Cartographic Knowledge Acquisition: From Practice to Theory

Kazemi, S., [1*] and Forghani, A.[2]
[1]Faculty of Education, Science, Technology and Mathematics, University of Canberra (UC), Bruce ACT 2601 Australia, E-mail: sharon_kazemi@hotmail.com, sharon.kazemi@canberra.edu.au
[2]School of Civil and Environmental Engineering, The University of New South Wales (UNSW), Kensington Sydney NSW 2033 Australia, E-mail: a.forghani@unsw.edu.au, alanforghani@yahoo.com.au
[*]*Corresponding Author*

ABSTRACT
This paper firstly provides a brief overview of expert systems and their application in Geographic Information Systems (GIS) with a particular emphasis on cartographic knowledge capture. Later, the paper describes the process of a heuristic natural knowledge transfer from cartographers for building an artificial intelligence system using an international Cartographic Generalisation Survey at several national and state mapping agencies as well as a number of geospatial software vendors, the survey results are utilised for defining the basis for a knowledge-based expert system.

1. Introduction

This paper firstly provides a brief overview of expert systems and their application in Geographic Information Systems (GIS) with a particular emphasis on cartographic knowledge capture. An expert system consists of four main components: (a) knowledge acquisition, (b) inference engine, (c) knowledge representation, and (d) user interface (Forghani, 1997). Expert systems have played an important role in automatic generalisation in different cartographic software applications. These are briefly discussed in Section 2. Key purpose of conducting this research is to capture the cartographers' knowledge about the principles of cartographic generalisation and their experience with existing generalisation software (Section 3). The survey results are utilised for defining the basis for a knowledge-based expert system which was presented by Kazemi et al., (2009). These authors formulated the key findings from this study to build a series of cartographic knowledge-based rules as part of their conceptual spatial databases generalisation framework. A survey of cartographic generalisation practices was conducted from November 2005 to May 2006 at several mapping agencies and a number of software vendors. The survey was designed to collect experts' recommendations in relation to new technologies and future generalisation research that could be undertaken by universities and the spatial information industry. The survey results were utilised to build a knowledge-based expert system as explained by Kazemi et al., (2009). Cartographers' feedback was provided in the form of broad qualitative statements, and was analysed to obtain the most pertinent comments. Statistical responses were assessed in quantitative terms and an analysis of the survey results were subsequently incorporated into GES software by Kazemi et al., (2009).

2. Components of Expert Systems

Expert systems have been widely discussed in the literature relevant to knowledge-based research for geospatial applications (e.g. Smith et al., 1987, Hintona, 1996, Bielawski, and Lewand, 1988, Walker and Moore, 1988, Tapiador, 2008, Stockwell, 1999, Almeida et al., 2008 and Mehmood and Tripathi, 2013). An expert system consists of four key components: (a) knowledge acquisition, (b) inference engine, (c) knowledge representation, and (d) user interface (Figure 1).

2.1 Knowledge Acquisition

Radke (1995 and 2007) noted that a key component of GIS is information which is translated into knowledge acquired from evidence and data. These evidences (facts) and data are characteristics abstracted from phenomena under investigation. By collecting, assembling and integrating these data, the GIS analyst derives knowledge and intelligence about the phenomena being studied.

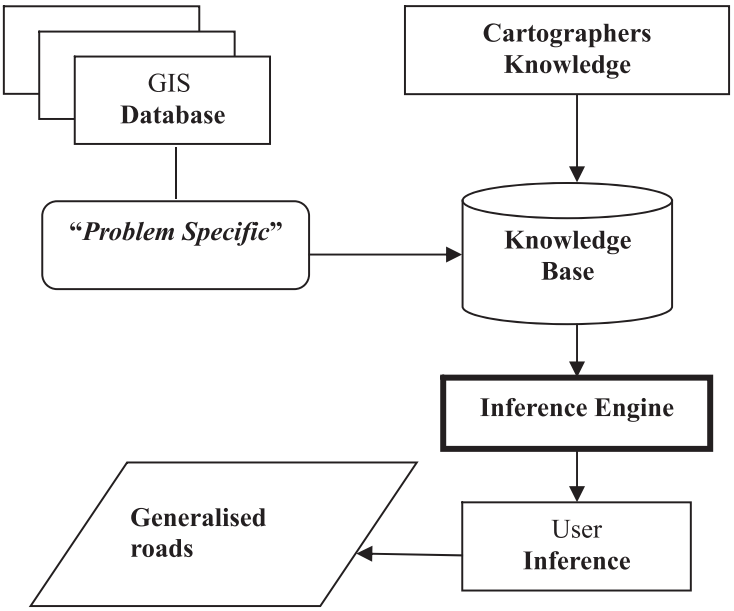

Figure 1 Conceptual expert systems architecture for road generalisation

This could include extraction of rules for problem solving from an expert (i.e. cartographer), and make it adequate for machine intelligence processing in order to perform complex search strategies and apply inference engines (Mason, 1995). The process of transferring knowledge from a cartographer to map generalisation software is not a natural knowledge transfer process, however transferring knowledge from lecturers to students or from parents to children, is a natural process. Several phases are noted in the knowledge acquisition stream, such as knowledge elicitation, knowledge extraction from the knowledge sources, knowledge encoding into symbolic form, knowledge-based organising, and modifying to gain the best performance (Marshall, 1990). In the context of digital cartographic generalisation it involves the examination and interpretation of manual generalisation processes. Translating cartographer's thoughts into a set of explicit and well-defined processes is a major challenge (Lee, 1992). Knowledge extraction approaches are sometimes termed 'methods of inference'. Giarrantano and Riley (1989) suggested that a knowledge extraction approach can be categorised into deduction, induction, intuition, heuristic, generate and test, abduction, default, auto-epistemic, non-monotonic, and analogy. This research will mainly build knowledge through interviews and surveys of cartographers to determine the basis on which the generalisation decisions can be made. This is the framework of knowledge engineering. It attempts to acquire from the cartographers all the elements of heuristic experience. A major challenge in this process is the intensity of the knowledge that is not well organised. Therefore a bottleneck in developing such expert systems is to compile it into a machine-readable format. This 'knowledge acquisition bottleneck' in the field of cartographic knowledge acquisition is noted in Weibel et al., (1995). However, in Mustiere et al., (2000) this problem is overcome by analysing different types of knowledge involved in the cartographic generalisation process. It is necessary to gauge the depth of knowledge, find the right amount of knowledge and accomplish map generalisation by adding some learning abilities to the software and database system. It is noted that cartographic rules are numerous, contradictory and often not formalised. This is because the knowledge 'mixture' is not easy to maintain and limits the comprehensibility of the reasoning done by the system. In Meng (1997), inadequate knowledge formalisation was identified as a problem in successful implementation of generalisation in geographic databases. Other researchers (Visvalingam and Herbert, 1999) have suggested that evaluation of computer-aided generalisation versus manual generalisation is required. Often generalisation algorithms ignore the role of cognitive issues through knowledge discovery techniques such as decision trees, fuzzy logic, data mining and neural networks to extract the hidden knowledge. Once knowledge is discovered, it can be represented in a suitable form to build an expert system. Examples of generalisation rules are: (1) contours never intersect; (2) water bodies are located in the bottom of valleys; (3) roads can cross

each other; and (d) symbology as well as colouring of land-use data is best displayed with colour-hue for visual perspective. These are used as general guidelines when a cartographer makes maps. But cognitive methods introduce more specific knowledge that is ignored in standard generalisation algorithms. Establishing cartographic rules dynamically provides a possible solution to automated generalisation. Different tasks have different rules and require a different knowledge base. Therefore creating a distributed knowledge base with respect to symbolisation, colour schemes, layout, object displacement, etc., is an essential part of this process. This study has evaluated generalisation tools and their functions in order to develop workflows and procedures for generalisation of geographical features such as road networks. The results need to be compared with maps of similar scales. Current generalisation systems, such as Intergraph's DynaGen□, formalised the learning process between cartographer and the generalisation function through an interactive mode. The data analysis (e.g. cluster analysis) and decision-making (e.g. the identification of critical points) are done visually. The drawback of this approach is the subjective nature of the generalisation. Manual generalisation operations are implemented in generalisation systems as functions. An example of a workflow for road data generalisation is: (a) *elimination* – very short branches and unimportant roads to be eliminated, (b) *simplification* – the complexity and the amount of data representing the roads to be reduced, and (c) *smoothing* – the simplified roads to be smoothed to improve visual impression of the output.

2.2 Knowledge Representation

Knowledge representation brings collected knowledge into a suitable form such as decision trees (Forghani, 2007, 2000a and 200b). For example, the Knowledge SEEKER algorithm (De Ville, 1990) produces the knowledge from example data and represents it in the form of decision trees, and offers the capability to convert it into both generic rules and programming statements. Rules are formulated as If-Then or If-And-Then statements into a knowledge base containing separate Conditions using Boolean operators (And, Or, inequalities such as >, <, =) and Actions. The Action segment of the rule is: If the condition is satisfied, then the relevant rule is executed. Rules consist of premise-action pairs, for example: *If* R_1 *&* ... *&* R_n, *Then* C_1 *&* ... *&* C_n. Which reads: *If* premises R_1 and ... and R_n are true, then actions C_1 ... C_n, are taken, where R_i and C_i are *'conditions'* and *'conclusions',* respectively. In cartographic generalisation knowledge representation mainly deals with symbolising geographic objects and is guided by the abstracted object such as how to represent a road so that it is legible (Mustiere et al., 2000). The research is more interested in intuition and heuristic approaches in generalising roads, and will be discussed in future work.

2.3 Inference Methods

A rule-based expert system requires control architecture to decide which rule would need to be applied first, or next, and which rules should be combined. As rules get more complex, computers face increasing difficulty in decision-making. To overcome this problem, forward and backward chaining (Watson, 1997) can be applied. The inference mechanism in the context of its use in spatial context problem solving has been investigated in a number of studies (e.g. Domenikiotis et al., 1995, Nishijima and Watanabe, 1997 and Forghani 1997).

Forward Chaining (Data-Driven): attempts to reach a conclusion through bottom-up reasoning where reasoning starts as the original state of problems from the evidence and facts, to the top-level conclusions that are based on facts. Bielawski and Lewand (1988) noted that the inference method does not compare the information in the goal database with the *Then* part of a rule in the knowledge base, but rather with its *If* statement. If all the conditions are satisfied, then the conclusion is reached (Giarrantano and Riley, 1989).

Backward Chaining (Goal-Driven): starts processing the data and associated rules from the hypotheses (top-down inference) to the lower level facts; in that it supports the choices and conclusions. This begins with a conclusion and proves the conclusion by providing the truth of each premise in a left to right, or top to bottom order. In contrast to forward chaining, the operator begins by assuming a conclusion to be true and then applies the rules to prove it (Giarrantano and Riley, 1989).

2.4 Interface

An expert system consists of a number of major system components and interface performing a range of functions. An expert system must offer a graphical interface so that even an inexpert user should be able to express the ideas, explanations, update and check the knowledge base when running the system in the absence of an operator (Boss, 1991). User interface can influence the applicability of an expert system.

User-friendlier interfaces make a reasonable use of menus, and windows, e.g. setting threshold parameters for line generalisation, etc.

3. Applied Cognitive Knowledge Acquisition Method

AI has played an important role in spatial data management and map production processes across GIS applications. Studies by the author show that knowledge acquisition within existing generalisation systems, e.g. generalisation of line and area features, has not been fully implemented (Kazemi et al., 2004a). Knowledge discovery has led to the amassing of very large repositories of customer, operations, scientific, and other types of data using a number of techniques such as predictive modelling (Provost and Kolluri, 1999). Survey research in general aims to collect information from representative samples of the total population of survey targets. Then the information gathered from the survey sample is used to make a generalisation about the view of the total population with the limitation of random errors. Two major criticisms are regularly made in the literature when discussing surveys of this nature, one is sampling errors due to the sample size, and the other is responses vs. non-response bias (Wunsch, 1986). It is essential that these two issues be addressed when designing a quantitative survey. A key benefit of quantitative survey research is the ability to use small samples to make inferences about the larger population that would be prohibitively expensive to study (Holton and Burnett, 1997). The question is, how large a sample is required to make valid inferences about the target population?. Statistical measurements are often used to determine the correct sample size for a survey, being one of four inter-related features of study design that can affect the discovery of significant differences, relationships or interactions (Peers, 1996). Bartlett et al., (2001) noted that survey design often attempts to minimise both an 'alpha error' (identifying a dissimilarity that does not actually exist in the population), and a 'beta error' (failing to find a difference that appears in the population). However, need to address problems associated with the survey process, such as no responses, no comments/opinions, missing data, and small size when the target population size is not large. The cartographic knowledge acquisition is undertaken through an International Cartographic Generalisation Survey in order to build a rule-based expert system. The next step will be to develop generalisation workflow to make generalisation as efficient as possible. The procedures should also highlight both essential and desirable steps for generating smaller scale maps in line with the production environment. These include topological relations between the object types and classes, how the objects have to be selected, how to generalise, when to smooth, when to delete, when to merge, how to reclassify roads, and so on. It should be noted that such developments and improvements will not be possible by the efforts of a single university, map producer, GIS software vendor or national mapping agency.

4. Sampling Techniques

A robust sampling method reduces the noise in the target population and in turn generates more sensible results. In this regard sample designs for a probabilistic approach include random, systematic, stratified, and cluster samples (Yates, 1981). In random sampling, each element has the same probability of selection and every combination of elements has the same probability of selection to ensure that the sample is representative. In fact all members of the population have an equal chance of being selected. The surveyor can use random number tables or statistical software tools to generate random numbers. In systematic sampling, however, each element has the same probability of selection, but not every combination can be selected (Foot-Retzer, 2003). When applying the stratified sampling method one must ensure that each sample represents some subgroup of the target population, e.g. female employees by age in the workforce, executive personnel by race, and so on. Finally, in the cluster sampling technique, which is generally used in face-to-face surveys, the target population is divided into clusters. The major advantages of this sampling method are that it is inexpensive, no standard sampling framework is required, and the sample size is *not* dependent on population size. Cluster sampling has been applied in a wide range of fields, from engineering (machine learning, artificial intelligence, pattern recognition, mechanical engineering, electrical engineering), computer sciences (web mining, spatial database analysis, textual document collection, image classification and segmentation), life and medical sciences (genetics, biology, microbiology, palaeontology, psychiatry, pathology), to earth sciences (geography. geology, remote sensing), social sciences (sociology, psychology, archaeology, education) and economics (marketing, business) (Hartigan, 1975 and Everitt et al., 2001). To manage, store and analyse a large amount of information an effective way of dealing with the data is to classify or group them into a set of categories or clusters. This can be done either by supervised or unsupervised processes, depending on whether new inputs are to be assigned to one of a

finite number of discrete supervised classes or unsupervised categories (Xu and Wunsch, 2005). The aim of clustering is to separate a large number of unlabelled data into a finite and discrete set of 'natural', hidden data structures, rather than to provide an accurate characterisation of unobserved samples generated from the same probability distribution (Cherkassky and Mulier, 1998). Clustering algorithms partition data into a certain number of clusters such as groups, subsets, and categories, although there is no universally agreed definition (Everitt et al., 2001). The discussion of algorithmic clustering methods is beyond the scope of this paper as classification of the target population was straightforward in this survey. Interested readers are referred to statistical textbooks for a detailed discussion (e.g. Everitt et al., 2001). The cluster sampling is combined with judgment sampling that selects the training data based on judgment, which is convenient for drawing conclusions from the entire target population. In judgment sampling the surveyor uses his/her judgment in selecting the units from the population for study based on the population's parameters. This sampling technique could be the most appropriate if the population to be studied is difficult to locate, or if some members are thought to be better (more knowledgeable, more willing, etc.) than others to interview. This determination is often made on the advice and with the assistance of the client. The target population refers to groups of clients for whom this survey is designed to survey.

4.1 Sample Size Selection

The determination of the appropriate sample size is an important task for conducting research surveys. Taking inappropriate, inadequate, or excessive sample sizes adversely influences the quality and accuracy of research. The survey of this study was constrained by the number of NMAs that participated in the survey (due to its nature). This paper describes the procedures used to conduct a cartographic generalisation survey based on a sample size for categorical variables using Cochran's (1977) formula:

$$n_1 = \frac{n_0}{\left(1 + n_0 / m\right)}$$

Equation 1

Where is required return sample size according to Cochran's formula, and is required returned sample size because sample > 5% of population. A table developed by Bartlett et al., (2001) was used as a guide to select the sample size for this research based on three alpha levels and a set error rate. In addition their procedures for determining the appropriate sample size for multiple regression and factor analysis, common issues in sample size determination, and non-respondent sampling matters were taken into consideration. Cochran's sample size formula and procedures were used here. For this particular survey, the alpha level is assigned a value of 0.05 with an acceptable error of 5% and a standard deviation of 0.5. In statistics, survey sampling is a random selection of a sample from a finite population. It is an important part of planning statistical research and design of experiments. Sophisticated sampling techniques that are both economical and scientifically reliable have been developed (Bartlett et al., 2001). Equation (2) computes:

$$n_0 = \frac{t^2 pq}{d^2}$$

Equation 2

where t is the value for the selected alpha level of 0.025 in each tail = 1.96 (the alpha level of 0.05 demonstrates the level of risk the surveyor is willing to take that the true margin of error may exceed the acceptable margin of error); pq is the estimate of variance = 0.25 (maximum possible proportion (0.5) * 1- maximum possible proportion (0.5) which generates maximum possible sample size) where p is the estimated proportion of an attribute that is present in the population; q is 1-p; and d is the acceptable margin of error for proportion being estimated = 0.05 (error the surveyor is willing to except). Thus, for a population of 174 the recommended sample size is 120. However, in this study, the sample size exceeds 5% of the population (174*0.05 = 8). Cochran's (1977) correction formula was applied to calculate the final sample size. The calculations are shown below (Bartlett et al., 2001):

$$n_1 = \frac{n_0}{\left(1 + n_0 / m\right)}$$

Equation 3

Where the population size is 174; is the required return sample size according to Cochran's formula= 384; and is the required returned sample size because sample > 5% of population. These measures result in a minimum returned sample size of 120. The *response rate* describes the extent to which the final data set includes all sample members (observations). It is the number of respondents divided by the total number of target organisations in the entire population, including those who refused to participate and those who were not available

(Smith, 1983). Conclusions were drawn using a target population of all agencies and software vendors who were a part of the process. These calculations are based on the following factors: Anticipated return rate is 25%, n_2 is sample size adjusted for response rate; minimum sample size (corrected) is 120. Therefore, n_2 is 120/0.25 which equals 480. The power analysis ideally requires 120 responses. The United Nations had 192 members in 2008 and the US State Department recognised 194 independent countries around the world (Worldatlas, 2008). Not all these countries' national mapping agencies were easy to access. During development of the cartographic survey questionnaire, the Australia's national mapping agency was approached when compiling the list of mapping agencies (state, national and international) and commercial spatial data producers. These entities are directly or indirectly involved in cartographic or digital map generalisation (Holland, 2005). It became apparent that there were at least 174 agencies / companies around the world directly or indirectly involved in cartographic or digital map generalisation. In this context the target population is therefore considered to be 174. Some agencies still using traditional cartographic methods for generalisation of spatial data and maps were selected, as well as those that employ a modern generalisation environment. Expressions of interest were received from 75 agencies and software vendors to participate in this survey. A total of 26 surveys were completed by 26 cartographers from several participant mapping agencies (Table 2). A number of authors (e.g. Donald, 1967, Hagbert, 1968 and Miller and Smith, 1983) have discussed the issue of sampling non-respondents. They suggested that the surveyors might take a random sample of 10-20% of non-respondents for use in non-respondent follow-up analyses. If non-respondents are considered as a potentially different population, it does not appear that this recommendation is valid or adequate. Instead, the surveyor could use Cochran's formula to determine an adequate sample of non-respondents for the non-respondent follow-up analyses. Prior to collecting data on the selected observations procedures on how the information will be captured need to be developed. In sample surveys of cartographic knowledge acquisition, the procedures may be the construction of a questionnaire in the form of telephone interviews, face-to-face interviews, or e-mail/postal surveys. How questions are phrased, the order in which they are presented, the time it takes to complete the questionnaire or interview, all influence how people answer. For example, two different versions of the same question can lead to different answers from an individual, potentially invalidating the survey. How data is collected also impacts the interpretation of the survey results. An explicit, precise protocol is needed for each type of data to be collected. The protocol may specify what type of technical questions to use, how to frame the question, the type and length of questions, as well as many other items. The determination of procedures is termed the 'response design'. The goal of the response design is to ensure the collection of consistent information for all sampling units. Those surveys received from the same mapping agencies and software vendors were merged into a single response in order to reflect the overall cartographic knowledge within the organisation. Cartographers from three major streams, including NMAs, SMAs and private industry (software vendors), would be responsible for completing the survey. The Cartographic Generalisation questionnaires were e-mailed to respective target population samples during the period November 2005 to May 2006, and recipients were asked to respond within 30 days. Reminders were e-mailed to all survey recipients that their completed questionnaire was due within the next 15 days. A second attempt was made to remind those who delayed sending the completed survey, in order to minimise the number of non-responding organisations. Out of 75 agencies who were contacted, only 26 cartographers completed the survey from 15 agencies. This response rate was attributed to the fact that the information gathering was conducted by e-mail survey due to a lack of funding and resources for face-to-face surveys / interviews. There were only two face-to-face surveys. The author had access to National Mapping in Australia and also visited Iran's National Cartographic Centre while attending a conference in that country. The same constraints also dictated that the sample sizes were smaller than statistical theory would suggest.

5. Survey Instruments

This section details the process of conducting the survey (Figure 2). The key participants have been identified; the survey design completed and survey testing undertaken. During each of these steps the best survey methodology, and the advantages and disadvantages of each method were considered (Table 1).

Survey Participants: Two types of mapping agencies were targeted; those using traditional cartographic methods for generalisation of spatial data and maps, and those operating in a modern generalisation environment (Table 1).

Table 1: Mapping agencies and software vendors participating in the survey research

Survey Participants	Sector
Kort-og Matrikelstyrelsen (KMS) Denmark	National Mapping
Institute Géographique National France	National Mapping
Geospatial and Earth Monitoring Division, Geoscience Australia	National Mapping
Land Information New Zealand	National Mapping
Iranian Cartographic Centre	National Mapping
BAE Mapping Australia	Private (map producer)
Department of Sustainability and Environment Victoria	State Mapping Agency
Land & Property Information NSW	State Mapping Agency
Department of Environment and Resource Management QLD	State Mapping Agency
National Cartographic Centre of Iran	National Mapping
Ordnance Survey of UK	National Mapping
Intergraph USA	Private (software vendor)
Laser Scan UK	Private (software vendor)
Sweden Lantmäteriet	National Mapping
ESRI USA	Private (software vendor)

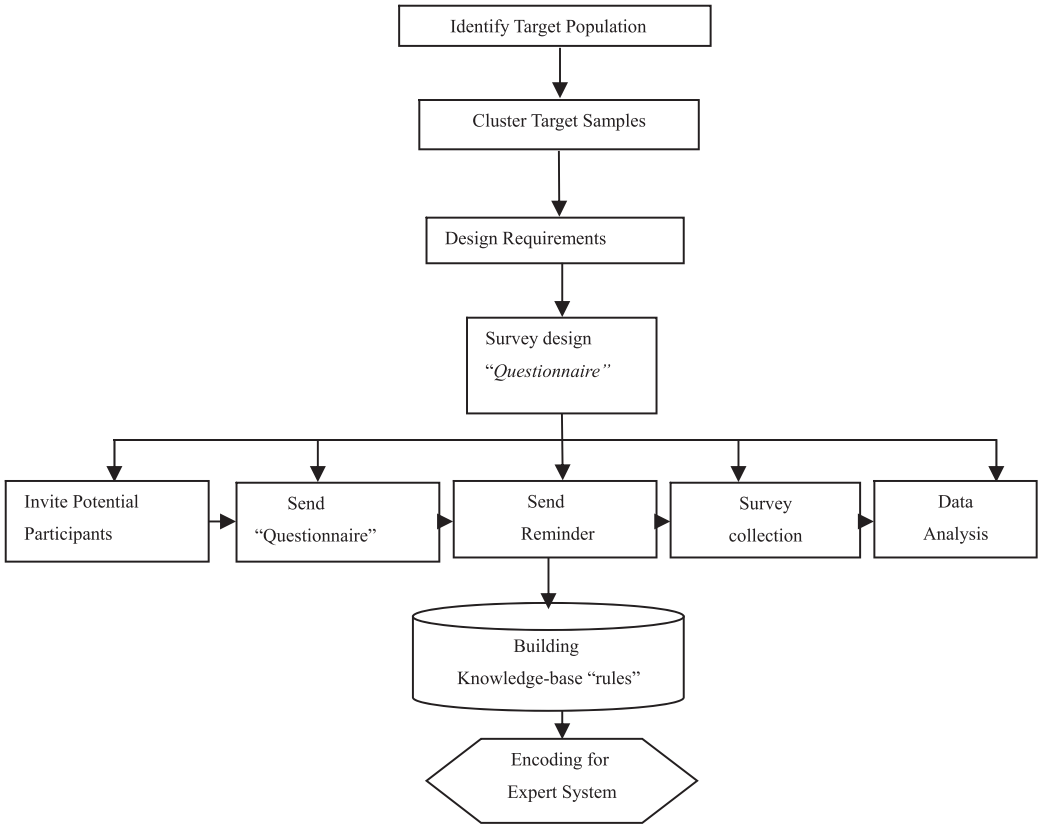

Figure 2: The process of conducting the cartographic generalisation survey

Selection of Survey Method: The face-to-face and e-mail survey methods were chosen here.

Beta Test Survey: The survey was pre-tested before sending it to survey participants, permitting the identification of any problems with the survey, and facilitating revision as necessary. The beta test survey was reviewed by 6 independent subjects. Their comments related to the content, flow, logic and structure. Some modifications were undertaken, resulting in an improved version of the questionnaire.

Initial Contact: Selected target agencies were initially contacted by e-mail to advise them of the survey and its purpose. The survey asked potential respondents if they were willing to participate in the survey, and assuring them that the responses would be treated in confidence. The survey was generally conducted by e-mail, but the opportunity was taken to carry out face-to-face interviews at two NMAs that were easily accessible. The potential respondents were advised of the timeframe for completion of the survey. Respondents were advised that the survey would be sent electronically, and it was asked if they could complete and return the survey within the designated time.

The survey questionnaire was primarily completed by mapping operators (cartographers) but not by their supervisors. This is because the operators possess detailed knowledge of cartographic mapping and generalisation functionalities, etc.

6. Data Analysis

For this cartographic knowledge capture work the summary may be as simple as the percentage of cartographers who completed the survey. How the percentage is calculated depends on the survey design used to collect the data. Among the four sampling techniques mentioned in this paper, it is considered that cluster sampling best fitted the purpose of the survey. When the survey design is a cluster sample, the target population is subdivided into three major streams: NMA, SMA, and private industry (software vendors).

Respondent's Subset: Cartographers' feedback was largely provided in the form of broad qualitative statements and was analysed to extract the most pertinent comments. The six themes that emerged will be discussed below. Of the 75 agencies that expressed interest in participating in the survey, only 26 responses were received from 15 agencies, representing a 20% response rate.

It represent the current status in people activated in cartographic applications with GIS.

The majority of respondents were from the NMAs and SMAs. The results indicated that a cross-section of participants' categories was represented (Figure 3). However, when the NMAs and SMAs categories were collapsed into one subset they were somewhat over-represented compared to the private industry subset. This was because a significant proportion of participants from the latter group declined to participate in the survey. Three incomplete surveys were received from state mapping agencies (QLD, VIC and NSW). The survey analysis results are graphically presented in Figures 3-11.

The responses to key questions are discussed below:
Time spent on generalisation for both modern and traditional cartographic environments: Figure 4 illustrates the time spent on the generalisation process. The time spent on processing the data has a significant effect upon delivery of the product in terms of both quantity and quality. Perhaps it is time that mapping agencies gave more serious consideration to automating the generalisation process where high quality generalisation should be an objective rather than a fast and simple approach. Time should be considered as a source of error and uncertainty by NMAs. Hunter and Goodchild (1995 and 1996) discussed uncertainty in spatial databases, and recommended that the sources of errors should be well explored. The complexity of the process is always an issue and represents a significant barrier to progress. Errors come from a wide range of sources, such as map registration, data conversion from raster to vector, interpretation, analysis, etc. It is useful to measure errors for every given application.

Generalisation Competency in Both Traditional Cartographic Environments: The majority of respondents demonstrated good or better generalisation competency: 36% had an extensive level of competency in generalisation tools, 29% reasonable, 21% moderate, only a small proportion (7%) had limited, minimal or no competencies (Figure 5). This is particularly important as the results of the process totally rely on the level of familiarity with the generalisation tools in order that the best results can be obtained. The respondents were given a clear indication of what is meant by 'generalisation tools', as many people will have different ideas of what are simple or sophisticated generalisation tools.

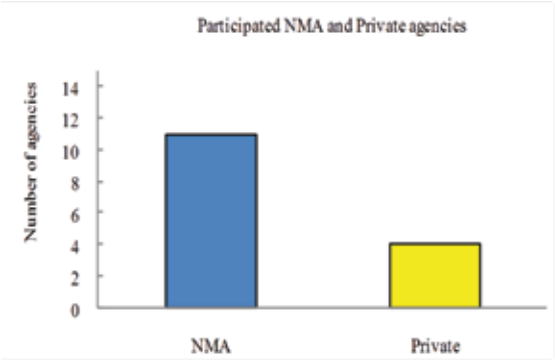

Figure 3: Respondents category

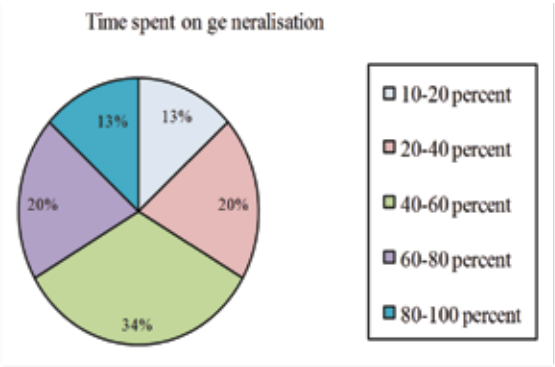

Figure 4: Time category

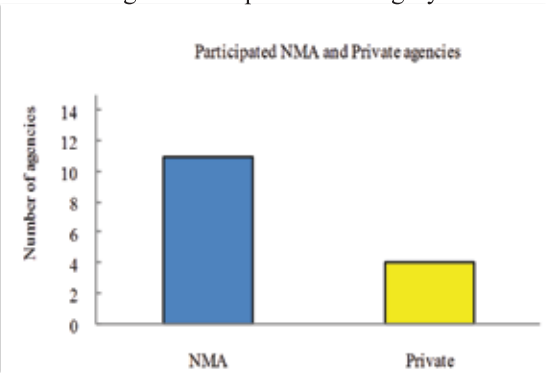

Figure 5: Respondents category

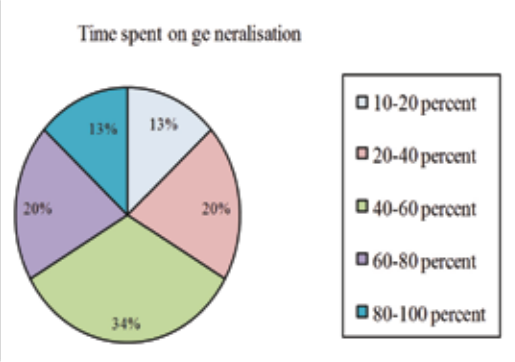

Figure 6: Time category

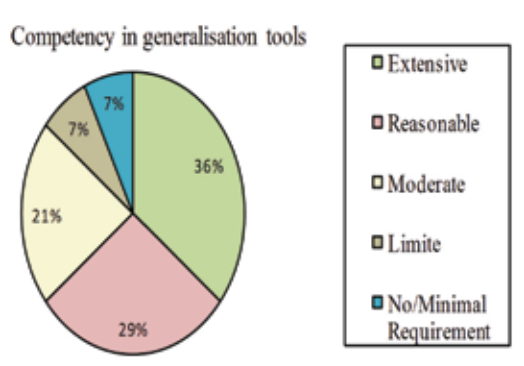

Figure 7: Generalisation tool competencies

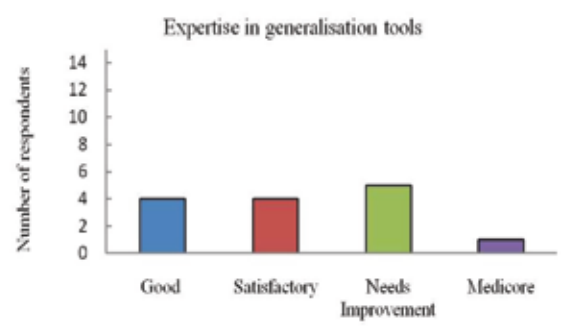

Figure 8: Expertise in generalisation tools

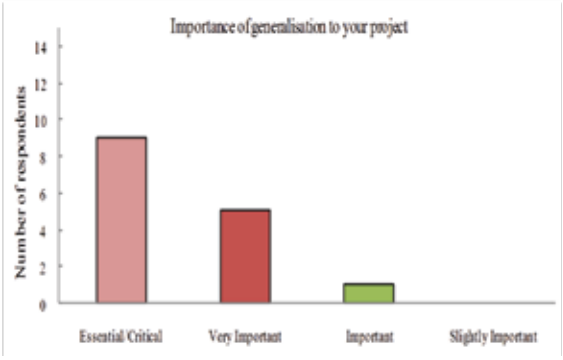

Figure 7: Importance of generalisation to target sample's project

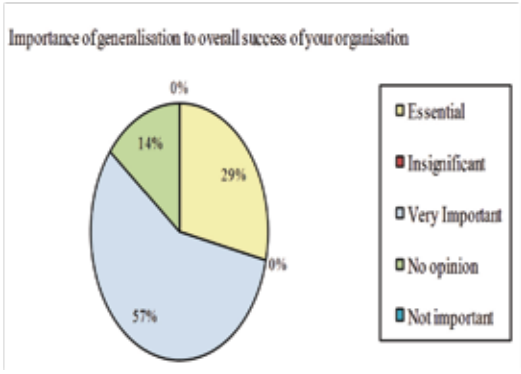

Figure 8: Importance of generalisation to overall success of your organisation

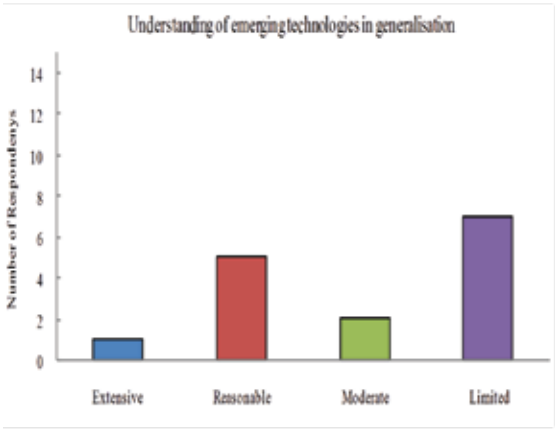

Figure 9: Understanding of emerging technologies in generalization

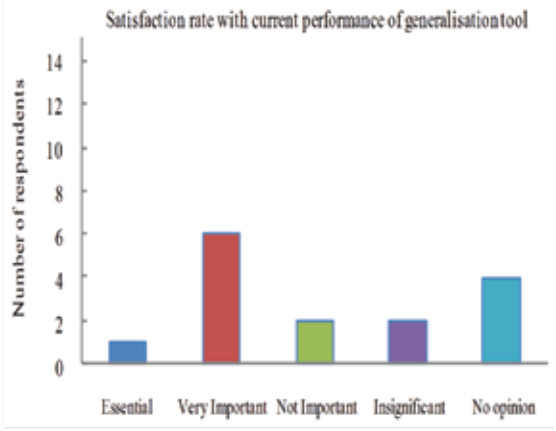

Figure 10: Satisfaction rate with current performance of generalisation

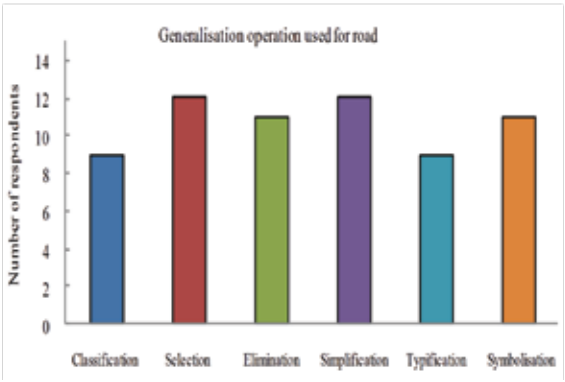

Figure 11: Performance of frequency use of generalisation operator

Their level of understanding would obviously influence their response to these qualitative questions. *Expertise in generalisation: In response to question 3 (Thinking about the transfer of map and database generalisation knowledge within your agency/company, how would you rate the expertise in generalisation there?)* Figure 6 shows the largest number recorded, i.e. that map generalisation technology needs improvement. Due to the rapid development of software and new technologies, it was required that the respondents clearly articulate their expertise in using generalisation tools.

Importance of Generalisation to your Project: The majority of respondents identified the critical importance of generalisation techniques to their project (Figure 7). However, even though agencies rated generalisation very highly, their expertise ranged from ‘beginner’ to ‘advanced’ level. This could be due to the fact that the technology is developing so quickly. There may also be economic impacts on the organisation's funds and human resources that may affect the rate of implementation.

Importance of Generalisation to Overall Success of Your Organisation: Figure 8 speaks for itself. It indicates that 57% of the respondents rated generalisation as being ‘very important’ to the overall success of their organisation, and 29% respondents categorised it as 'essential'. Only 14% answered 'no opinion '. Most of the projects in NMAs involve the updating of various databases from road to vegetation, and other map features. All these processes are time-consuming, costly and, possibly, lead to inconsistent results (e.g. due to the use of different operators).

Understanding of Emerging Technology in Generalisation: This question draws attention to the importance of an awareness of generalisation techniques or technologies, and their progress. It is often worth testing new techniques in a pilot project so as to evaluate its usefulness, and also to introduce and educate the relevant staff. Figure 9 illustrates that the majority of respondents had limited knowledge of emerging techniques/technologies in generalisation.

A key conclusion is that this lack of understanding will affect the results, causing inconsistency and errors, and therefore affecting the accuracy and integrity of the database. It was taken into account that the respondents have a good grasp of 'emerging technologies' such as cutting edge R&D and new products on the market. It was also assumed that respondents have a similar level of understanding of the concept.

Satisfaction Rate with Current Performance of Generalisation: Virtually all the respondents recorded a positive rating for the increasingly important role of generalisation in current and future projects, although a small proportion rated it as 'not important'. Also it should be acknowledged that there were a large number of 'no opinion' responses. This inconsistency implies that the new technology should be capable of satisfying a range of needs (Figure 10).

Generalisation Operation used for Roads: A significant proportion of respondents exploit symbolisation, simplification and selection operations in their generalisation process (Figure 11). The survey focused on the most commonly used generalisation operators. The components of a generalisation process fit together like a 'jigsaw puzzle', and the whole process should work in harmony to deliver an accurate, consistent and well-structured result. There is a wide range of software offering a choice of algorithms and operations, and choosing the right one for the data in hand is not a straight forward or automatic process. However, various software tools were tested in order to select the optimal one to meet the requirements for a given dataset. A significant proportion of respondents exploit symbolisation, simplification and selection operations in their generalisation process. They focus on the most commonly used operators. There is a range of software offering a choice of algorithms and operations, and choosing the right one for the data in hand is not an automatic process.

Handling selection of appropriate/optimal tolerance in simplifying lines using simplification: This question was intended to draw attention to the range of approaches taken to the selection of the right/proper tolerance value(s). Two respondents had 'no opinion', but others made comments and suggestions for improvement which are summarised here:

- By testing various tolerances, and selecting the appropriate tolerance by viewing the results.
- Using detailed specifications and applying size criteria to the algorithms.
- By setting the tolerance at a smaller value there will be less displacement.
- By manually setting the tolerance.
- By editing at 1:250,000 and comparing with reference datasets.
- In general, test and trial is the optimal way since each data set requires a specially tailored tolerance to minimise the manual work.

It was revealed that there is no documented process regarding the selection of the right/proper tolerance. In general, 'test and trial' was the optimal way to determine the tolerance value that minimises the manual work.

Do you use Cartographic Rules, Guidelines, Workflow when Undertaking Generalisation: Overall the answers were divided into two broad categories: rule based decision-making; and practical decision-making without reference to specific rules:

a) Using experience and not a list of rules.
b) Referring to spatial data specifications.
c) Rules for roads include: roads can intersect; short segments less than 1250 m with a dangle cannot exist. Priority is given to road hierarchy such as retaining principle roads. Tracks are often generalised.
d) All generalisation is output scale driven.
e) Maintaining the topological relationships between features.
f) Using combination of experience and rules.
g) Mostly according to cartographic rules.
h) Experience rather using just written rules.
i) Sometimes there are no rules to follow just relying on experience.
j) Mostly cartographic rules plus common sense.
k) By knowing the common errors and trying to fix them according to cartographic rules.

These responses indicate the important role played by cognition, or common sense, in problem solving. If a particular type of error has not previously occurred, a problem solving strategy cannot be developed for it. This means National Mapping Agencies (NMA) and State Mapping Agencies (SMA) must place considerable weight on the cartographer's initiative and experience to ensure the quality of maps (e.g. Lee, 1992 and Kilpelainen, 2000).

In response to the question of *'How would you evaluate accuracy',* the following are the most important recommendations by respondents:

1. Because a 100% automatic generalisation is not able to be achieved, the output that requires the least manual editing afterwards is considered the most valuable. Sometimes it is not the most advanced generalisation which generates the most valuable output. The assessment of the output is important.

2. Evaluation during the process (AGENT technology) that offers some measures of displacement.

3. Critical – most of this decision-making is undertaken: a) with a reasonable appreciation of the region, and b) with a cognitive approach to what are major and minor roads. There is no set standard, each situation is considered separately based on available information and 'gut feel'.

4. This assessment is extremely important. Ideally the user wants to have a scale-less mapping environment for map production with an output of varying map scales. In the case of roads deriving the hierarchy of features between scales will increase or decrease the number of features per scale but will maintain the relationship between databases.

5. Accuracy (and tolerance) is of critical importance in large scale mapping output.

6. Scale plays an important part in map production. When producing a map of 1:500,000 from 1:250,000, some of the map details will be removed due to limited space, but most of them will stay. Producing a map of 1:1,000,000 from 1:250,000, most of the details in 1:250,000 national topographic data will be removed; the vertex and sharp angles in arc features will be removed; the displacements between features will increase. The degree/processes of generalisation will be higher. As a result the quantity of features in the map will decrease. However, if the details of the map present correctly in topological relationships, the quality of the map can be maintained.

7. By referring to the small scale and final checking.

8. Using 1:25,000 as a source map.

9. Double checking, which is time consuming.

10. Mostly rely on software and, in some cases, checking by a different cartographer.

11. Not all the errors are related to process. In some cases they are related to the data source, but will be checked and controlled several times.

Major comments to *'what emerging generalisation research and development area will have the greatest impact on the future direction of automated generalisation technology'.* Examples of feedback are:

- Decision-making systems rather than new specific algorithms.
- The development of Clarity .
- Roads and stream features – much beyond this and you are moving into purely subjective decision-making.
- Annotation.
- Internal development of spatial information resources (Arc database connection).
- The method of handling Annotations.
- With regards to the cost of this sort of research it depends on budgets.
- With new technology it is good if universities and NMA's put some effort in this regard.
- Quality is better than quantity.
- More research in particular area.

Is there any specific topic area that you feel universities and the spatial information industry should be pursuing for future research and development? Examples of feedback are:

- The universities should look at how to solve the complexity when generalising a dataset with many object types, rather than looking at making a good algorithm for a specific object type seen in isolation.
- Label selections.
- Feature link annotation, size of file structures, and advance technologies, e.g. communications.
- Locating of budget in these fields.
- Automation of the process as much as possible.
- Automation rather than manual editing is preferred.
- More automatic process.
- 100% automatic process.

How satisfied are you with the current overall performance of the generalisation tools? The majority of respondents selected 'satisfied' and 'no opinion'. *What is the most effective generalisation framework to derive various types of maps at different scales, e.g. at scales ranging from 1:250,000 to 1:10,000,000?*

There are many generalisation frameworks developed by researchers and cartographers that emphasise the graphical and semantic properties of certain features. These generalisation frameworks enable cartographers to effectively handle spatial information and represent it in multiscale databases. Semi-automated frameworks and procedures are both in use for cartographic and database generalisation applications. However, the quality assessment in map generalisation, formalising knowledge in the generalisation process, and feature conflict detection and resolution are examples of challenging issues for researchers, in order to develop efficient automated generalisation workflows for deriving various types of maps at different scales. Virtually all the respondents recorded a positive rating for the increasingly important role of generalisation. However there were a large number of 'no opinion responses'.

Exploiting the Results: Fundamentally a sample survey is only effective when its results are shared with the intended audience, and if it is used to solve a particular problem. A carefully conducted cartographic knowledge collection survey can be negated by a poorly written report. NMAs hope to convey the message to both participants and to computer scientists to highlight the views of cartographic practitioners so that the heuristic knowledge derived can be drawn upon to build an expert system.

7. Remarks

The cartographers' knowledge acquisition is an integral part of generalisation systems that facilitates the development of a powerful, flexible and robust expert system, which is capable of generating (composing and editing) and manifesting (exhibiting and demonstrating) an innovative method for semi-automated road network generalisation. Without doubt, a comprehensive evaluation of generalisation systems and their performance is essential to embed the cartographic knowledge from experts and bring it into a generalisation framework. While obtaining representative samples in this survey has proven to be a difficult process, the survey revealed that cartographers' knowledge is not being consistently communicated to generalisation software developers. Nor is that knowledge documented consistently across mapping agencies. Out of the 75 agencies that expressed interest in participating in the survey only a total of 26 responses were received from 15 agencies, representing a 20% response rate. The majority of respondents were from the NMAs and SMAs. Both cartographic knowledge and rule-based decision-making are used. However, it was suggested that knowledge-based rules should be formulated in the software. The survey responses indicate the important role played by cognition, or common sense, in problem solving. If a particular type of error has not previously occurred, a problem solving strategy cannot be developed for it. This means that considerable weight must be placed on the cartographer's initiative and experience. It is important to reach widespread agreement among cartographers in relation to specific knowledge about map and spatial data generalisation. The agreed methodological guidelines and procedures could then be incorporated into future software tools to make generalisation operations and algorithms fairer and more reliable. In order to achieve this goal, a set of tools, guidelines and protocols that incorporates a standardised cartographic generalisation methodology needs to be developed and made available to the cartographic and GIS software communities.

Acknowledgements

The lead author would like to acknowledge that this research was not possible without sponsorship and assistance of several individuals and organisations. In particular, the University of New South Wales (UNSW) granted the lead author with the Faculty of Engineering's Research PhD Scholarship and Supplementary Engineering Award (SEA). The authors are grateful to several reviewers including Professor Chris Rizos of UNSW, Professor Gary Hunter of University of Melbourne, Professor Graeme Wright of Curtin University of Technology, Mr Graham Baker of Geoscience Australia (retired) and Dan Lee of ESRI Redlands. They participated in various discussions on the topic of generalisation and provided constructive suggestions toward the lead author's doctoral research. Special thanks are also extended to the University of California Berkeley for hosting Dr Sharon Kazemi as Visiting Scholar and Dr Alan Forghani as Visiting Fulbright Scholar from 2006 to 2007. We thank also the anonymous reviewers for their excellent feedback. The paper is extracted and modified from previous publications.

References

Almeida, C. M., Gleriani, J. M., Castejon E. F. and Soares-Filho, B. S., 2008, Using Neural Networks and Cellular Automata for Modelling Intra-Urban Land-Use Dynamics. *International Journal of Geographical Information Science, Vol. 22, No. 9, 943 - 963.*

Armstrong, M. P., 1991, Knowledge classification and Organization. Map Generalization. Ed. Barbara P. Buttenfield and Robert B. McMaster. Making Rules for Knowledge Representation, *Longman Scientific and Technical*, ISBN 0-582-08062-2.

Bartlett, J. E., Kotrlik, J. W. and Higgins, V. C., 2001, Organisational Research: Determining Sample Size in Survey Research. *Information Technology, Learning, and Performance Journal, Vol. 19, No. 1, 43-50.*

Bielawski, L. and Lewand, R., 1988, Expert Systems Development. Building PC-Based Applications. QED Information Sciences, Inc. Wellesley, Massachusetts, USA.

Boss, R. W., 1991, What Is an Expert System? ERIC Digest. ERIC Clearing House on Information Resources, Syracuse, New York, USA. Accessed online 20 May 2006: http://www.ericdigests-.org/pre-9220/expert.htm

Carrico, M. A, Girard, J. C. and Jones, J. P., 1989, Building Knowledge Systems. McGraw-Hill Book Co., New York, USA.

Cherkassky, V. and Mulier, F., 1998, Learning from Data: Concepts, Theory, and Methods, Wiley, New York, USA.

Cochran, W. G., 1977, Sampling Techniques, 3rd Edition, John Wiley and Sons, New York, USA.

De Ville, B., 1990, Applying Statistical Knowledge to Database Analysis and Knowledge-Base Construction. *Proceedings of the 6th IEEE Conference on Artificial Intelligence Applications, IEEE Computer Society,* 5-9 March 1990, Washington D.C., USA, 30-36.

Domenikiotis, C., Lodwick, G. D. and Wright, G. L., 1995, Intelligent Interpretation of SPOT Data for Extraction of a Forest Road Network. *Cartography,* Vol. 24, No. 2, 47-55.

Donald, M. N., 1967, Implications of Non-Response for the Interpretation of Mail Questionnaire Data. *Public Opinion Quarterly,* Vol. 24, No. 1, 99-114.

Everitt, B., Landau, S. and Leese, M., 2001, Cluster Analysis. Edward Arnold, London, UK.

Foote Retzer, K., 2003, Introduction to Survey Sampling. Survey Research Laboratory University of Illinois at Chicago. Accessed online 20 December 2006: http://www.srl.uic.edu/seminars/Spr03_UIUC-/samplingS03.PDF

Forghani, A., 1997, A Knowledge-Based Approach to Mapping Roads from Aerial Imagery using a GIS Database. PhD Dissertation, Department of Surveying and Spatial Information Science, the University of Tasmania, Hobart, Tasmania, November 1997, Australia, 1.-299.

Forghani, A., 2000a, Decision Trees for Mapping of Roads from Aerial Photography Employing a GIS-Guided Technique. *Proceedings of the 10th Australasian Remote Sensing and Photogrammetry Conference,* 21-25 August 2000, Adelaide, Australia, CD-ROM procs, 12.

Forghani, A., 2000b, Identification of Roads from Aerial Photography Using an Artificial Intelligence Technique. Proceedings of the XIXth ISPRS Congress, Amsterdam, Netherlands, July 2000, pp. 1-3.

Giarrantano, J. and Riley, G., 1989, Expert Systems: Principles and Programming. PWS-Kent, Boston, Massachusetts, USA.

Hagbert, E. C., 1968, Validity of Questionnaire Data: Reported and Observed Attendance in an Adult Education Program. *Public Opinion Quarterly*, Vol. 25, 453-456.

Hartigan, J., 1975, Clustering Algorithms. John Wiley and Sons Inc., New York, USA, ISBN 0-471-35645-X.

Hintona, J. C., 1996, GIS and Remote Sensing Integration for Environmental Applications. *International Journal of Geographical Information Systems,* Vol. 10, Issue 7, 1996, 877-890.

Holland, P., 2005, Verbal Communication. Geoscience Australia, Canberra, Australia.

Holton, E. H. and Burnett, M. B., 1997, Qualitative Research Methods. In R. A. Swanson, and E. F. Holton (Eds.), Human Resource Development Research Handbook: Linking Research and Practice, Berrett-Koehler Publishers, San Francisco, California, USA. ISBN 978-1-57675-314-9.

Hunt, E. B., 1962, Concept Learning: An Information Processing Problem. John Wiley and Sons, New York, USA, ISBN 0-471-14890.

Hunter, G. J. and Goodchild, M. F., 1995, Dealing with Error in Spatial Databases: A Simple Case Study. *Photogrammetric Engineering and Remote Sensing,* Vol. 61, No. 5, 529-537.

Hunter, G. J. and Goodchild, M. F., 1996, Communicating Uncertainty in Spatial Databases. *Transactions in GIS.* Vol. 1, No. 1, 13–24.

Iwaniak, A. and Paluszynski, W., 2001, Generalisation of Topographic Maps of Urban Areas. *Proceedings of the 20th International Cartographic Conferences, 6-10 August 2001, Beijing, China,* CD-ROM procs, 12pp. Accessed online 19 July 2003:http://icaci.org/documents-/ICC_proceedings/ICC2001-/icc2001/ author.htm

Kazemi, S., Lim, S. and Rizos, C., 2004a, A review of Map and Spatial Database Generalisation for Developing a Generalisation Framework. *Proceedings of the XIX Congress of 2004 the International Society for Photogrammetry and Remote Sensing,* 12-23 July 2004, Istanbul, Turkey, 1221-1226.

Kazemi, S., Lim, S. and Paik, H., 2009, Generalisation Expert System (GES): A Knowledge-Based Approach for Generalisation of Line and Polyline Spatial Datasets. *Proceedings of the Surveying and Spatial Sciences Institute Biennial International Conference,* 28 September-2 October 2009, 717-729.

Khoshnevis, B. and Parisay, S., 1993, Machine Learning and Simulation: Application in Queuing Systems. *Simulation,* Vol. 61, No. 5, November, 294-302.

Kilpelainen, T., 2000, Knowledge Acquisition for Generalisation Rules. *Cartography and Geographical Information System*, Vol. 27, No. 1, 41-50.

Krejcie, R. V. and Morgan, D. W., 1970, Determining Sample Size for Research Activities. *Educational and Psychological Measurement,* Vol. 30, 607-610.

Lee, D., 1992, Cartographic Generalisation. Intergraph Corporation Press, USA.

Marshall, G., 1990, Advanced Students' Guide to Expert Systems. Heinemann Newnes, Oxford, UK.

Mason, S., 1995, Expert System-Based Design of Close-Range Photogrammetric Networks. *ISPRS Journal of Photogrammetry and Remote Sensing,* Vol. 50, No. 5, 13-25.

McKeown, D. M., Harvey, W. A. and McDermott, J., 1985, Rule-based interpretation of aerial imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 8, 532-542.

Mehmood, H. and Tripathi, N. K., 2013, Cascading Artificial Neural Networks Optimized by Genetic Algorithms and Integrated with Global Navigation Satellite System to Offer Accurate Ubiquitous Positioning in Urban Environment. *Journal of Computers, Environment and Urban Systems,* Volume 37, Pages, 35-44.

Meng, L., 1997, An In-depth Investigation on the Long Lasting Problem of Cartographic Generalisation. Technical Report, May 1997, *Swedish Armed Force*, 1-75.

Michie, D., Spiegelhalter, D. J. and Taylor, C. C., 1994, Machine Learning, Neural and Statistical Classification. Prentice Hall, Englewood Cliffs, N.J., USA, ISBN-10: 013106360X.

Miller, L. E. and Smith, K. L., 1983, Handling Non-Response Issues. *Journal of Extension,* Vol. 21, 45-50.

Mitchell, T. M., 1997, Machine Learning, McGraw-Hill, New York, USA, ISBN-10: 0070428077.

Mustiere, S., Zucker, J. and Saitta, L., 2000, An abstraction-Based Machine Learning Approach to Cartographic Generalisation. *Proceedings of the 9th International Symposium on Spatial Data Handling,* 10-12 August 2000, Beijing, China, CD-ROM procs, 10.

Nikolopoulos, C., 1997, Expert Systems: Introduction to First and Second Generation and Hybrid Knowledge Base Systems, Marcel Dekker, Inc., New York, ISBN 10: 0824799275.

Nishijima, M. and Watanabe, T., 1997, A Cooperative Inference Mechanism For Extracting Road Information Automatically. *Proceeding of the 3rd Asian Conference on Computer Vision*, Volume II, 8-10 January 1998, Hong Kong, 217-224.

Peers, I., 1996, Statistical Analysis for Education and Psychology Researchers. Falmer Press, Bristol, UK.

Provost, F. and Kolluri, V., 1999, A Survey of Methods for Scaling up Inductive Algorithms. *Data Mining and Knowledge Discovery,* Vol. 2, 131–169.

Quinlan, J. R., 1983, Learning Efficient Classification Procedures and Their Application to Chess and Games. In R.S. Michalski, J.G. Carbonell and T.M. Mitchell (ed), Machine Learning: An Artificial Intelligence Approach, Tioga Publishing Co. PaIo Alto, California, USA.

Quinlan, J. R., 1986, Induction of decision trees. *Machine Learning.* Vol. 1, 81-106.

Quinlan, J. R., 1990, Learning Logical Definition from Relations. *Machine Learning,* Vol. 5, 239-266.

Radke, J., 2007, Lecture Notes - Geographic Information Systems. Department of Landscape Architecture and Environmental Planning/ Department of Geography, University of California Berkeley, California, USA, 1-50.

Radke, J., 1995, Modeling Urban/Wildland Interface Fire Hazards within a Geographic Information System. *Geographic Information Sciences,* Vol. 1, No. 1, 7-20.

Rdenas, S. N., Wang, L. and Zhan, F. B., 2009, Representing Geographical Objects with Scale-Induced Indeterminate Boundaries: A Neural Network-Based Data Model. *International Journal of Geographical Information Science,* Vol. 23, No. 3, March 2009, 295–318.

Smith, T. R., Peuquet, D., Menon, S. and Agarwal, P., 1987, KBGIS-II: A Knowledge-Based Geographical Information System. *International Journal of Geographical Information Systems.* Vol. 1, 149-7.

Smith, T. W., 1983, The Hidden 25 Percent: An Analysis of Non-Response on the 1980 General Social Survey. *Public Opinion Quarterly,* Vol. 47, 386-404.

Stockwell, D., 1999, The GARP Modelling System: Problems and Solutions to Automated Spatial Prediction. *International Journal of Geographical Information Science,* Volume 13, Issue 2, 1999, 143-158.

Tapiador, F. J., 2008, Management Tools: Geographical Information Systems (GIS) and Expert Systems, Rural Analysis and Management, Springer Berlin Heidelberg.

Visvalingam, M. and Herbert, S., 1999, A Computer Science Perspective on the Bend Simplification Algorithm. *Cartography and Geographic Information Science,* Vol. 26, 253-270.

Walker, P. A. and Moore, D. M., 1988, SIMPLE: An Inductive Modelling and Mapping System for Spatially Oriented Data. *International Journal of Geographical Information Systems,* Vol. 2, 347-63.

Wang, F. and Newkirk, R., 1988, A Knowledge-Based System for Highway Network Extraction. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 26, No. 5, 525-531.

Watson, L., 1997, Applying Case-based Reasoning, Morgan Kaufmann Publishers, San Francisco, California.

Weibel, R., Keller, S. and Reicheenbacher, T., 1995, Overcoming the Knowledge Acquisition in Map Generalisation: The Role of Interactive Systems and Computational Intelligence. In A. Frank and W. Kuhn (Eds.), Spatial Information Theory, Springer, Berlin, Germany, 139-156.

Worldatlas, 2008, How Many Countries? Accessed online October 2008: http://www.worldatlas-.com/-nations.htm

Wunsch, D., 1986, Survey Research: Determining Sample Size and Representative Response. *Business Education Forum*, Vol. 40, No. 5, 31-34.

Xu, R. and Wunsch, D., 2005, Survey of Clustering Algorithms. *IEEE Transaction on Neural Networks*, Vol. 16, No. 3, 645-648.

Yates, F., 1981, Sampling Methods for Censuses and Surveys, 4th Edition, Griffin, London, 1-254.