

Discerning Landslide Hazard Using a Rough Set Based Geographic Knowledge Discovery Methodology

Colin H Aldridge

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

Presented at SIRC 99 – The 11th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
December 13-15th 1999

ABSTRACT

This study undertook an investigation of a 64 km² area near Dunedin, New Zealand, with the principal objective of inducing a geographic model that would identify the extent to which locations are at risk of landslides. The study area has several geographic, geological, and pedological themes available. The condition attributes available in these extensional knowledges were: *Coal seam, Elevation, Faults, Fault zone, Slope zone, Subsidence, Lithology, and Soils*. The single decision attribute was *Landslide*, a binary attribute indicating either the presence or absence of a landslide at a location. Throughout the knowledge induction process, a 'location' was a 100 metre by 100 metre (one hectare) raster element.

For each of the datasets, the induction of a geographic model, as production rules, took place according to the RS-GKDD methodology. The first step was an heuristic search for a subset of optimal condition attributes. These were found to be *Elevation* and *Lithology*. The second stage took these optimal attributes and induced a model consisting of a set of 'nearest' statistically significant rules. The nineteen rules of this model achieved 40-fold cross validation classification accuracies of 55.6% (standard deviation 2.0%) for landslide present decisions, and 81.0% (standard deviation 0.9%) for landslide absent decisions. For all decisions, the classification accuracy was 78.9% (standard deviation 0.7%).

The study concludes that, firstly, the rules making up these knowledges are of potential interest to people needing to make land use decisions about locations in or near the study area. Secondly, the RS-GKDD methodology appears applicable to the induction of geographic knowledge about landslides in other districts, provided at least the landscape themes *Elevation* and *Lithology* are available.

Keywords and phrases: rough set theory, landslide hazard, knowledge discovery in databases, GIS

The real voyage of discovery is not in seeking new landscapes but in having new eyes—Marcel Proust

1 INTRODUCTION

This paper reports on the application of a knowledge discovery methodology to multiple theme geographic, geological, pedological and historic landslide data from a 64 square kilometre area south-west of Dunedin, New Zealand (Figure 3). The purpose of this research was to demonstrate that *rough set based geographic knowledge induction*—the RS-GKDD methodology—could generate a model able to usefully indicate locations at risk of landslide. To do this, the investigation attempted to discover interesting and potentially useful relationships between landscape variables and the locations of pre-existing landslides in the study area.

In the region concerned, a reliable and objective means of identifying the factors associated with unstable land, and hence estimating the degree of landslide hazard, would be of importance to both property owners and territorial authorities. There are many other parts of New Zealand that have land stability problems and a

method for estimating landslide hazards in those regions would assist local authorities in meeting their obligations for the management of hazards under the Resource Management Act 1991 (N.Z Government 1991).

In the following sections, after outlining the research leading up to this paper, the RS-GKDD methodology is briefly described. The objectives of the landslides study are then introduced. The geological context is given, 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 PREVIOUS WORK

This paper extends research into the development of a hazards information system for the Dunedin region, New Zealand — what was formerly known as the Dunedin Pilot Hazards study (Aldridge and Benwell 1993; Aldridge and Benwell 1993a; Aldridge et al. 1993; Glassey et al. 1994). Previous work had identified a need for, amongst other things, a Landslide Hazards Analysis and Modelling System (Aldridge and Benwell 1993) to infer the existence and extent of landslide hazards from relevant thematic map data.

In 1994, as part of the Dunedin Pilot Hazards study, Hancox (cited in Glassey et al. 1994) developed a landslide susceptibility zonation method based on the landscape attributes: lithology; rock strength and stability; pre-existing landslides; land slope; groundwater; and the existence of landfill on slopes. Values and relative weightings were assigned to these factors according to their perceived influence on the incidence of landslides. The weighted sum of the factor values at a location became the landslide susceptibility rating. It was proposed that these susceptibility ratings should be used to compile a map of landslide hazard zones. With regard to the use of a weighted sum of factors, the mathematical operations of addition and multiplication cannot be validly applied to nominal and ordinal scale data (Krantz et al. 1971). Unfortunately, several of the Hancox model inputs are measures having these scales. Therefore, the validity of the modelling method and its resulting susceptibility ratings must be seriously questioned.

Subsequent to the above research, Aldridge (1988) developed a methodology for Rough Set Based Geographic Knowledge Discovery in Databases (RS-GKDD). Because of its use of rough set theory, this approach to model development is particularly suited to processing digital map data that includes nominal scale measures. The availability of this methodology created an opportunity for applying a computer-based method to create a predictive model for landslide hazards by utilising the multiple-theme land-based data available for the earlier Dunedin Pilot Hazards Study area.

3 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 choropleth 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 choropleth maps and their digital representation (Aldridge 1988). The RS-GKDD methodology also has objectives, scope, constraints, assumptions, theory, procedures, algorithms, and tools (below). 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).

3.1 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* extensional 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*.

3.2 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. The input data are the digital equivalents of choropleth maps that, together, comprise a sufficient number of themes as to render manual interpretation difficult or impossible. New knowledge is induced as a production rule model, one that should fulfil its user's requirements for nontrivial, previously unknown, and potentially useful information (c.f. Frawley et al. 1992).

3.3 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 fully utilise the additional information in ordinal, interval and ratio scale measures.

3.4 Assumptions

The principal assumptions of the RS-GKDD methodology are:

- A *data validity assumption*, namely that the input data—the digital equivalents of choropleth 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 (choropleth 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.

3.5 Theory

3.5.1 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.

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?

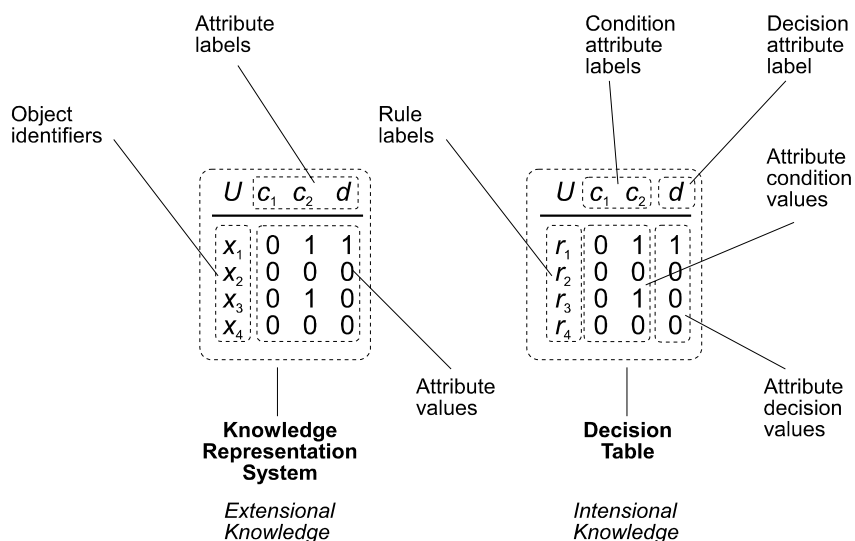


Figure 1 Knowledge representation system generalised into a corresponding decision table

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 also provided to illustrate the application of the theory.

3.5.2 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.

3.6 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.

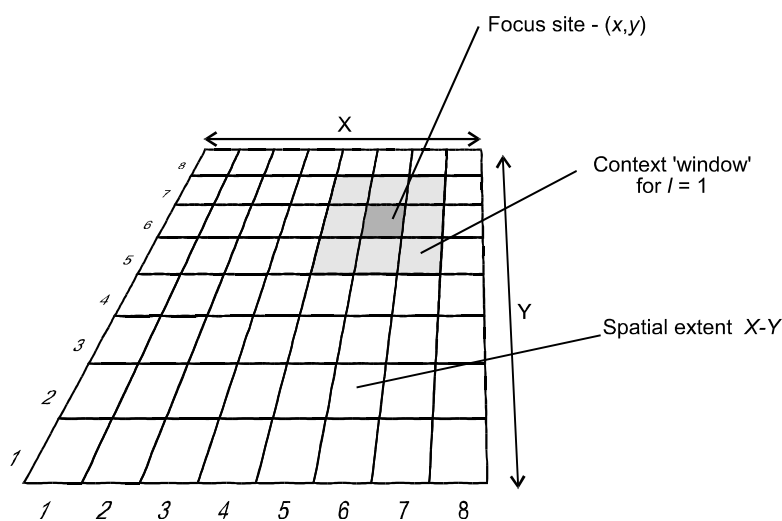


Figure 2 Relationship of context 'window' to a raster spatial extent

3.6.1 Rasterisation

Knowledge induction from several choropleth 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.

3.6.2 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).

3.6.3 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 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:

1. Attribute category codes as a look up table for use in assigning meaningful category names to the numeric codes used in computations.
2. 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.

3.6.4 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 is weighted towards positive instances.

3.6.5 Rough Set Based Rule Induction

The rough set knowledge induction process is implemented in an algorithm that applies the theory described above. 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.

$$\text{For example, KRS, } S = \begin{bmatrix} U & a & b & c & d & e \\ x_1, y_1 & 1 & 2 & 0 & 1 & 1 \\ x_2, y_1 & 1 & 2 & 0 & 1 & 1 \\ x_3, y_1 & 0 & 0 & 1 & 1 & 0 \\ x_4, y_1 & 2 & 1 & 0 & 2 & 1 \\ x_5, y_1 & 2 & 1 & 0 & 2 & 1 \\ x_1, y_2 & 0 & 0 & 0 & 2 & 2 \\ x_2, y_2 & 0 & 0 & 1 & 1 & 0 \\ x_3, y_2 & 0 & 1 & 0 & 2 & 1 \\ x_4, y_2 & 2 & 1 & 0 & 2 & 2 \\ x_5, y_2 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \text{ would be reduced to decision table, } D = \begin{bmatrix} c_1 \rightarrow e_0 \\ b_2 \rightarrow e_1 \\ a_2 \rightarrow e_1 \\ a_0 \wedge b_1 \rightarrow e_1 \\ b_0 \wedge c_0 \rightarrow e_2 \\ a_2 \rightarrow e_2 \end{bmatrix}.$$

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.”

3.6.6 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. *Support* 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 can select *interesting rules* 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).

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.

3.6.7 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).

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.

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.

3.6.8 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.

3.7 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.

3.8 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.

4 LANDSLIDE STUDY OBJECTIVES

The first objective of the study was to take existing geographic, geological, and pedological thematic data from the study area and use it to induce (discover) a rule-based model that is:

- Interesting and potentially useful to those wishing to better understand the relationships between landscape attributes and the occurrence of landslides in the district.
- Able to take available data about a location and reliably infer the relative risk of landslide events occurring there. That is, the model should be able to successfully discriminate between locations at significant risk and those relatively safe from landslides.
- Able to achieve the above while being minimal in terms of the number of landscape attributes used.

The second objective was to demonstrate that the RS-GKDD methodology is potentially applicable to discovering useful landslide knowledge in other districts for which suitable thematic data is available.

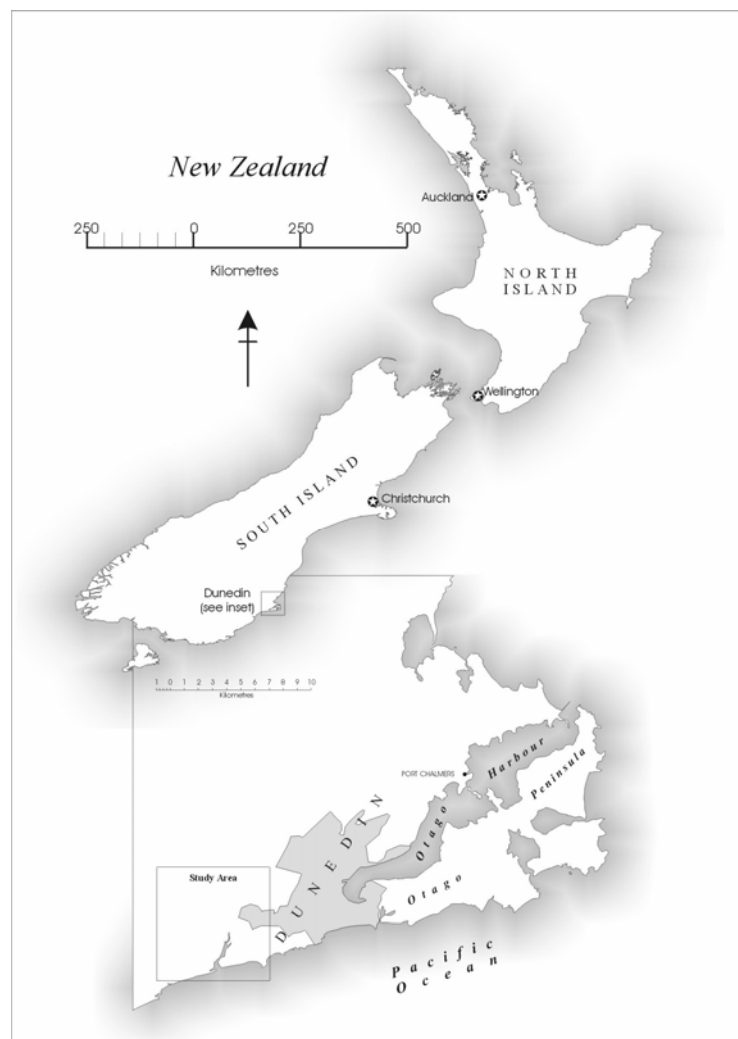


Figure 3 Landslide hazard case study location map

5 STUDY AREA

The geographic scope of the study was the eight kilometre by eight kilometre area shown in Figure 3. This is a rectangular region with vertices defined by New Zealand Map Grid coordinates I44/030710, 110710, 110790, 030790. The data used in the study were obtained from digitised thematic maps of the area of interest. Although a significant amount of knowledge about the study area is held in geological reports, field notes, drilling logs, etc., it is not readily compatible with the RS-GKDD methodology and was not used for the development of models.

5.1 Geological Setting

McKellar (1990: 5) summarises the geology of this and the immediately surrounding area as follows:

Outliers of volcanic rocks are underlain by eastward dipping...marine sediments resting on a thin sequence of ...quartz gravel, sand, and clay with some coal seams. These in turn rest on a basement of...[schist]. [More recent] alluvial gravel and sand formations occupy the [main] valleys. Most of the map area is mantled by colluvium and loess regolith.

And the geological hazards:

Various geological hazards place constraints on urban development. The most important of these being landsliding which originates mostly in clay-rich marine sediments of the Saddle Hill Sandstone and Abbotsford Formation, but also occurs in the regolith. The latter is derived partly from these sediments, and from loess and colluvium which mantle the hill slopes. The landslide hazard which exists in some places was highlighted in 1979 by the east Abbotsford landslide.

5.2 The Abbotsford Landslide

The east Abbotsford landslide was an event that resulted in the destruction of 69 houses. It displaced four hundred and fifty residents, fortunately without loss of life (Anonymous 1980; McKellar 1990). Not surprisingly, there has been a significant interest in assessing landslide hazard in the surrounding district, leading in part to the Dunedin Hazards Information System research.

6 DATA

6.1 Data Assessment

The following geological themes were made available for the case study by the Institute of Geological and Nuclear Sciences (IGNS), Dunedin: landslides; lithologies; faults; and mine subsidence. These had been digitised from McKellar's (1990) 1:25,000 geological map.

Also obtained through the IGNS was a digital version of the relevant part of the Terralink New Zealand Ltd. 1:50,000 topographic map. The map had been digitised at a scale of 1:25,000 into the themes: contours, hydrology, and rivers. The IGNS had generated a slope coverage from the contours data by using an intermediate Digital Elevation Model (DEM). Details of the transformation are given in Glassey et al (1994).

A digital soils map for the area was purchased from Landcare Research.

All of the above data was received in the form of Arc/Info export files. Table 1 summarises the data types of these "coverages" (i.e themes).

6.2 Data Preparation

6.2.1 Rasterisation

Since none of the data were in raster form, the first task of data preparation was to decide on an appropriate raster element dimension. An examination of the frequency distribution of landslide polygon areas showed they had a median size of 0.43 hectares, with a range of 0.01 to 73.3 hectares. The substantial skewing of the frequency distribution toward towards smaller polygons is illustrated by Figure 4 Table 4. A choice of a raster element size of 0.02 hectares, sufficient to support representation of all landslide polygons would result in a computationally costly 320,000 element raster. A decision was therefore made to focus on the more significant landslides by adopting a compromise one hectare raster element. This size supports representation of landslides over 0.5 hectares and corresponds to a raster spacing of one hundred metres. During data preparation, all map themes were extracted to cover the same 8100 by 8100 metre region and rasterised at the chosen element size, resulting in an 81 row by 81 column raster having a total of 6, 561 elements.

Table 1 'Raw' land-based data obtained for the landslides study

Coverage	Spatial Data Type	Attribute Data Type	No. of Attribute Classes
Topographic contours	Line (Arc)	Interval	-
Hydrology	Polygon	Nominal	4
Rivers	Line	Nominal	2
Lithology	Polygon	Nominal	25
Coal seam	Polygon	Nominal	2
Faults	Line	Nominal	4
Subsidence	Point	Nominal	2
Soils	Polygon	Nominal	44
Landslides	Polygon	Nominal	5

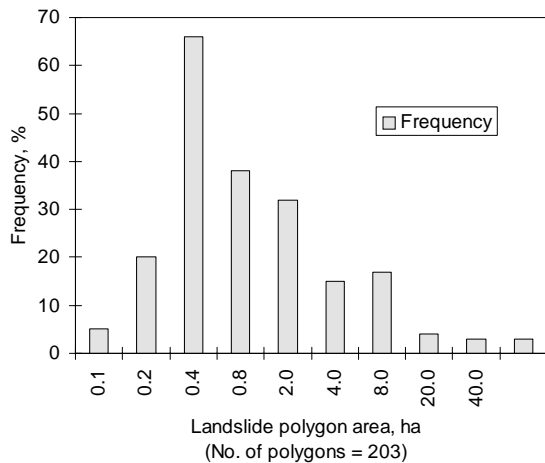


Figure 4 Frequency distribution of landslide polygon areas

inevitably results in their conversion to region features, in this case with a nominal one hundred metre width corresponding to the raster cell size. This result was considered acceptable for the coal seam and substance themes in view of their presumed localised influence on land instability. Allowance for a possible wider influence of fault activity was made by creating arbitrary buffer zones around fault-lines, resulting in the fault zone theme (Figure 6).

6.2.2 Re-classification (Generalisation)

The topographic data required a more extensive treatment. In the first instance, a digital elevation model (DEM) was developed from the contours data, based on the chosen raster size. The transformation was carried out using the Spatial Analyst module of ESRI's ArcView 3.0. The resulting DEM expressed altitudes in terms of floating point numbers, which is a much too specialised knowledge for the discovery of comprehensible

The landslides data as received was work in progress and the classification of landslide polygons was far from complete. As a consequence, the only landslide classification that could be used consistently across the study area was the presence versus the absence of a landslide. The original dataset was therefore reclassified by grouping all landslides together, making the landslide decision a binary true-false one (Figure 5).

The next problem with the data as received was with those describing linear features: contours, coal seam, subsidence, and faults. Rasterization of linear features

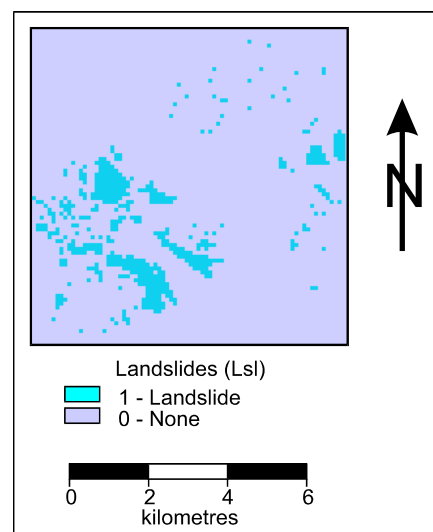


Figure 5 Map of decision theme - landslides

rules.² A simple approach to generalising a set of such numbers is therefore to reduce the number of significant figures. For instance, the altitudes could be rounded to 10 metre intervals. However, rather than taking such an arbitrary approach, the method of Cassie (1954), which is part of the RS-GKDD methodology, was used to analyse the frequency distribution of altitudes. This identified values that are able to separate component groups in the distribution. The generalised altitude intervals chosen on this basis are given in the elevation theme legend in Figure 6.

Slope was considered likely to be a significant determinant of landslide hazard. As a consequence, the digital elevation model previously developed from the topographic data was used to derive a corresponding slope model. Again, the ArcView 3.0 spatial analysis module was used for the transformation. As with the elevation model, Cassie's method (op. cit.) was used to identify sub-population groups in the frequency distribution of floating point slope values. These groups are generalisations about slopes in the study area, creating classes that are termed flat, low, moderate, and steep in the slope zone legend in Figure 6.

Finally, the lithology and soils themes have 25 and 44 attributes, respectively. They therefore comprise relatively specialised knowledges. The possible generalisation of lithological or soil classes was outside the scope of this research as the task would require specialised geological and pedological knowledge and skills. However, the use of fine-grained data does have the advantage of offering potential for a fine grained description of landslide conditions. In this regard, then, the use of the rock type and soils data without reclassification is not necessarily a problem. If the induced knowledge proved too fine-grained then the question of generalising these inputs could be revisited.

The hydrology and river themes were excluded from the knowledge induction phase on the basis that they describe parts of the landscape that are irrelevant to a consideration of landslide hazard. Of the available and transformed data, there remained eight condition attributes: *CoalSeam*, *Elevation*, *Faults*, *FaultZone*, *Lithology*, *SlopeZone*, *Soil*, and *Subsidence*, and one decision attribute, *Landslides*. The data for each of these attributes was collated into a single tabular *knowledge representation system* (KRS), with columns corresponding to attributes and rows corresponding to raster element objects.

6.2.3 Sampling

The number of raster element objects (6561) meant that applying the algorithm *Hill-Climbing Search For Optimal Attributes* to all raster elements was computationally costly, despite the dataset having 'only' eight attributes. Even though the search was able to terminate after examining subsets of just three attributes, it still took 55 minutes on the 233 MHz Pentium PC used. This suggested that using randomly sampled subsets of the data could be advantageous, provided satisfactory models could still be induced.

Another consideration suggesting that sampling the dataset might be desirable was the relatively small proportion of positive examples (8.23%) of landslides. In this study, positive decision examples are those raster elements with landslides. The 'signal' from raster elements that provide examples of landslides could possibly be swamped by the 'noise' of a surfeit of examples of the non-existence of landslides. A immediately appealing response to this problem, would be to concentrate solely on positive examples, as, indeed, the focus of the study is on landslides. Unfortunately, this would logically lead to inducing a single rule: "Always a landslide." This rule would guarantee a 100% classification accuracy of the (positive) examples considered. On the other hand, the rule would never predict the *absence* of landsliding. Thus the superficial response of only inducing knowledge from positive examples must be rejected. Consideration of both positive and negative examples is needed to induce a satisfactory model. A compromise was achieved in this case study by using the RS-GKDD algorithm *Create Sample Of Binary KRS*. This algorithm provides for not only specifying the proportion of examples in the sample, but also the ratio of positive to negative examples. When applying the algorithm to the landslide dataset, a ratio of 1.2:1 positive to negative examples was adopted. This was seen as a compromise that emphasised positive instances, without doing so overwhelmingly. The sample size specified was 5% for all runs.

² In effect, the class sizes of a set of floating point numbers is determined by their precision.

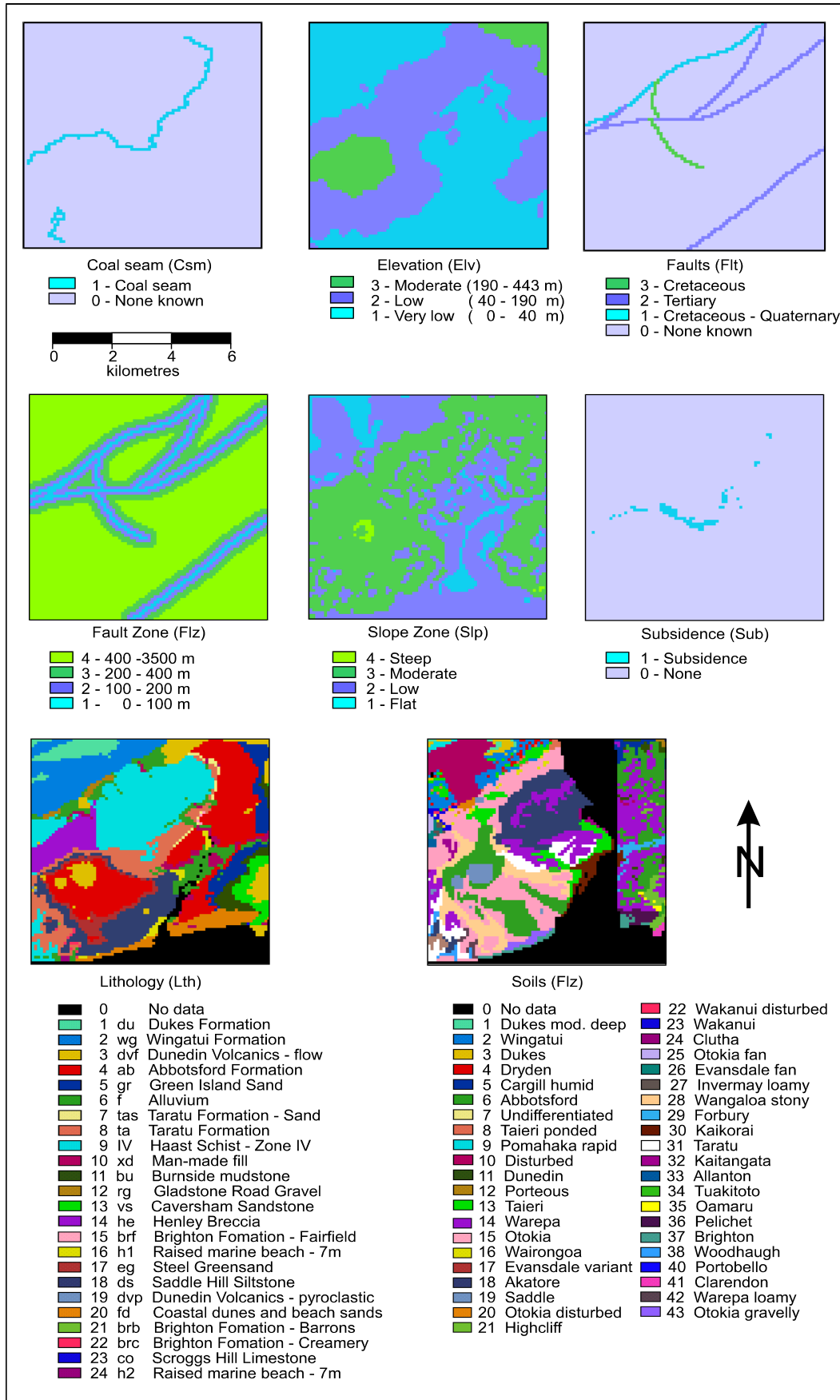


Figure 6 Maps of condition themes for the landslide hazard study area

7 KNOWLEDGE INDUCTION, REDUCTION AND VALIDATION

This phase of the case study used the RS-GKDD methodology algorithms described earlier.

7.1.1 Induction of Rule-Based Knowledge

The algorithm *Rough Set Based Rule Induction*, when applied to the *entire* landslides dataset comprising 6561 raster element “objects”, *all* eight condition attributes, *and* the one decision attribute, induced an initial 280 rules. *Rule Evaluation and Selection* was used to reduce these to 173 ‘nearest’ statistically significant rules. The latter rules achieved 40-fold cross validation classification accuracies of 80.7% on positive examples, 79.4% on negative examples, and 79.6% overall. An alternative view of these results is that false positive decisions were 19.3% and false negative decisions 20.6%. Of the ‘nearest’ statistically significant rules, the

Table 2 ‘Nearest’ statistically significant rules for all attributes of the landslides data – showing the forty positive and negative rules with largest supports

Condition Attributes								Decision	Absolute Support	Gen. Acc.%			
Coal Seam	Elevn	Fault	Fault Zone	Lithology	Slope Zone	Soil Type	Subsidence	LandSld	Absolute Support	Gen. Acc.%			
None	Mod		Distant	Abbotsford Formation		Abbotsford		LandSld	121	38.8			
	Mod			Abbotsford Formation		Otokia		LandSld	110	15.5			
	Low			Abbotsford Formation		Mod		Abbotsford	LandSld	101	46.5		
	Low			Abbotsford Formation		Mod		Warepa	LandSld	84	15.5		
	Low			Saddle Hill Siltstone		Mod		Abbotsford	LandSld	78	65.4		
	Low			Saddle Hill Siltstone		Mod		Otokia	LandSld	67	16.4		
	Mod			Dunedin Volcanics - flow		Mod		Saddle	LandSld	39	17.9		
	None			Low		Distant		Taratu Formation	Mod	Warepa	LandSld	36	22.2
	Vlow			Distant		Saddle Hill Siltstone		Mod	Otokia	LandSld	35	25.7	
	Low			Distant		Green Island Sand		Mod	Warepa	LandSld	34	41.2	
None	Low		Distant	Taratu Formation	Mod	Taratu	LandSld	34	14.7				
	Low			Steel Greensand	Mod	Otokia	LandSld	33	12.1				
	Low			Green Island Sand	Mod	Abbotsford	LandSld	27	33.3				
	Low			ModClose	Abbotsford Formation	Mod	Abbotsford	LandSld	27	14.8			
				Distant	Henley Breccia	Mod	Otokia	LandSld	26	19.2			
	Mod			Distant	Steel Greensand	Mod	Abbotsford	LandSld	23	69.6			
				ModClose	Henley Breccia	Mod	Abbotsford	LandSld	22	77.3			
				ModClose	Abbotsford Formation	Mod	Otokia	LandSld	18	16.7			
	Low			Distant	Abbotsford Formation	Low	Abbotsford	LandSld	16	18.8			
	Low			Saddle Hill Siltstone	Low	Abbotsford	LandSld	15	93.3				
Low	Steel Greensand	Abbotsford	LandSld	14	85.7								
None				NoData				NoSlide	641	100.0			
				Wingatui Formation				NoSlide	548	100.0			
				Alluvium				NoSlide	355	100.0			
				Abbotsford Formation				Low	NoSlide	210	99.5		
				Coastal dunes and beach sands				NoSlide	204	100.0			
				Haast Schist - Zone IV				NoSlide	185	100.0			
				Dunedin Volcanics - flow				NoSlide	169	100.0			
				Dukes Formation				NoSlide	166	100.0			
				Haast Schist - Zone IV				Otokia	NoSlide	150	100.0		
				Taratu Formation				NoSlide	114	99.1			
				Man-made fill				NoSlide	112	100.0			
				Green Island Sand				NoSlide	112	97.3			
				Abbotsford Formation				Low	Warepa	NoSlide	101	97.0	
				Abbotsford Formation				Low	Warepa	NoSlide	91	97.8	
				Haast Schist - Zone IV				Distant	Warepa	NoSlide	81	97.5	
				Saddle Hill Siltstone				Distant	Wangaloa stony	NoSlide	70	94.3	
				Green Island Sand				ModClose	NoSlide	59	100.0		
				Green Island Sand				ModClose	NoSlide	58	98.3		
Dunedin Volcanics - flow	Low	NoSlide	57	100.0									
Haast Schist - Zone IV	K - Q	NoSlide	57	98.2									

twenty positive and twenty negative rules having the largest supports for their respective decisions are given in Table 2.

Table 3 shows the results of a *sample*-based search for the landslide dataset optimal attributes. Each sample search used a 70% *F* statistic confidence level for one-way analysis of variance. Notice that for all samples, searching concluded each time after identifying a two-attribute subset that gave a maximum classification accuracy for statistically significant induced rules. This suggests that other condition attributes are either poorly related to the presence or absence of landslides, or are substantially correlated with other condition attributes.

For the sampling strategy adopted, the question arises as to whether a sample is sufficiently representative. This issue was addressed by repeating the sampling and optimal attribute search algorithms twenty times. The most frequently occurring sample optimal attribute subset was selected as the optimal one for the dataset. The model classification accuracy on positive examples was used as a tie-breaker where more than one subset occurred with the same frequency. The computational cost of this approach remained significant.

Table 3 Optimal attributes for twenty random data samples

Optimal Attributes	No. of Samples	20-Fold Cross Validated Classification Accuracy					
		All Instances		Positive Instances		Negative Instances	
		Mean %	S. D. %	Mean %	S. D. %	Mean %	S. D. %
{ <i>Elevation, Lithology</i> }	14	80.5	1.47	89.9	1.30	67.9	2.41
{ <i>Elevation, Soil</i> }	3	80.5	1.51	89.4	2.51	67.5	6.94
{ <i>Lithology, Slope Zone</i> }	2	81.5	2.47	90.9	0.21	69.4	7.35
{ <i>Slope zone, Soil</i> }	1	81.5	-	81.2	-	81.2	-

Since fourteen of the samples had the optimal attributes {*Elevation, Lithology*}, these were chosen as the attribute subset on which to develop a set of rules using the full dataset. It should be remembered however that the classification accuracies achieved on these samples are not representative of the full dataset. As discussed above, the rule discovery process was weighted towards examining positive decisions.

Table 4 Summary of 'nearest' statistically significant rule model for optimal attributes

	{ <i>Elevation, Lithology</i> }	
Number of rules	19	
Number of positive rules	4	
Number of negative rules	15	
	Mean %	S.D.
Number of attributes/rule	1.53	-
Support/rule	3.54	-
Mean generalisation acc.	80.1	-
<i>Classification accuracies on training data</i>		
Positive decisions	54.6	-
Negative decisions	82.7	-
All decisions	80.4	-
<i>40-fold cross validated classification accuracies</i>		
Positive decisions	55.6	2.00
Negative decisions	81.0	0.85
All decisions	78.9	0.72

The 'nearest' statistically significant rules obtained for these attributes are shown in Table 5. It is interesting to observe that, even though four different attribute subsets are found to be optimal over the 20 samples evaluated, the classification accuracies of their corresponding rule models are relatively consistent. This is perhaps a result of correlation between the lithology and soil attributes since at least one of these attributes appears in every sample's optimal subset.

7.2 Discussion and Conclusions

With 19 rules, the rule model characterised in Table 4 is only moderately comprehensible. On the other hand a number of rules of potential geological interest are apparent in Table 5, even if they have relatively low generalisation accuracies.

With a maximum generalisation accuracy of 30.9%, none of the four rules in Table 5 giving a landslide decision is a particularly

strong generalisation. The rules fall well short of being deterministic. However, they clearly identify the circumstances associated with substantial landslide hazard. Support for these positive rules amounts to 20.5% (1,340 hectares) of the study area.

Table 5 'Nearest' statistically significant rules for optimal attributes

Condition Attributes		Decision	Support, %	Gen. Acc., %
<i>Elevation</i>	<i>Lithology</i>	<i>Land-slide</i>		
Low	Abbotsford Formation	Landsld	8.3	15.7
Mod	Abbotsford Formation	Landsld	4.7	23.5
Low	Saddle Hill Siltstone	Landsld	4.2	30.9
Low	Henley Breccia	Landsld	3.3	24.9
Low	Haast Schist - Zone IV	NoSlide	13.2	97.1
	NoData	NoSlide	9.8	100.0
	Wingatui Formation	NoSlide	8.4	100.0
	Alluvium	NoSlide	5.4	100.0
	Dunedin Volcanics - flow	NoSlide	5.0	95.7
Vlow	Abbotsford Formation	NoSlide	4.6	98.0
	Coastal dunes and beach	NoSlide	3.1	100.0
Vlow	Haast Schist - Zone IV	NoSlide	2.9	99.5
	Dukes Formation	NoSlide	2.5	100.0
	Caversham Sandstone	NoSlide	1.9	96.1
Vlow	Taratu Formation	NoSlide	1.7	99.1
	Man-made fill	NoSlide	1.7	100.0
Vlow	Green Island Sand	NoSlide	1.7	97.3
Mod	Green Island Sand	NoSlide	0.9	100.0

The generalisation accuracies of the rules give a potentially useful indication of the level of landslide risk (i.e. hazard) in the case of positive decision rules. Conversely, the generalisation accuracies of negative rules suggest the extent to which a location can be described as safe from landslide.

Since the geographic landslide model "discovered" in this study resulted from a "generic" application of the RS-GKDD methodology, it is reasonable to assume that the same methodology could be applied equally effectively in other districts. In such cases a minimum requirement for thematic data would be *Elevation* and *Lithology*. There is also no reason, in principle, preventing the use of a more specialised input landslide classification. This would support a finer-grained understanding of the spatial distribution of landslide hazards.

With regard to possible extensions to this study it is noted that there is geological evidence that surface features away from immediate slip sites may have causal influences on landslide events. This suggests that a knowledge discovery method capable of inducing spatial relationships may enable a more effective model to be induced. Such an approach is part of the RS-GKDD methodology and an initial application to the dataset used in this study is reported by Aldridge (1998).

ACKNOWLEDGEMENTS

Thanks to Bruce McLennan for compiling the data sets and providing the location map. The Institute of Geological and Nuclear Sciences (IGNS), Dunedin, kindly provided most of the thematic data. The constructive feedback on the discovered models from IGNS geologist, Phil Glassey is much appreciated.

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.
- Aldridge, C.H., and Benwell, G.L., 1993, *Dunedin Pilot Hazards Information System: Logical Design*. An unpublished report of the Spatial Information Research Centre, University of Otago, Dunedin, N.Z.
- Aldridge, C.H., and Benwell, G.L., 1993, *Dunedin Pilot Hazards Information System: System Proposal*. An unpublished report of the Spatial Information Research Centre, University of Otago, Dunedin, N.Z.
- Aldridge, C.H., Benwell, G.L., Turnbull, I., Glassey, P., Henderson, J., Harris, M., and Tay, A.L., 1993, Dunedin Pilot Hazards Information System: a system analysis and proposal. In Benwell, G.L., and Sutherland, N.C., editors, *Proceedings: Fifth Annual Colloquium of the Spatial Information Research Centre, 17-19 May 1993, Dunedin, New Zealand* (Dunedin, N.Z.: University of Otago), 247-264.

- Anonymous, 1980, *Report of the Commission of Inquiry into the Abbotsford Landslip Disaster* (Wellington, N. Z.: N. Z. Government Printer).
- 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, 513-522.
- Cohen, P.R., and Feigenbaum, E.A., editors, 1982, *The Handbook of Artificial Intelligence - Vol. 3* (Reading, MA, U.S.A.: Addison-Wesley Publishing Co).
- Düntsch, I., and Gediga, G., 1998, Uncertainty measures of rough set prediction. *Artificial Intelligence*, no. 106, 109-137.
- Efron, B., 1982, *The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics* (Philadelphia, PA, U.S.A.: Society for Industrial and Applied Mathematics), no. 38.
- 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, 27-34.
- Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C., 1992, Knowledge discovery in databases: an overview. *AI Magazine*, 57-70.
- Gawrys, M., and Sienkiewicz, J., 1993, *Rough Set Library User's Manual (Version. 2.0, September 1993)* (Warsaw, Poland: Institute of Computer Science, Warsaw University of Technology).
- Glasse, P., Forsyth, P., Aldridge, C.H., Clements, Ryan, and Benwell, G.L., 1994, Dunedin Pilot Hazards Information System - Trial by GIS. In Benwell, G.L., and Sutherland, N.C., editors, *Proceedings: Sixth Annual Colloquium of the Spatial Information Research Centre, 17-19 May 1994, Dunedin, New Zealand* (Dunedin, N.Z.: University of Otago), 105-116.
- Gunter, B., 1997, Tree-based classification and regression - Part 2: Assessing classification performance. *Quality Progress*, Dec, 83-84.
- Hancox G.T., 1994, *Landside zonation in southwest Dunedin. (In progress 1994)*.
- 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.)* (Berlin: Springer), no. 951, 47-66.
- 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.
- Krantz, D.H., Luce, R.D., Suppes, P., and Tversky, A., 1971, *Foundations of Measurement - Volume 1: Additive and Polynomial Representations* (San Diego, CA, U.S.A.: Academic Press).
- Maddison, R.N., 1983, *Information System Methodologies*.
- McKellar, I.C., 1990, *Miscellaneous Map of New Zealand - Southwest Dunedin Urban Map 1:25 000 (1 sheet) and Notes (64 p.)* (Wellington, N. Z.: New Zealand Geological Survey).
- Mitchell, T.M., 1983, Learning and problem solving. In *IJCAI-83: Proceedings of the Eighth International Joint Conference On Artificial Intelligence, 8-12th August, 1983, Karlsruhe, West Germany*, vol. 2, 1139-1151.
- N.Z. Government, 1991, *Resource Management Act 1991* (Wellington, N.Z.: NZ Government Printer).
- Pawlak, Z., 1982, Rough sets. *International Journal of Computer and Information Sciences*, v. 11, no. 5, 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* (Dordrecht, The Netherlands: Kluwer Academic Publishers), no. 9.
- Shannon, C.E., 1949, Communication in the presence of noise. *Proceedings of the IRE*, v. 1949, no. Jan, 10-21.
- Stone, M., 1974, Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B: Applied Statistics*, v. 36, no. 2, 111-147.