical information be already at least partly classified. This is because structure without some content information (*i.e.* classification) is too vague to identify categorisation for any but the simplest of cases. For that reason this structure matching approach is ideally applied after a partial classification has already been achieved. Figure 3 illustrates the basic system architecture.

The first process involves selecting suitable collections of polygon information to be passed on to the analogy process itself, and this involves much more than ensuring only one unclassified polygon is included in each of these clusters. The basic cluster of polygon information used in this project is referred to a *locality*, consisting of a root polygon, all adjacent polygons and all adjacency information between these polygons. Localities are the basic unit of process for all subsequent structure matching activities. A number of factors motivated this choice of structural primitive. Firstly, structure matching is an NP-complete problem, and thus using small information domains is considerably more efficient than large domains. Even though large domains may support more robust classification, the resultant computational expense would be too severe for practical application on a typical desktop computer. Secondly, localities include sufficient information for a variety of ambiguous classification tasks - including error detection as well as classification. Third, explicit storage of one locality can expedite computation for adjacent localities. Fourth, localities offer the possibility of easily including more detailed information on individual polygons with structure information at some later date.

Separating locality identification from the structure matching process itself has some additional benefits. Under the described architecture, we explicitly store each locality and additionally we index each locality according to the category of each polygon contained within that locality. Though there is a large degree of overlap between adjacent localities, explicit storage helps expedite the subsequent matching process and the locality identification process itself. Before proceeding to structure matching itself, we see how certain types of localities support categorisation without applying computationally expensive structure matching algorithms.

More importantly, a very significant classification advantage emanates from explicit storage and indexing of locality information. This advantage concerns classification polygons with *flexible structure matching*. This classification process extends the power of the matching process by allowing a number of problems with differing structure to all match the same source domain. For example, a number of structures can be identified as erroneous without recourse to detailed structure matching. Regardless of the structure of certain localities, we may identify the error based purely on the contents of the locality itself. Such errors can be identified by direct examination of the locality information, and this process is referred to Adjacency Matrix Classification - see Figure 3.

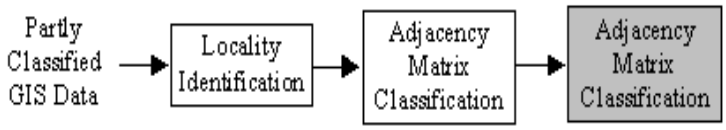

Figure 3 : System Architecture

Clearly, one cannot have a “waterway” polygon completely enclosed within a polygon classified as a building - regardless of the structure of these or any other polygons with the locality. A house located entirely within another house can be flagged as a probable error, and this surprisingly accounts for several errors detected in published topographic data. This technique offers the unique capability of altering the original map data - this is vital for extending existing boundaries to ensure feature codes don’t *run* into adjacent objects, causing misclassification.

We see two different uses of this structure matching approach to topographical classification: first for completing classification of partly classified polygons, and secondly for correcting certain categories of classification error. Examples of both uses are described later in this paper.

A major problem with topographic data is known as “category bleeding”, where the categorisation of one object bleeds into attached objects because not all objects perfectly enclose an area due to data capture errors. This results in two polygons being connected by a narrow neck of land - often invisible to the cartographer. We see the identification of known localities as providing a potential solution. By recognising a known structure, particularly a problematic one, we are often able to detect the error and even offer the ability to extend one boundary to stop this category bleeding.

4. Detecting Adjacency Errors

In this section we examine in detail two applications of the flexible structure matching technique. These examples illustrate the technique used in identifying (pre-existing) classification errors - this activity being a necessary precursor to reliability estimation and to error correction.

Consider the illegal-prototype representing “no building may directly and completely enclose another building” - though intuitively obvious, this simple rule cannot be represented by isolated object identification. The flexible structure matching technique identifies (amongst others) the following misclassification. In figure 4 (below) the lighter colored objects identify building, while the darker objects represent any other categorization.

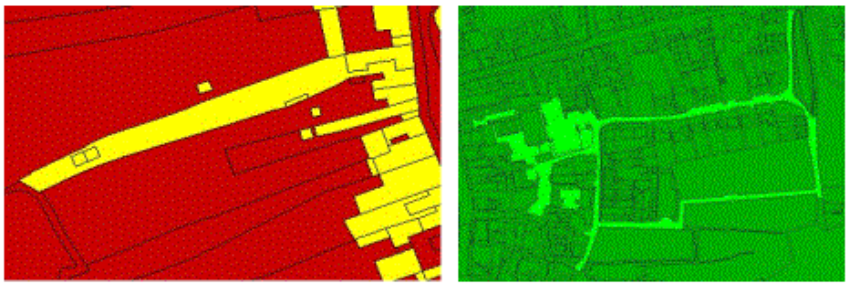

Figure 4: Examples of the “Building within building” error

Another illegal neighborhood that is identifiable from the adjacency matrix is the “no road may be completely detached from all other roads”. This effectively enforces the rule that all roads must be attached to other roads - hence a road’s usefulness. This template identifies the error, which can be easily corrected by isolated shape-recognition techniques.

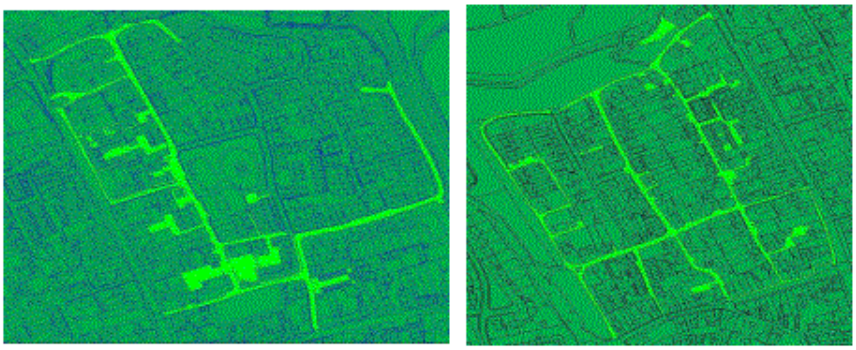

Figure 5 : Examples of the “Isolated road” error

5. Structure Matching

Detailed structure matching is more expensive to perform, but is more generally applicable to polygon classification. Structure matching allows us to infer and assign a classification to unclassified polygons. The matching process identifies a 1-to-1 correspondence between some problem structure (containing an unidentified polygon) and a similar template. Given the identical structures and sufficient similarity in each polygon’s categorisation, we may infer the identity of the unclassified polygon.

Figure 6 illustrates various templates for a buildings. Again we define (a number of) template structures for each class of object we wish to classify. The dots in this diagram indicate that neighbouring polygons are adjoined at one point rather than being joined by a line - thus diagonally adjoining polygons are included in a locality.

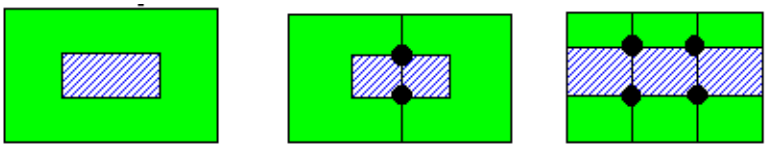

Figure 6 : Simple Geographic objects

We represent all the *adjoins* relationships in the chosen locality. The incremental matching algorithm [Keane et al., 1994] identifies the structural isomorphism between the problem and template. If an isomorphism exists, wherein every problem object is matched to one (and only one) template object - and if the same juxtapositions exist in each domain, then we have a structural match between domains. This is the first part of our requirements.

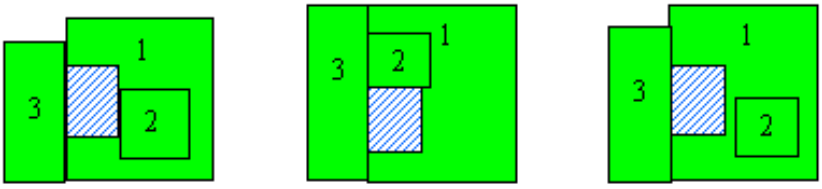

Figure 7 : Labelled Geographic localities

The second requirement is that matching objects are of the same categories; this requirement obviously does not apply to the problem polygon, which must be unclassified. Given the structural match, and the subsequent category matching, we infer the class of the unclassified polygon.

The structure matching algorithm for matching topographic structures is based on the *incremental mapping model* [Keane et al, 1994]. However, matching is this domain presents some unique problems. Firstly, each domain represents adjacency information between locality objects. Structure matching domains described entirely with commutative predicates introduces problems of structural ambiguity - because `adjacent(a,b)` may also be written `adjacent(b,a)` and the matching algorithm must be aware of this problem. Secondly, we need to integrate attribute information into the matching process - this requires a fundamental extension to the theory of Gentner [1983]. Finally, different source domains must be applied to solve different problem structures. Thus our mapping process must include a retrieval phase to select the most appropriate template. However, we shall not describe our algorithm further. The significant factor here is that it identifies the largest possible 1-to-1 correspondence between the problem and the template data.

Results are not yet available for detailed structure matching, but initial results are very promising. The extra precision provided by detailed structure matching brings extra refinement to the classification process, and thus many ambiguous objects are categorised by virtue of the classes of the surrounding objects.

6. Conclusion

Current techniques for automatically classifying topographical data are based on analysis of isolated objects - and thus cannot process ambiguous polygons. We described the technique of structure matching and how it may sustain topographical object classification. We also showed how contextual information can be used to identify topographical classification errors, even without detailed information on individual objects. This is achieved by matching object neighbourhoods against known templates - representing problematic clusters of topographic information.

Acknowledgements

We would like to thank the Ordnance Survey (Southampton) for access to samples of their topographic database and for their support for this project.

References

Bohan, A. and O'Donoghue, D. 2000, "LUDI: A Model for Geometric Analogies using Attribute Matching", *AICS-2000 11th Artificial Intelligence and Cognitive Science Conference*, Aug. 23-25, NUI Galway, Ireland.

Crean B. P. and O'Donoghue, D. "Features of Structure for Analogy Retrieval", *Applied Informatics 2001*, Innsbruck, Austria.

Duncker, K. "On problem solving", *Psychological Monographs*, 58 (whole, no. 270), 1945.

Evans, T. G. 1968, "A program for the solution of a class of geometric-analogy intelligence test questions", in *Semantic Information Processing*, (Ed.) M. Minsky, MIT Press.

Falkenhainer, B. Forbus, K. and Gentner, D. 1989 "The Structure Mapping Engine: Algorithm and Examples", *Artificial Intelligence*, 41, 1-63.

Forbus, K., Gentner, D. and Law, K. "Simulating Similarity-Based Retrieval: A Comparison of ARCS and MAC/FAC", *Proc. 14th Cognitive Science Society*, 543-548, 1994.

Gentner, D. 1983 "Structure-Mapping: A Theoretical Framework for Analogy", *Cognitive Science*, 7, 155-170.

Gick, M. and Holyoak, K. 1980, "Analogical Problem Solving", *Cognitive Psychology*, 12, 306-355.

Keane, M. T. and Brayshaw, M. 1988, "Indirect Analogical Mapping: A Computational Model of Analogy", in *Third European Working Session on Machine Learning.* Ed. D. Sleeman, London Pitman,.

Keane, M. T., Ledgeway, T. and Duff, S. (1994). "Constraints on analogical mapping: A comparison of three models." *Cognitive Science*, 18, 387-438.

Keyes L. and Winstanley, A.C. 2000, "Applying Computer Vision Techniques to Topographic Objects", *XIXth International Archives of Photogrammetry and Remote Sensing*, 33 (B3), 480-487.

Plate T. "Structured operations with Distributed Vector Representations", in *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational and Neural Sciences*, New Bulgarian University, Sofia, Bulgaria, July, 1998.

Salvucci D. D. and Anderson J. R. 2001, "Integrating Analogical Mapping and general problem solving: the path-mapping approach", *Cognitive Science*, 25, 67-110.

Veale, T. O'Donoghue, D. and Keane, M. T. 1999, "Computability as a limiting cognitive constraint: complexity concerns in metaphor comprehension about which cognitive linguists should be aware", in *Cultural, Psychological and Typological Issues in Cognitive Linguistics*, Ed. M. Hiraga, C. Sinha and Wilcox., S. John Benjamins Publ. Amsterdam/Philadelphia. pp 129-155 - ISBN: 90 272 36569.