A Rough Set Based Methodology for Geographic Knowledge Discovery

Colin H Aldridge
Spatial Information Research Centre
University of Otago. Dunedin, New Zealand
Phone: +64 3 479-7391 Fax: +64 3 479-8311
caldridge@infoscience.otago.ac.nz

Abstract. This paper presents the RS-GKDD (rough set geographic knowledge discovery in databases) methodology. The purpose of the methodology is to take digital geographic data and induce a model that, expressed as production rules, constitutes non-trivial, previously unknown, and potentially useful information. The methodology is outlined in terms of its philosophy, objectives, scope, constraints, assumptions, theory, procedures, tools, and phases.

The principal theoretical treatment introduces rough sets, rough set-based knowledge induction, and spatial context. Where appropriate, the procedures associated with the RS-GKDD methodology are described using algorithms. The various aspects of the methodology are drawn together as four phases: (i) project analysis and design; (ii) data preparation; (iii) knowledge induction, reduction and validation; and (iv) knowledge interpretation and application.

The paper outlines the results of applying the methodology to ecological data relating the spatial distribution of greater glider possums to the quality of their habitat. In the case study geographic knowledge is sought because the ability of the animals to glide between trees suggests that spatial relationships between habitat variables may be important. The dataset has already been analysed by other researchers using several knowledge induction methods to generate both spatial and non-spatial models. The results of these earlier studies are used as benchmarks against which to compare the results of the present work. The analysis of results shows that, of the models for predicting the population density of greater glider possums, the four best, when selected for classification accuracy and comprehensibility, include a rough set based model that was induced by considering spatial context.

1. INTRODUCTION

In the following sections, the RS-GKDD methodology is briefly described first. The objectives of the Greater Gliders study are then introduced. The ecological context of this study provided, followed by a discussion of the data used, its properties and the preparation necessary to ready it for analysis. The results obtained by applying the RS-GKDD knowledge induction, reduction and validation methods to the data are outlined. Finally the results are analysed, conclusions drawn and suggestions made for further research.

2. THE RS-GKDD METHODOLOGY

The *rough set (based) geographic knowledge discovery in databases* (RS-GKDD) methodology provides a means for inducing rule-based knowledge from the digital database equivalents of chloropleth maps. It combines many of the elements of the comprehensive information system development "methodologies" of Kennedy (1993) and Maddison (1983) with the tasks of the knowledge discovery "process" described by Fayyad et al. (1996). The "tasks" of Fayyad et al. therefore become "phases" of an encompassing methodology, which also has ascribed to it a philosophy and a theory based around, firstly, the concept of rough sets proposed by Pawlak (principally 1982, 1991) and, secondly, a theory of geographic knowledge centering on chloropleth maps and their digital representation (Aldridge 1998). The RS-GKDD methodology also has objectives, scope, constraints, assumptions, theory, procedures, algorithms, and tools. Knowledge discovery is accomplished in four principal phases: project analysis and design, data preparation, rough-set-based knowledge induction, and

knowledge interpretation/application. The scope for iterative looping and back-tracking in the execution of these phases is acknowledged. The methodology is described in detail by Aldridge (1998).

2.1 RS-GKDD Philosophy

Pawlak's emphasis on classification (Pawlak 1991) as the essence of knowledge about objects is fundamental to this knowledge discovery methodology. Consequently, the basic philosophy of RS-GKDD knowledge discovery is to make maximum use of such *a priori* factual knowledge, while minimising the introduction of additional assumptions. This position is stated succinctly by Düntsch and Gediga (1998: 110)

"Rough set analysis uses only internal knowledge, and does not rely on prior model assumptions as fuzzy set methods or probabilistic models do. In other words, instead of using external numbers or other additional parameters, rough set analysis utilises solely the granularity structure of the given data, expressed as classes of suitable equivalence relations."

In short, the philosophy of RS-GKDD is to endeavour to *let the data speak for itself, so far as is practical.*

2.2 Objectives

The principal objective of *geographic knowledge discovery in databases* (GKDD) is to augment the capacity of humans to obtain useful knowledge from multi-theme, 2-dimensional, geographic databases (Aldridge 1998). The input data are the digital equivalents of chloropleth maps that, together, comprise a sufficient number of themes as to render manual interpretation difficult or impossible. New knowledge is induced as a knowledge model, one that should fulfil its user's requirements for nontrivial, previously unknown, and potentially useful information (c.f. Frawley et al. 1992).

2.3 Constraints

The principal constraint on *rough set-based* GKDD is the size of the candidate rule space encountered during knowledge induction. The cost of examining this space is exponentially related to the number of examples and the number of themes (i.e. it is NP-hard) (Aldridge 1998). A major challenge is to develop effective search strategies and algorithms.

Inevitably, the choice of language elements and grammar inherent in RS-GKDD constrains the domain of knowledge that can be "discovered" using the methodology. In particular, while the underlying rough set theory deals well with nominal scale data, it cannot utilise the additional information in ordinal, interval and ratio scale measures.

2.4 Assumptions

The principal assumptions of the RS-GKDD methodology are:

- A *data validity assumption*, namely that the input data–the digital equivalents of chloropleth maps–are, in fact, a sufficiently valid representation of reality.
- A *closed world assumption* which asserts that the knowledge contained in the knowledge representation system used for induction (the training data) is complete and is closed. It asserts internal validity.
- An *open world assumption* applicable when the induced model is used on new examples. It asserts external validity. This is well-recognised as the necessary and potentially dangerous assumption needed to apply knowledge induced from one set of examples to new, previously unseen examples.
- A *data representation assumption*, which asserts that the input geographic knowledge (chloropleth maps) used for knowledge discovery are capable of being represented by a tabular knowledge representation system (Pawlak 1991: 55) and is, therefore, amenable to treatment using the rough set theory outlined below. A knowledge representation system (below) closely resembles a conventional relational database table.

2.5 Theory

2.5.1 Knowledge Induction

Knowledge that leads to contradictory conclusions is an all-too-common real-world experience. The need for a theory that copes with incomplete knowledge and inconsistent rules is an important motivation behind the development of rough set theory (Pawlak 1982, 1991). In the RS-GKDD methodology, the theory of rough sets is adopted to the extent necessary to support the analysis of inconsistent rules, and the induction of useful rule-based knowledge.

Rough set-based knowledge induction commences with a knowledge representation system (KRS) table (Pawlak 1991), as is illustrated in Figure 1. A *first stage of generalisation* is achieved by deleting object identifiers from the KRS (c.f. Cohen and Feigenbaum 1982). This operation transforms the KRS into a decision table (Figure 1), as defined by Pawlak (op. cit.). The problem of inductive learning subsequently becomes one of reducing the decision table to an abbreviated table able to be interpreted as a minimal production rule system.

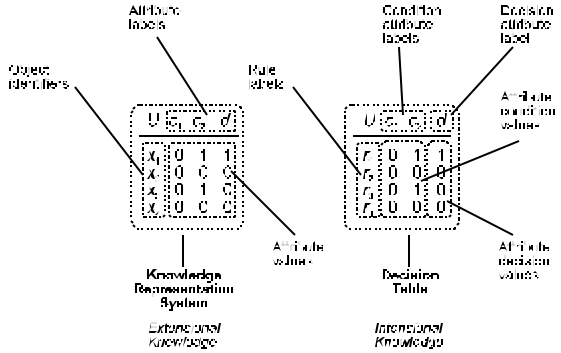

Figure 1 Knowledge representation system generalised into a corresponding decision table

Decision table reduction using rough set theory attempts to answer, for a particular input data set, the following questions:

1. Are all the input data condition attributes necessary to define the set of example decisions? And if not, which attributes should be used?
2. What is the minimal set of condition values necessary to make any particular decision?

Such questions reflect a need to reduce knowledge to what is a necessary/sufficient description for effective and efficient decision making. This, in turn, requires an examination of how one knowledge depends on other knowledges. This is the grist of rough set theory. The aspects of this theory relevant to knowledge induction are reviewed by Aldridge (1998), where an extended worked example is used to illustrate the application of the theory.

2.5.2 Spatial Context

The *spatial context* of a particular raster element (n_x, n_y) is defined as the set of spatial elements $Context_{x,y}$ standing relative to (n_x, n_y), as in Figure 2. When the focus of attention moves to another focus element, this defines a new set of spatial context elements in the same *relative* standing with respect to the new focus element.

There are many ways in which spatial context knowledge might be used to supplement knowledge about a specific element. The spatial context adopted in RS-GKDD is the simple one that includes all attributes of all proximal elements and which extends as far from the focus element as is computationally practical. Underlying this choice is the assumption of *localisation*, which is that nearer elements contribute more useful predictive knowledge than do distant elements.

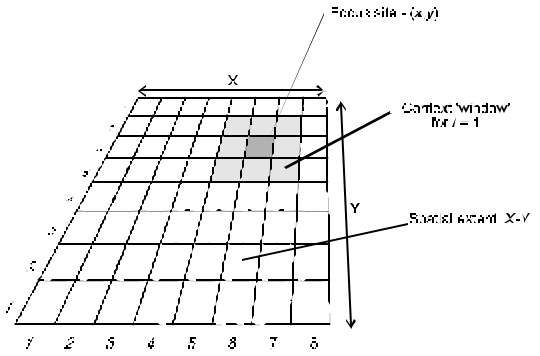

Figure 2 Relationship of context 'window' to a raster spatial extent

2.6 Procedures

RS-GKDD takes place in a number of steps. Procedures for rasterisation, discretisation, random sampling and *n*-fold cross validation are well established and have been adapted to the particular needs of the RS-GKDD methodology. Other procedures are specific to RS-GKDD. They are: KRS compilation, instance sampling, rough set based rule induction, attribute subset search, rule validation, and rule evaluation and selection.

2.6.1 Rasterisation

Knowledge induction from several chloropleth map themes is facilitated by using a common raster. This ensures that the spatial objects used in analysis are constant in shape and size. The use of raster elements also enables the convenience of a numerical representation of spatial relationships through the associated coordinate system. A raster representation also has the considerable advantage that analysis can focus on differences between attributes. In addition, in context-based analysis, the spatial relationships between raster cells are used prior to knowledge induction to assemble context window attributes, and after knowledge induction, to determine the spatial relationships

contained in discovered rules. As a result of these considerations, thematic vector data is rasterised prior to further analysis using standard GIS software.

In choosing an element size when digitising existing maps, analogy with Shannon's Sampling Theorem (Shannon 1949) suggests that to fully describe all map regions, the chosen raster cell size (c.f. wavelength) should be half the smallest region dimension in the database. However, practical considerations, particularly related to the size of search space in inductive analysis, are likely to preclude this ideal.

2.6.2 Attribute Discretisation

Chloropleth maps typically have a relatively small number of legend categories. Usually, these categories are nominal-scale measures, such as soil type, land use class, etc.. Occasionally, they are ordinal measures. However, in analysis, some continuous attributes such as altitude and slope are significant variables. In such cases, the spatial data is not only rasterised, but the attribute data is also mapped into discrete classes, that is, it is 'discretised'. The method for discretisation used in RS-GKDD is that of Cassie (1954).

2.6.3 Knowledge (KRS) Compilation

For each theme in a geographic dataset, the output of rasterisation and, where used, discretisation, can be represented by a three-column spatial knowledge representation system (KRS) table. The first two columns contain two-dimensional coordinates identifying raster cells. The third column contains the raster cell attribute category values for the theme. Also required are:

- Attribute category codes as a look up table for use in assigning meaningful category names to the numeric codes used in computations.
- Spatial parameters describing raster cell dimensions, raster boundaries and coordinates. These are used to relate the RS-GKDD model to geographic reality.

Compilation of a single KRS that collates all chloropleth map themes in the dataset is simply a matter of appending an additional attribute column for each theme.

2.6.4 Instance Sampling

The number of instances (i.e. raster cells) in a rasterised spatial dataset may be too many to enable processing in an acceptable time. Consequently, some means of reducing the dataset size is required—one that will at the same time maintain its essential characteristics for knowledge induction. The strategy adopted in RS-GKDD is a simple one of extracting a random sample of cases, while, if necessary, focusing on those cases specified by the user to be the most relevant. For instance, if the positive examples of binary decision data are relatively few in the dataset but are important for decision making, sampling can be weighted towards positive instances.

2.6.5 Rough Set Based Rule Induction

The rough set knowledge induction process is implemented in an algorithm that applies the theory described above (Aldridge 1988). The output of the algorithm is a reduced, tabular decision table from which has been eliminated any condition attribute (column), rule (row), or value that does not contribute towards discerning between input data objects. The eliminated condition attributes and values can be regarded as having been replaced by a "wildcard" value that can match any condition attribute value.

For example,

KRS, $S =$

U	a	b	c	d	e
x_1, y_1	1	2	0	1	1
x_2, y_1	1	2	0	1	1
x_3, y_1	0	0	1	1	0
x_4, y_1	2	1	0	2	1
x_5, y_1	2	1	0	2	1
x_1, y_2	0	0	0	2	2
x_2, y_2	0	0	1	1	0
x_3, y_2	0	1	0	2	1
x_4, y_2	2	1	0	2	2
x_5, y_2	0	0	1	1	0

would be

reduced to decision table, $D = \begin{bmatrix} c_1 \rightarrow e_0 \\ b_2 \rightarrow e_1 \\ a_2 \rightarrow e_1 \\ a_0 \wedge b_1 \rightarrow e_1 \\ b_0 \wedge c_0 \rightarrow e_2 \\ a_2 \rightarrow e_2 \end{bmatrix}$.

For example, the fourth rule in *D* could be written out in full as, "*If* condition attribute *a* has a value 0 *and* condition attribute *b* has a value 1, *then* the decision attribute *e* has a value 1."

2.6.6 Rule Evaluation and Selection

Because of the unconstrained approach taken, the rough set based rule induction phase of RS-GKDD typically generates many more rules than would meet requirements for non-trivial, informative, and potentially useful knowledge. Some rules will be weak ones because they are supported by only a few examples in the knowledge base. Others will be poor generalisations because their decisions are inconsistent with other rules having the same pre-conditions. A real possibility is that some rules may not even be better at making decisions than a random choice. A sound and effective method for selecting rules from the reduced decision table to give a single set of deterministic rules is required.

A significant contribution of the RS-GKDD methodology is in providing an effective and statistically sound basis for evaluating and selecting the rules produced by rough set knowledge induction.

The principal measures of rule quality are support and generalisation accuracy. S*upport* is the proportion of all available examples that satisfy a given rule—that is, those database instances both satisfying the rule's precedent (conditions), and its consequent (decision). *Generalisation accuracy*[1] is the proportion of examples that satisfy a given rule's precedent (conditions) and its consequent (decision), compared with the number of examples that can satisfy its precedent.

In statistical rule evaluation and selection, rules are selected on the basis that their support and generalisation accuracy measures are unlikely to have arisen by chance. The resulting pruned rule set comprises the *statistically significant rules.* From these, the person controlling the knowledge discovery process may select *interesting rule*s by arbitrarily defining selection thresholds for support and generalisation accuracy. Appropriate combinations of support and generalisation accuracy thresholds may be interpreted as defining *strong rules* (Koperski and Han 1995).

[1] Usually called "confidence" in KDD literature, but this invites a potential confusion comparison with the widely used terms "confidence interval" and "confidence limit" from the field of statistics.

Finally, the statistically significant, interesting rules are filtered using a proximity measure based on the idea that for a given database example, its "nearest" rule in a rule set is the one with the smallest number of unmatched attributes (Gawrys and Sienkiewicz 1993). Using this criterion, all rules can be compared with each example and the *"nearest" rules* to the example set chosen. An important consequence of nearest rule pruning is that, because the algorithm forces the selection of one rule for each training data object, the output rule set is comprised entirely of consistent rules.

2.6.7 Ruleset Validation

Once a pruned ruleset has been obtained, the extent to which it is valid is evaluated. A widely used measure for evaluating classification models is *classification accuracy*, which is the proportion of objects (raster cells in this case) that are correctly classified by the model. The method for determining classification accuracy used in RS-GKDD is a refinement of cross-validation (Stone 1974) known as *k*-fold cross-validation (Efron 1982, Gunter 1997).

The essence of cross-validation is as follows: First, the output ruleset is induced as described above, using *all* the available data. Second, the dataset is then randomly divided into two parts. The first part, a fraction 1-1/*k* of the data is used as "training data" on which to induce another ruleset. The second data fraction, being 1/*k* of the data, is used as "test data" to calculate the classification accuracy of the model developed on the first dataset. The second phase is repeated *k* times to give *k* estimates of classification accuracy, from which a mean classification accuracy and standard deviation can be calculated.

2.6.8 Attribute Subset Search

In RS-GKDD an 'optimal' subset of condition attributes is found using a simple hill-climbing search heuristic. *k*-fold cross validated accuracies and their variances are used to guide the search.

2.7 Tools

Tools used to implement the RS-GKDD methodology are: GIS software, which is used principally for converting vector map representations to raster data, and for displaying input data and output data; RS-GKDD application software written by Aldridge (1998), which implements the procedures described above, and

incorporates core rough set functions from a library written by Gawrys and Sienkiewicz (1993); simple statistical analysis software for discretisation; and a text editor for manipulating KRS data.

2.8 Phases

RS-GKDD is carried out in four phases: project analysis and design; data assessment and preparation; knowledge induction, reduction and validation; knowledge interpretation and application. These phases are likely to be iterative.

Central to the knowledge induction, reduction and validation phase is a knowledge induction method designed to cope with large data sets. In this context, the term “large” means having a size where the computational cost of applying the rough set rule induction algorithm to the complete dataset would be impractical. The method includes sampling and attribute selection strategies for dealing with datasets that are large in terms of both numbers of objects, and numbers of attributes. The method is described in a flow chart (Algorithm 1).

Algorithm 1: Specification of rough set-based geographic knowledge induction and cross-validation

INPUT

$K = (U, C, D)$ — Initial KRS

PREPARE DATA

$K_{CW} = (U, C_{CW}, D) = CW(U, C, D, d_{cw})$ — Assemble KRS that includes "context window" attributes (non-spatial if context distance $d_{cw} = 0$))

$K_{SA} = (U_{SA}, C_{CW}, D) = SA(K_{CW}, s)$ — Define sample KRS (no sampling if sample proportion $s = 1$)

$C_{OPT} = HC(K_{SA}, k, c)$ — Hill-climbing search for optimal condition attributes

INDUCE RULES FROM ALL DATA

$T_{RS} = (U_{RS}, C_{RS}, D) = RS(U_{SA}, C_{OPT}, D)$ — Define rough set reduced decision table

$T_{SS} = (U_{SS}, C_{RS}, D) = SS(T_{RS}, K_{SA}, c)$ — Prune to statistically significant rules

$S = SUP(T_{SS}, K_{SA})$ — Compute absolute supports for stat. sig. rules

$T_{SS,NR} = (U_{SS,NR}, C_{RS}, D) = NR(T_{SS}, S)$ — Prune to "nearest" stat. sig. rules

$T_{SS,NR,IN} = (U_{SS,NR,IN}, C_{RS}, D) = IN(T_{SS,NR}, S, A, I, J, th_s, th_a, th_I, th_J)$ — Prune to "interesting", "nearest", stat. sig. rules

PARTITION DATA INTO k SUBSETS

$\{(U_{SA_1}, C_{OPT}, D), \mathrm{K}, (U_{SA\,j}, C_{OPT}, D)\}$ — Randomly partition sample data U_{SA} into k equal parts. (Adjusted for non-zero remainders.)

LOOP OVER DATA SUBSETS

$j = 1, \mathrm{K}, k$

INDUCE RULES FROM SUBSET j DATA

$T_{SS,NR,IN\,j} = IN(NR(SS(RS(U_{SA\,j}, C_{OPT}, D), U_{SA\,j}, C_{OPT}, D, c), S_j), thr_J, thr_I, Thr_S, Thr_G)$

LOOP OVER DATA SUBSET INSTANCES

$i = 1, \mathrm{K}, o_j$ — o_j is the number of objects in the data subset j.

$d(u_i') = CL(T_{SS,NR,IN\,j}, S_j, u_i).$ — Classify objects in the data subset j

If $d(u_i') = d(u_i)$ then $m = m + 1$ — Count correct decision predictions

end of i loop

$acc_j = m / card(o_j)$ — Classification accuracy of subset j rules

$sum\_acc = sum\_acc + acc_j$ — Sum of classification accuracies

$ssq\_acc = ssq\_acc + acc_j^2$ — Sum of squares of classification accuracies

end of j loop

$mean\_acc = MEAN(sum\_acc, k)$ — Average of k classification accuracies

$sdev\_acc = SDEV(mean\_acc, ssq\_acc)$ — Standard deviation of k classification accuracies

OUTPUT

$T_{SS,NR,IN}$ — Predictive model as a decision table

$mean\_acc$ — Cross-validated mean classification accuracy of model

$sdev\_acc$ — Standard deviation of model's cross-validated classification accuracy

3. THE GREATER GLIDER CASE STUDY

3.1 Study Objectives

This case study used spatially referenced ecological data. The data was sourced from a number of habitat studies focused on a species of Australian tree-living marsupial. The aim of knowledge discovery was to find non-trivial, potentially useful—perhaps causal—relationships between habitat qualities and the distribution of the animals of interest. What was known of the behaviour of the animals, particularly their ability to glide, suggested that spatial relationships between habitat variables may be important. Hence the need for *geographic* knowledge discovery.

The case study had two principal research objectives. The first was to evaluate the validity of the RS-GKDD methodology through applying it to a realistic knowledge discovery problem. The second objective was to place rough set based knowledge induction in the context of other approaches to the machine induction of spatial and non-spatial ecological knowledge.

The size of the dataset for this study was not large (400 elements, seven attributes). However, seeking knowledge about spatial relationships was still computationally non-trivial.

3.2 Phase I - Analysis and Design

3.2.1 Study Background

This study focused on the distribution of arboreal marsupials in the forests of south-eastern Australia. Of particular interest was a species of possum *Peteroides volans,* commonly known as the greater glider. This species is regarded as, "an indicator of the suite of arboreal marsupials present in the area." (Stockwell et al. 1990: 36.) The 1600 hectare study area is located in the region's coastal ranges and tableland and is part of a wildlife reserve that has been intensively studied.

Stockwell et al. (1990) collated data from earlier investigations and used them to develop several models aimed at predicting the distribution of greater gliders. The best classification accuracies in this study were achieved using the machine induction algorithms ID3 and CART. These gave classification accuracies of 57.3 • 2.2% and 54.8 • 3.9%, respectively, on randomly chosen test data.

Pearson and McKay (1996) used the Rulefinder decision tree algorithm, which they describe as similar to CART (above), to conduct a series of experiments using the greater gliders data from the Stockwell et al. (1990) study. An element-specific baseline experiment gave a similar classification accuracy to the earlier CART results (52.3%, standard deviation 6.74%). By subsequently incorporating spatial relationships, classification accuracies of up to 71.3% were obtained.

Table 1 Modelling methods used by Stockwell et al. (1990)

Method/Algorithm	Model	Code
Multiple regression (stepwise)	Linear	MR
Principal components analysis	Decision table	PC
Knowledge acquisition	Rule base (describing expert's mental model)	KA
Machine induction - CN2	Decision tree	CN2
Machine induction - ID3	Decision tree	ID3
Machine induction - CART	Decision tree	CART
Maximally general classification	Decision table	MGC
Maximally specific classification	Decision table	MSC

Whigham (1966) also used the gliders dataset to develop a series of propositional grammars that were used to induce programs in the form of if–then–else knowledge using a genetic algorithm. An initial non-spatial grammar was extended to enable the genetic algorithm to induce statements about the conditions of elements *within some distance* of the focus element. The use of this spatial relation significantly improved the average classification accuracy of the six best induced programs to 64.1 • 2.0%.

A grammar allowing for distances up to 5 units and incorporating two "bias" rule statements brought the classification accuracy for the best induced model up to 67.2 • 1.7%.

Thus the greater gliders data had been investigated non-spatially and spatially using a number of methods, resulting in several moderately effective models. Pearson and McKay (1996) argued that their best models approached the limit set on classification accuracy by inconsistencies inherent in the dataset (i.e. noise).

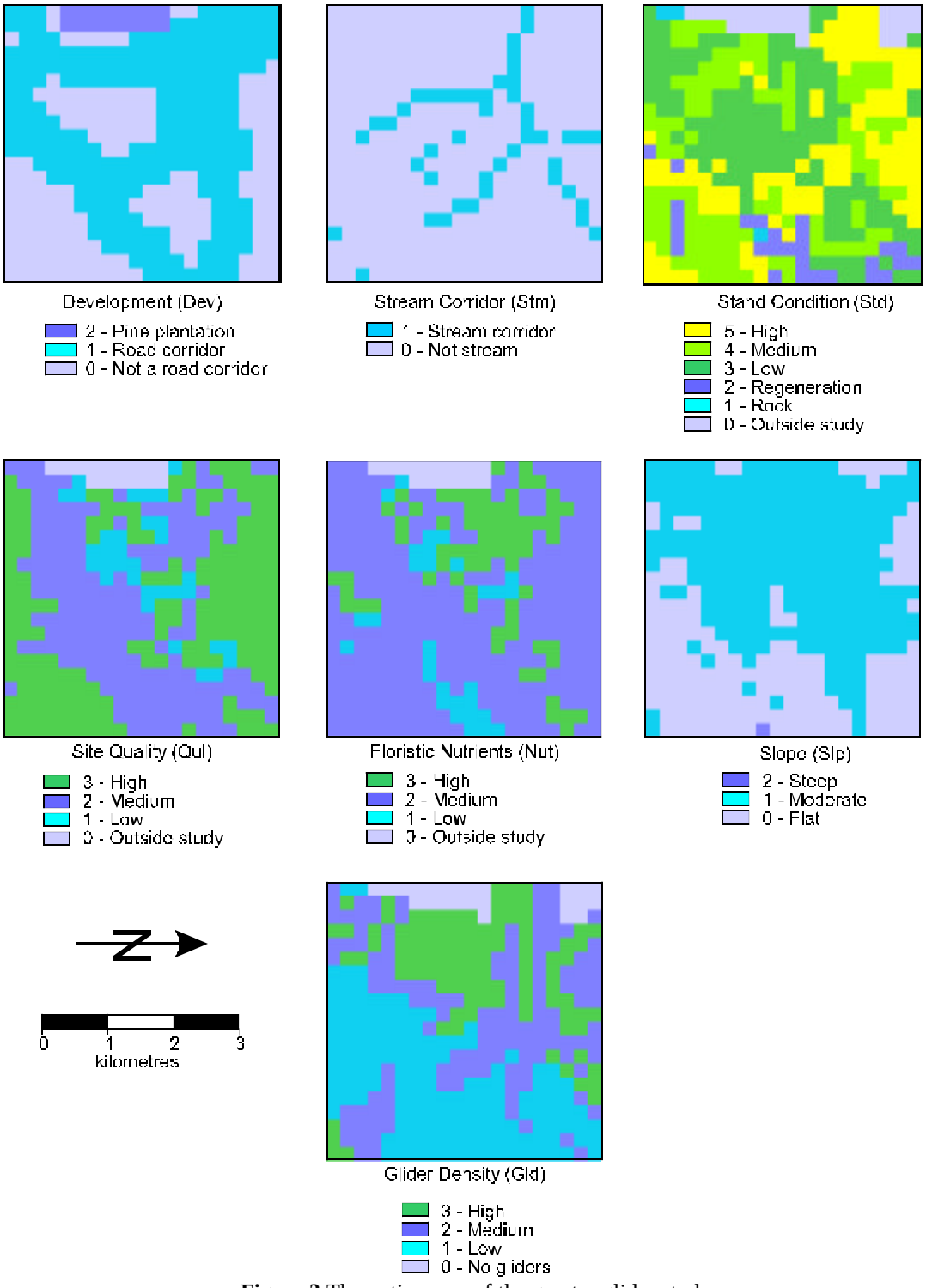

Figure 3 Thematic maps of the greater glider study area

3.2.2 Study Scope and Objectives

The principal objective of the case-study per se was to induce an rule-based geographic knowledge (model) that would accurately predict the population density of greater gliders in similar ecosystems. The prediction should call upon a minimal set of inexpensively acquired extensional knowledge. Both spatial and non-spatial output knowledges were investigated.

A secondary objective was to identify interesting associations that may lead to understanding causal mechanisms underlying the distribution of the gliders.

The scope of the knowledge discovery was confined to the study area covered by the Stockwell et al. (1990) dataset.

3.2.3 Data Assessment

For the present research, the original Stockwell et al. (1990) dataset was obtained from Whigham (pers. comm.). Figure 3 shows the forest inventory and gliders data as thematic maps. Each map is a $20 \bullet 20$ raster of $200 \bullet 200$ metre cells.

With regard to the measurement scales, there are no ratio or interval measures. Therefore, the data are amenable to direct analysis using the rough set approach. On the surface, all attributes except *Slope* and *Glider density* have nominal (categorical) measures. However, *Site quality* and *Floristic nutrients* could be turned into ordinal measures if the elements outside of the study area were eliminated. The presence of ordinal data is not a particular issue when applying RS-GKDD as the ordering relation is not used during knowledge discovery.

The size of the non-spatial rule space was calculated from the domain cardinalities of the six condition attributes and the single decision attribute. It is

$$(3+1)\times(2+1)\times(6+1)\times(4+1)\times(4+1)\times(3+1)\times 4 = 33{,}360,$$

which is quite moderate. Also, none of the condition attributes has a domain cardinality of more than six. The knowledge is therefore not excessively fine-grained and is suitable for use as is. It was recognised, however, that when using a context window, the number of attributes and therefore the rule space would increase substantially. There seemed little to be gained from generalising the data further in the hope of reducing the computational cost of spatial knowledge discovery.

3.3 Phase II - Data Preparation

There are no values missing from the data sets. This made the task of data preparation simply one of assembling the numerical data captured from the thematic maps into data files in the format expected by the RS-GKDD software. For non-spatial analysis, this consisted of building a text file with a column for each of the seven element attributes. Each row in the table corresponds to a focus raster cell, or spatial element. This dataset was called *NonSpatialKRS*, where KRS denotes "Knowledge Representation System".

For spatial knowledge discovery, the data was assembled with an additional set of condition attributes for each contextual element. Two spatial context data files were prepared, one with a window length parameter of one unit and consisting therefore of nine elements, including a focus element. This was dataset *Context1KRS*. A second data file, *Context2KRS*, was prepared for a context distance of two units. This resulted in 24 context elements and one focus element, giving 150 condition attributes and one decision attribute.

3.4 Phase III - Knowledge Induction, Reduction and Validation

For each of the three datasets above, knowledge induction was carried out using software that executes the RS-GKDD algorithms for:

- Optimal attribute selection, and rule induction.
- Selection of rules that are statistically significant, and 'near' the example data.
- Cross-validation of generated rules.

The selection of rules deemed 'interesting' was left to the interpretation and application phase of the methodology.

The greater gliders dataset was small enough (400 elements) for knowledge induction to be carried out without sampling for a spatial context of up to a distance of two units.

To enable comparison of rough set based rule induction with earlier greater glider machine learning experiments, the validation procedure adopted by Stockwell et al. (1990) was also used. This approach to validation required randomly splitting the dataset into a 'training' set of 200

elements and a 'testing' set of 200 elements. This was repeated to give six dataset pairs. Each model generated was used to classify both its 'training' and its 'testing' datasets. This enabled a mean classification accuracy and standard deviation to be calculated over the six paired datasets, for each induction method. Following Stockwell et al. (1990), Pearson and McKay (1996) and Whigham (1996), the relative accuracies of the models were compared using a Student's *t* test to evaluate the significance of differences between classification accuracy means. This approach to model generation and validation is inferior to *k*-fold cross validation. This is because it only uses half the dataset for model induction, compared with all the data in the *k*-fold approach. It also uses only half the dataset for each validation, compared with the alternative of $n(1-k)$ cases where k is typically 10, 20, or 40. As a consequence, validation using six paired data splits was only used here to enable comparison with the earlier research referred to above.

3.5 Phase IV - Knowledge Interpretation and Application

Application of the RS-GKDD methodology to the pre-existing greater gliders dataset over spatial context distances of zero[2], one, and two raster units (Table 2) enabled the induction of rule-based knowledges that are informative, in that most of the rules are not otherwise readily apparent; and potentially useful, because they achieve moderately effective predictions of the target variable, glider density.

Including spatial context supported the induction of spatial relationships that include distance and direction. With the gliders dataset, the knowledge obtained using spatial context is simpler than that obtained through element-specific knowledge discovery. As shown by Table 2, the knowledge has fewer attributes participating in a smaller number of induced rules. A comparison of classification accuracies shows that including spatial context results in models with better predictive power.

The stated objective of inducing an rule-based knowledge that accurately predicts the population density of greater gliders in similar ecosystems is achieved only to a moderate degree. The best models achieve no more than approximately 69.5% classification accuracy on training data. This appears to be approaching a limit set by noise inherent in the data, a conclusion reached by other workers (e.g. Pearson and McKay 1996). Allowing for this noise level, the cross-validated classification accuracy of 63.3% obtained using the distance-two context model is an encouraging outcome, particularly considering that this is achieved with a knowledge consisting of only five rules, each with two condition attributes. These rules are:

- **If** *Floristic nutrients* a distance of 2 units and in an eastern direction are *medium* **and** *Slope* a distance of 2 units and in a south-western direction is *flat* **then** *Glider density* is *low*.
- **If** *Floristic nutrients* a distance of 2 units and in an eastern direction are *high* **and** *Slope* a distance of 2 units and in a south-western direction is *moderate* **then** *Glider density* is *high*.
- **If** *Floristic nutrients* a distance of 2 units and in an eastern direction are *medium* **and** *Slope* a distance of 2 units and in a south-western direction is *moderate* **then** *Glider density* is *medium.*
- **If** *Floristic nutrients* a distance of 2 units and in an eastern direction are *high* **and** *Slope* a distance of 2 units and in a south-western direction is *flat* **then** *Glider density* is *medium.*
- **If** *Floristic nutrients* a distance of 2 units and in an eastern direction are *low* **and** *Slope* a distance of 2 units and in a south-western direction is *flat* **then** *Glider density* is *low*.

When applying rules such as these, a classification method must be used that deals with instances that do not exactly match any one rule. The basis used for matching instances and rules was discussed as part of the RS-GKDD methodology.

The results of the spatial versus non-spatial knowledge discovery confirm the hypothesised influence of glider mobility in determining population density at a location. Of particular interest is the importance of non-element characteristics in determining density at a element, as is the case with knowledge discovered using a context distance of two raster units.

The search for optimal attributes, which was undertaken to reduce computational cost when using spatial context, resulted in useful observations being made about which variables and relationships are important. The consequential reduced number of attributes needed to describe the target variable, *Glider density*, achieves one of the objectives of machine learning, namely to

[2] Knowledge induction with spatial context window of zero units is, in effect, non-spatial.

reduce the cost of acquiring knowledge about an expensive variable by using a minimal set of variables that more easily acquired, or readily available, and therefore typically less costly.

Table 2 Summary of spatial and non- spatial knowledge discovery outcomes for the gliders dataset

Ruleset	No. of Cond. Attrs.	No. of Rules	Training Data	20-Fold Cross Validation		Comp-utation Time, s
			Class. Acc., %	Mean Class. Acc., %	Std.Dev. Class. Acc., %	
Non-spatial - all attrs.						35.8
All rules	6	92	69.5	58.2	2.19	
Stat. sig. rules	6	69	69.2	58.0	2.16	
Stat. sig., 'nearest'	6	48	69.2	57.8	2.25	
Non-spatial - optimal						194.5
All rules	2	23	57.0	57.0	2.19	
Stat. sig. rules	2	10	56.5	53.2	2.74	
Stat. sig., 'nearest'	2	9	56.5	53.2	2.74	
Spatial dist. 1 - optimal						328.1
All rules	2	23	58.0	58.1	3.19	
Stat. sig. rules	2	9	58.0	58.1	3.19	
Stat. sig., 'nearest'	2	8	58.0	58.1	3.19	
Spatial dist. 2 - optimal						582.8
All rules	2	19	62.9	62.9	2.32	
Stat. sig. rules	2	6	62.9	62.9	2.32	
Stat. sig., 'nearest'	2	5	61.3	63.3	2.31	

Notes: Class. Acc. = Classification accuracy; Cond. Attrs. = Condition attributes; Stat. Sig. = Statistically significant

The improved result using spatial context carries a cost in terms of the computational effort required to obtain it (Computing time, Table 2). For instance, computing optimal attributes and inducing corresponding rules took approximately sixteen times longer for the two raster unit contextual data, compared with the non-spatial data.

The computational cost of knowledge discovery in the case of greater gliders is principally the cost of rule induction. The processing needed for data preparation, including compilation of context window attributes, is by trivial comparison.

3.6 Comparison of Knowledge Induction Methods

This section focuses principally on the machine learning aspects of the RS-GKDD methodology. As was discussed earlier, a number of machine learning methods have been tested on the gliders dataset, in both spatial and non-spatial contexts. To enable comparison between previous work and the present, experiments were carried out using six randomly selected paired datasets, an approach used by Stockwell et al. (1990). Their method for comparing the accuracy of models by using the Student's *t* test for significance of the difference between means was also adopted. As a method for determining classification accuracies, this method is inferior to *k*-fold cross validation because it makes less use of the information contained in the training data when inducing models. However, its use in this case does allow comparisons to be made with earlier work.

Stockwell et al. (1990) evaluated eight non-spatial models on the basis of their accuracy, comprehensibility, and efficiency. Accuracy was measured by the classification accuracy of the model on training and test data sets, as described above. Means and standard deviations were reported. These were used in the Student's *t* test to evaluate the null hypothesis that the models predict no better than random assignment to decision values. A cross-comparison of models was not done by these researchers. Such a comparison was undertaken by Whigham (1996) and his results are extended by the present work

in, as shown in Table 4 and Table Table 6, below.

Table 3 Non-spatial knowledges induced from the greater glider data

Intensional Knowledge Type (Induced Model)	Code	Classification Accuracy, %				Rules Equivalent	
		Training		Test			
		Mean	S.D.	Mean	S.D.	Mean	
Maximally general classification[1]	MGC	36.7	0.8	33.2	1.6	1.0	0.0
ID3 (version 3)[1]	ID3	60.7	0.8	57.3	1.6	11.2	0.2
Principal components analysis [1]				41		11.0	
Multiple regression[1]				45		4.0	
CN2[1]	CN2	50.8	2.6	45.2	2.2	5.2	0.5
Expert system[1]				48		29.0	
Maximally specific classification[1]	MSC	75.8	1.2	48.3	1.6	65.7	1.2
Rulefinder – Expt. 1[2]				52.5		26.0	
Program from CFG – G_{ggd} [3]	CFG	57.9	1.4	52.5	3.4	6.0	
CART[1]	CART	61.2	3.5	54.8	3.9	5.0	0.9
RS-GKDD[4]							
{Slp}	RS1	46.8	2.9	46.8	2.9	2.0	0.0
{Std, Slp}	RS2	57.1	3.7	52.0	3.7	7.8	0.4
{Std, Qul, Slp}	RS3	63.6	3.9	57.4	3.9	14.8	1.5
{Std, Qul, Nut, Slp}	RS4	68.1	4.8	59.6	4.8	21.7	1.9
{Dev, Std, Qul, Nut, Slp}	RS5	70.2	4.6	58.4	4.6	32.7	2.0
{Dev, Stm, Std, Qul, Nut, Slp}	RS6	71.3	3.3	58.6	3.3	35.2	2.0

Notes:
1. Stockwell et al. (1990).
2. Pearson and McKay (1996). Number of rules deduced from reported decision tree.
3. Whigham (1996). Program was evolved from a context free grammar using a genetic algorithm. Number of rules deduced from reported program. Model uses an ordering relation (>).
4. This research.

The measure of comprehensibility used by Stockwell et al. (op. cit.) was the number of conjunctive rules, or their equivalent in the case of tree models. They treated alternatives in disjunctions as separate rules. This measure is consistent with the number of rules reported by the rough set induction algorithm. Neither Pearson and McKay (1996) nor Whigham (1996) report a rule number measure for the models they induce. The rule numbers given in the Table 3 and Table 5 below were derived by applying Stockwell et al.'s criteria to instances where models are reported in sufficient detail.

The measure of efficiency described by Stockwell et al. is the induction time complexity of a decision tree algorithm. Unfortunately, this is not readily applied to non-tree algorithms. As a consequence, an efficiency comparison between models reported in the literature and the rough set based models generated here is not attempted.

3.6.1 Non-Spatial Knowledges

The various non-spatial models developed on the gliders data during this and previous research are drawn together in Table 3. The table summarises the accuracy and comprehensibility measures for a number of non-spatial knowledges (i.e. models).

Table 5 confirms and extends the statistical analysis of non-spatial model significances presented in Whigham 1996 (as Table 4.18). Abbreviated codes used to identify models in Table 1 are matched with their descriptions in Table 3.

Table 4 Comparison of accuracies of induced non-spatial models on test data

	MGC	CN2	RS1	MSC	RS2	CFG	CART	ID3	RS3	RS5	RS6	RS4
MGC		–	–	–	–	–	–	–	–	–	–	–
CN2	**Sig**			-	-	–	–	–	–	–	–	–
RS1	**Sig**				-	-	-	-	-	-	-	-
MSC	**Sig**	Sig				-	-	–	–	–	–	–
RS2	**Sig**	Sig	Sig					–	–	–	–	–
CFG	**Sig**	**Sig**	Sig	Sig				-			-	-
CART	**Sig**	**Sig**	Sig	Sig								
ID3	**Sig**	**Sig**	**Sig**	**Sig**	Sig	Sig						
RS3	**Sig**	**Sig**	**Sig**	**Sig**								
RS5	**Sig**	**Sig**	**Sig**	**Sig**	Sig							
RS6	**Sig**	**Sig**	**Sig**	**Sig**	Sig	Sig						
RS4	**Sig**	**Sig**	**Sig**	**Sig**	Sig	Sig						

Notes:
1. Each cell answers the question, "Is the row model a significant improvement on the column model in terms of mean test data classification accuracy?"
2. The statistical significance of the difference between the means is assessed using Student's *t* statistic.
3. "Sig" denotes significant improvement at a confidence level of 95%.
4. "**Sig**" denotes a significant improvement at a confidence level of 99%.
5. "-" indicates that the row model is significantly worse than the column model at a confidence of 95% (at 99% if "–").
6. A blank entry indicates that test has failed to show that the mean classification accuracies of the models are significantly different at the 95% level.

In the columns without "Sig" codes identify the superior non-spatial models—ones that no others can significantly improve upon. At 95% confidence, they are ID3, CART, RS3, RS4, RS5, and RS6. The models most frequently achieving significantly better classification accuracies are identified by the table rows with the most "Sig" entries. Also at 95% confidence, these are ID3, RS4 and RS6. These can be viewed as the 'best' model choices in terms of classification accuracy. In terms of comprehensibility, and referring to Table 3, the RS4 knowledge has with an average of 94% more rules equivalents than the ID3 one. Thus for the induction of "interesting" rules, the method ID3 appears the best choice for this non-spatial analysis of gliders data. However RS3 has a mean accuracy not significantly different from ID3 and only has an average of 3.6 more rules in its model.

More generally, the blank region in Table XXXT5 formed by the intersection of the six superior models identified above shows that the classifying powers these models are not significantly different (95% confidence). They therefore cannot be excluded from consideration when dealing with datasets having similar characteristics. Other matters such as model comprehensibility, model type (i.e. rules, decision trees, logic programs, etc), resource requirements, efficiency, and even software availability, are factors which will guide the choice of the most suitable knowledge induction strategy, or strategies, for a given knowledge discovery task.

3.6.2 Spatial Knowledges

Table 5 includes spatial models generated during this work, along with those reported by previous researchers. A statistical analysis comparing all these methods is presented in Table 6.

Table 5 Spatial knowledges induced from greater glider data

Knowledge Type (Induced Model)	Code	Classification Accuracy, %				Rules Equivalent	
		Train		Test			
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Non-Spatial							
ID3 (version 3)[1]	ID3	60.7	0.80	57.3	1.60	11.2	0.20
RS-GKDD[4]							
{*Std, Qul, Nut, Slp*}	RS0-4	68.1	4.80	59.6	4.80	21.7	1.90
{*Dev, Stm, Std, Qul, Nut, Slp*}	RS0-6	71.3	3.30	58.6	3.30	35.2	2.00
Spatial							
Rulefinder[2]							
Expt. 3a	RF3a	-	-	70.5	5.68	14.0	-
Expt. 4a	RF4a	-	-	71.3	4.64	-	-
Expt. 5a	RF5a	-	-	65.5	6.78	20.0	-
Program from CFG[3]							
Gggd–spatial	CFGs	67.3	2.00	64.1	2.00	5.0	-
Gggd–spatial–biased	CFGsb	71.8	1.30	67.2	1.70	7.0	-
RS-GKDD[4]							
{*1-SWSlp*}	RS1-1	51.7	1.59	49.5	1.59	2.0	0.00
{*1-SWSlp, 0-L-Std*}	RS1-2	59.4	3.91	53.4	3.91	7.7	0.52
{*1-SWSlp, 0-L-Std, 1-SWDev*}	RS1-3	62.3	2.91	58.4	2.91	12.5	0.55
{*1-SWSlp, 0-L-Std, 1-SWDev, 1-W-Slp* }	RS1-4	65.7	2.67	58.9	2.67	17.3	1.37
{*1-SWSlp, 0-L-Std, 1-SWDev, 1-W-Slp, 1-SE-Flo*}	RS1-5	71.1	2.80	59.5	2.80	23.2	1.47
{*1-SWSlp, 0-L-Std, 1-SWDev, 1-W-Slp, 1-SE-Flo, 1-SW-Ste*}	RS1-6	74.9	1.49	61.9	1.49	30.2	2.14
{*2-SWSlp*}	RS2-1	55.7	2.95	54.8	2.95	2.0	0.00
{*2-SWSlp, 2-E-Ste*}	RS2-2	62.2	2.89	61.8	2.89	5.2	1.17
{*2-SWSlp, 2-E-Ste, 2-E-Flo*}	RS2-3	67.6	4.35	64.7	4.35	9.0	0.89
{*2-SWSlp, 2-E-Ste, 2-E-Flo, 2-NWStm*}	RS2-4	70.6	5.37	69.7	5.37	11.7	1.63
{*2-SWSlp, 2-E-Ste, 2-E-Flo, 2-NWStm, 0-L-Std*}	RS2-5	76.7	2.40	69.4	2.40	24.8	1.83

Notes:
1. Stockwell et al. (1990).
2. Pearson and McKay (1996). The number of rules is deduced from the reported decision tree. Only 10-fold cross validated results from all data are reported.
3. Whigham (1996). Program was evolved from a context free grammar using a genetic algorithm. Number of rules deduced from reported program. Model uses an ordering relation (i.e. >).
4. This research.

Table 6 Comparison of accuracies of induced spatial models on test data

	ID3	**RS0-6**	**RS0-4**	**RS1-1**	**RS1-2**	**RS2-1**	**RS1-3**	**RS1-4**	**RS1-5**	**RS2-2**	**RS1-6**	**CFGs**	**RS2-3**	**RF5a**	**CFGsb**	**RS2-5**	**RS2-4**	**RF3a**	**RF4a**
ID3				**Sig**						-	–	–	-	-	–	–	–	–	–
RS0-6				**Sig**								-	-		–	–	–	–	–
RS0-4				**Sig**											-	–	-	-	–
RS1-1	–	–	–			-	–	–	–	–	–	–	–	–	–	-	-	-	-
RS1-2								-	-	–	–	–	–	–	–	–	–	–	–
RS2-1				Sig					-	–	–	–	-	-	–	–	–	–	–
RS1-3				**Sig**							-	-	-		–	–	–	–	–
RS1-4				**Sig**	Sig							-	-		–	–	–	–	–
RS1-5				**Sig**	Sig	Sig						-			–	–	–	–	–
RS2-2	Sig			**Sig**	**Sig**	**Sig**									-	–	-	-	–
RS1-6	**Sig**			**Sig**	**Sig**	**Sig**	Sig								–	–	-	-	–
CFGs	**Sig**	Sig		**Sig**	**Sig**	**Sig**	Sig	Sig	Sig						-	–		-	-
RS2-3	Sig	Sig		**Sig**	**Sig**	**Sig**	Sig	Sig											
RF5a	Sig			**Sig**	Sig	Sig													
CFGsb	**Sig**	**Sig**	Sig	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	Sig	**Sig**	Sig							
RS2-5	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**							
RS2-4	**Sig**	**Sig**	Sig	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	Sig	Sig								
RF3a	**Sig**	**Sig**	Sig	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	Sig	Sig	Sig							
RF4a	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	**Sig**	Sig							

Notes:

1. Each cell answers the question, "Is the row model a significant improvement on the column model in terms of mean test data classification accuracy?"
2. The statistical significance of the difference between the means is assessed using Student's *t* statistic.
3. "Sig" denotes significant improvement at a confidence level of 95%, (of 99% if "**Sig**").
4. "-" indicates that the row model is significantly worse than the column model at a confidence of 95% (of 99% if "–").
5. A blank entry indicates that test has failed to show that the mean classification accuracies of the models are significantly different at the 95% level.
6. The best of the non-spatial models are shown shaded.

Referring to the pattern of results in Table 6, the superior models are, as before, those without significant competing methods. These are the seven models RS2-3, RF5a, CFGsb, RS2-5, RS2-4, RF3a, and RF4a. Of these, the best models are CFGsb, RS2-5, RF3a, and RF4a. At 95% confidence these are significantly more accurate than all models except the seven superior ones already identified. Because the row-column intersections in Table 6 for these seven models are blank, there is no reason on the basis of these statistical tests to regard any one of them as the best in terms of classification accuracy.

When comparing spatial and non-spatial models, it is particularly interesting to note that, in terms of classification accuracy, the three best non-spatial models (ID3, RS0-4 and RS0-6) are significantly better than only one of the spatial models (RS1-1). What is more, the spatial models can often achieve their superior classification accuracies while requiring fewer rules than similar but non-spatial models derived using the same induction method.

4. CONCLUSIONS

With regard to methods of inducing geographic knowledges (or models), the rough set based algorithm forming the central part of the RS-GKDD methodology generated models that can be ranked with the best published models, in terms of classification accuracy. This was the case for both spatial and non-spatial models. However, in achieving these competitive classification accuracies, the rough set based models did tend to have more rules.

The ID3 method is able to produce a non-spatial model that achieves a superior classification accuracy with relatively few rules. This method is therefore responsible for the "best" non-spatial model in terms of accuracy and comprehensibility. It has yet to be tested with the gliders dataset for the induction of models that include spatial relationships. Such an investigation appears a worthwhile research task.

For the greater gliders dataset, the inclusion of spatial relationships while inducing knowledge almost invariably yields models that achieve better classification accuracies, and with relatively fewer rules, or rule equivalents. That is, the induced *spatial* knowledge is more interesting and useful than the *non-spatial* knowledge. However, these benefits come at a significantly increased computational cost.

REFERENCES

Aldridge, C.H. (1998) *A Theory Of Empirical Spatial Knowledge Supporting Rough Set Based Knowledge Discovery in Geographic Databases.* PhD Thesis, University of Otago, Dunedin, New Zealand.

Cassie, R.M. (1954) "Some uses of probability paper for the graphical analysis of size frequency distributions," *Australian Journal of Marine and Freshwater Research*, no. 5, pp. 513-522.

Cohen, P.R., and Feigenbaum, E.A., editors (1982) *The Handbook of Artificial Intelligence - Vol. 3* Addison-Wesley Publishing Co: Reading, MA, U.S.A.

Duntsch, I., and Gediga, G. (1998) "Uncertainty measures of rough set prediction", *Artificial Intelligence*, no. 106, pp. 109-137.

Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling Plans. *CBMS-NSF Regional Conference Series in Applied Mathematics, no. 38*. Society for Industrial and Applied Mathematics: Philadelphia, PA, U.S.A.

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. (1996) "The KDD process for extracting useful knowledge from volumes of data", *Communications of the ACM*, v. 39, no. 11, pp. 27-34.

Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C. (1992) "Knowledge discovery in databases: An overview." *AI Magazine*, pp. 57-70.

Gawrys, M., and Sienkiewicz, J. (1993) *Rough Set Library User's Manual (Version. 2.0, September 1993).* Institute of Computer Science, Warsaw University of Technology, Warsaw, Poland.

Gunter, B. (1997) "Tree-based classification and regression - Part 2: Assessing classification performance." *Quality Progress*, Dec, pp. 83-84.

Kennedy, G.J. (1993) *A Systematic Approach to the Specification of an Information Systems Development System.* PhD Thesis, Department of Information Science, University of Otago, Dunedin, New Zealand.

Koperski, K., and Han, J. (1995) "Discovery of spatial association rules in geographic information systems." In Egenhofer, M.J., and Herring, J.R., editors, *Advances in Spatial Databases: 4th International Symposium, SSD '95, Portland, ME, U.S.A., August 1995, Proceedings. Lecture Notes in Computer Science*: Goos, G. and Hartmanis, J. (Eds.) no. 951, pp. 47-66. Springer: Berlin.

Maddison, R.N. (1983), *Information System Methodologies*.

Pawlak, Z. (1982) "Rough sets." *International Journal of Computer and Information Sciences*, v. 11, no. 5, pp. 341-356.

Pawlak, Z. (1991) *Rough Sets: Theoretical Aspects of Reasoning About Data.* Theory and Decision Library. Series D: System Theory, Knowledge Engineering and Problem Solving, no. 9. Kluwer Academic Publishers: Dordrecht, The Netherlands.

Pearson, R.A., and McKay, R.I. (1996) *Spatial Induction for Natural Resource Problems: A Case Study in Wildlife Density Prediction*. Technical Report CS0298, School of Computer Science, Australian Defence Academy: Canberra, Australia.

Shannon, C.E. (1949) “Communication in the presence of noise.” *Proceedings of the IRE*, v. 1949, no. Jan, pp. 10-21.

Stockwell, D.R.B., Davey, S.M., Davis, J.R., and Noble, I.R. (1990) “Using induction trees to predict greater glider density,” *A I Applications in Natural Resource Management*, v. 4, no. 4, pp. 33-43.

Stone, M. (1974) “Cross-validatory choice and assessment of statistical predictions,” *Journal of the Royal Statistical Society B: Applied Statistics*, v. 36, no. 2, pp. 111-147.

Whigham, P.A. (1996) Grammatical Bias for Evolutionary Learning. *PhD Thesis, University College, University of New South Wales: Canberra, Australia.*