

From Concepts To Data And Back Again: Connecting Mental Spaces With Data And Analysis Methods

Mark Gahegan, Xiping Dai, James Macgill and Sachin Oswal

GeoVISTA Center, Department of Geography,
Pennsylvania State University,
State College, PA 16802, USA
Telephone: +1-814-865 2612
FAX +1-814-863-7943
Email: mng1@psu.edu

Abstract

This paper describes our efforts to build a system that encompasses the full extents of the classification problem as it applies in geography. We use as our example the specific task of creating an informative land cover map, taking into account both the abstract conceptualization of the landscape and the emergent properties of the dataset used. The system employs a range of computational and visual components, from a concept or ontology browser, through a range of different classifiers to some exploratory visualization methods. We specifically recognise that successful classification involves the dynamic interaction of these kinds of tools within a cycle of scientific investigation. We present a use-case scenario to show these tools, how they might be used and how they interact.

1 Problem Setting

Within the geographic domain there has been much interest recently in the design, implementation and use of two distinct sets of tools and techniques, those that encode and depict conceptual structures, such as ontologies and concept maps (Smith and Mark, 1998; Rodriguez and Egenhofer, 2003), and those that support exploration and knowledge discovery activities, such as geovisualization and exploratory spatial data analysis (ESDA) (MacEachren et al., 1999; Dykes, 1996; Gahegan 2001). These two types of tools represent both ends of a continuum from mentally-held concepts¹ and their relationships (a top-down view of the world) through to the actual data to be analysed and its emergent properties (a bottom-up view of the world). However, these sets of tools are largely separated from each other, with no means of interaction, in fact they typically reside in quite distinct software products with poor integration between them. But activities at either end of this continuum of science activities should not be artificially isolated by the systems we use because they are intimately connected in a conceptual sense. For example, geography is both a descriptive and a discovery science; a person's understanding of concepts both helps to shape, and is in turn shaped by, interaction with data. Indeed it is well known that the creation of categories throughout the earth sciences is a compromise between the user's conceptualisation of a problem and the structure and

¹ Here we use the word 'concept' to indicate a mental notion of some set of like entities. An example might be the mental idea of forest. We use the word 'category' to describe an implementation of a concept, which can be thought of in two senses: the first is *intensionally*, for example as the normal distribution function produced by a Gaussian classifier, or the interval produced by a quartile classifier; the second is *extensionally*, as the set of examples or that are assigned to a specific a category. We use the word 'classifier' to describe a tool by which concepts and categories are mapped to each other.

distribution inherent in the data; and therefore a good classification scheme should both *impose* structure and *reveal* the structure already present within the data (Anderberg, 1973).

Visualization has the potential advantage to be an interface between human and machine-learning since it creates graphical images of data, helps humans to explore, reason and learn effectively, and usually enables an interactive visual exploration of the data. Blenkinsop *et al.*, (2000) employed exploratory visualization tools to help improve the expert’s understanding of uncertainty in a classified image scene. Lucieer and Kraak (2002) implemented visualization tools together with a supervised fuzzy classification algorithm to improve a geoscientist’s insights into uncertainty in remotely sensed image classification. Both of the above research emphasized the visualization tools with functionalities such as dynamically linked views and geographic brushing. An overview of possible connections between visualization and knowledge discovery is given by Gahegan *et al.*, 2001.

Considering the specific case of land cover classification, ontological tools that describe hierarchies of concepts (such as might be drawn from the Anderson land cover classification taxonomy) can offer sets of candidate categories from which a classifier might be trained (see Figure 1), or conversely, exploring the clustering of sample points in attribute space might lead one to hypothesize suitable mental concepts to represent these points (see Figure 2)—in other words to create or modify an ontology.

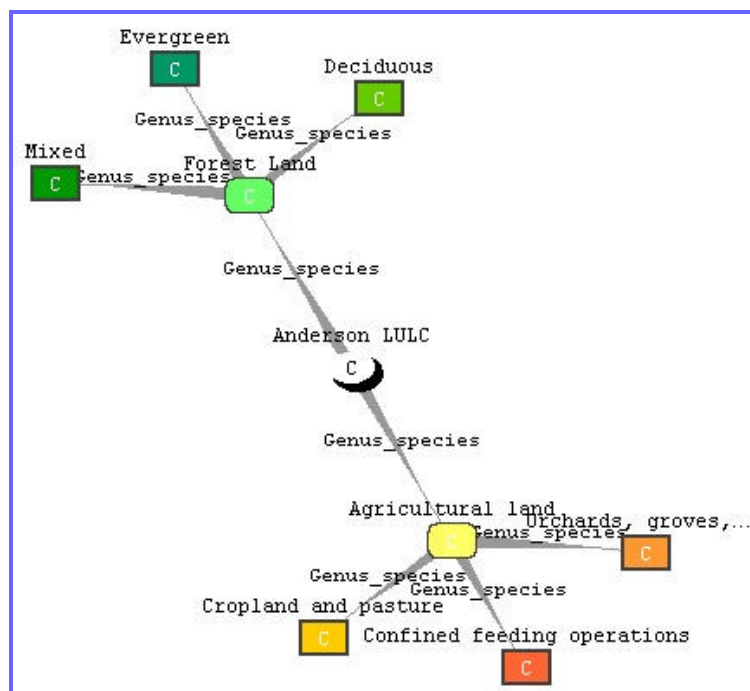


Figure 1. Part of the Anderson land use-land cover taxonomy, displayed in *ConceptVISTA*. This fragment shows the hierarchies pertaining to agriculture and forestry.

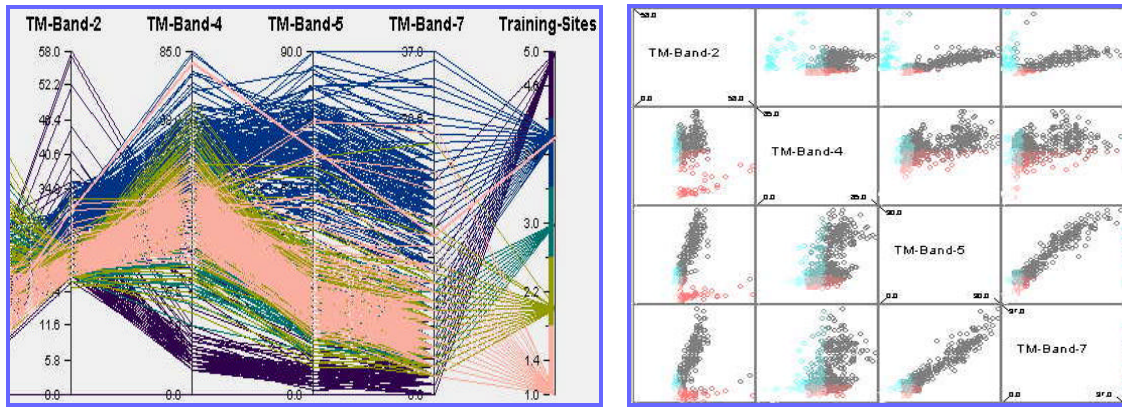


Figure 2. A parallel coordinate plot (left) and scatterplot matrix plot (right) detailing the spectral characteristics of different land cover types in a forested region.

2 The Process of Category Construction

It is common in cartography, geovisualization, GIS and remote sensing for users to construct categories that help them better discover, understand and communicate the structure of the situations they encounter. These categories may be act to summarise and simplify the data for presentation purposes (e.g. choropleth mapping) or may represent an attempt to impose labels that relate to specific mental concepts (e.g. soils classification). Constructing categories that are indeed a good compromise between mental conceptualization and the emergent properties of the dataset is a complex task that involves as much mental activity as it does computational analysis.

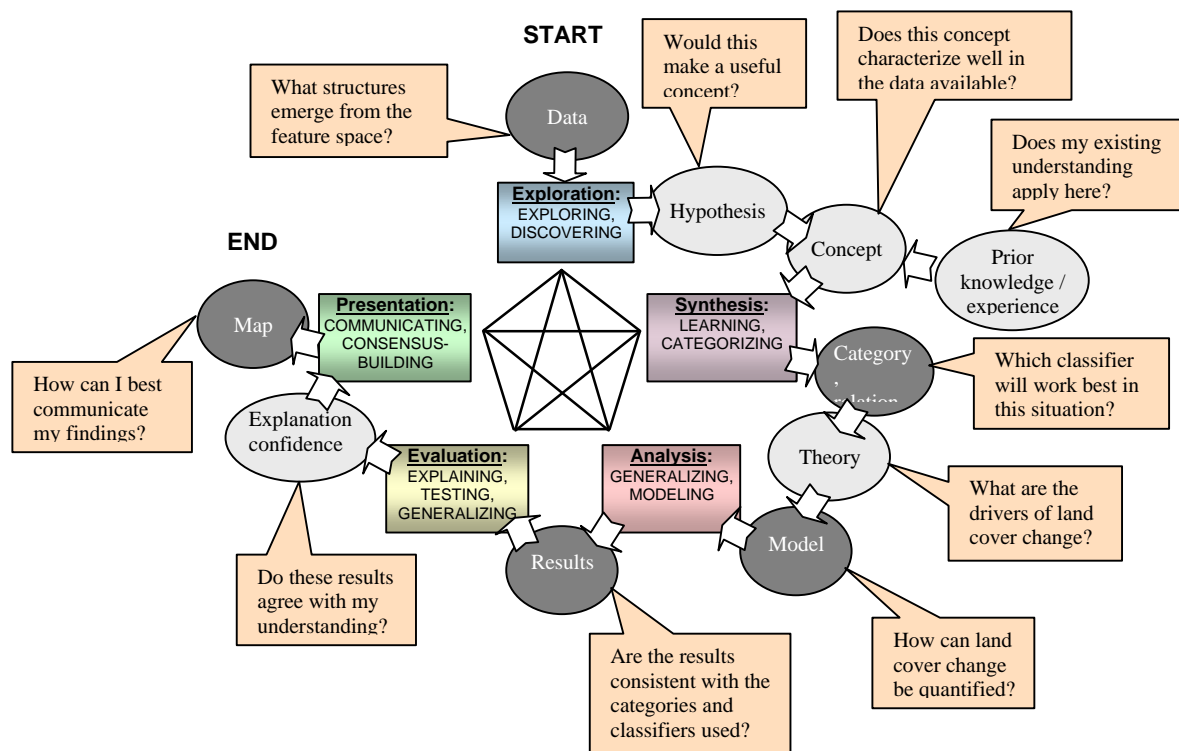


Figure 3. A conceptual map of the process of synthesising a suitable set of categories for a land use / land cover mapping exercise. See text for details.

An overview of the process is given in Figure 3, showing in the middle a cycle of the general stages of scientific investigation, (coloured square boxes) and a series of more specific products that form the interfaces between stages (grey ellipses: the light grey ellipses represent mental structures, the dark grey ones are products that can be held in a computer. The outer layer of captions show the kinds of questions that an analyst might ask to establish a suitable classification scheme for a given land use / land cover classification task. Many other analysis tasks in geography, cartography and throughout the natural sciences involve similar kinds of processes; for example identifying regions of urban gentrification, generating seismic anomaly maps and mapping the vulnerability of places to climate change.

3 Our Solution

We can conceive of the integration of classification and visualization techniques according to the depth of integration, so three such levels would be as follows. In the first level, visualization techniques are applied independently from clustering or classification methods. This is commonly the case at the moment in many commercial packages. For the second level, visualization techniques are implemented to display the workings of clustering and classification methods, such as visualizing the training of a neural network or k-means classifier with the resulting clusters or classes displayed in exploratory analysis tools. For the third level, visualization and classification techniques directly interact with each other; visualization tools not only display the input and intermediate steps, but also provide a dynamic interface for analysts to directly control and steer the processes.

To this last end we have developed a suite of tools, developed in the GeoVISTA *Studio* problem-solving environment (www.geovistastudio.psu.edu; Gahegan *et al.*, 2002) that facilitate and connect together the processes of (i) creating and browsing concepts in ontological and taxonomic browsers, (ii) selecting concepts to use in a specific analysis exercise, (iii) operationalising the concepts with classifiers (iv) exploring the data to help formulate concepts from emergent structures (v) dynamically modifying the concepts, the classifiers or the data used as a result of poor categories being produced (i.e. categories that do not align well with mental concepts or are not clearly differentiable in the data). Figure 4 shows the actual tools developed in *Studio* to facilitate these activities, with arrows being used to indicate schematically some of their interactions. We now describe each of the relevant tools briefly and give examples of their coordinated use to better understand aspects of land cover / land use classification and the construction of demographic indicators.

The concept browser / ontology editor (*ConceptVISTA*) is based on the *TouchGraph* graph visualization package (www.tuuchtraph.com), which we have augmented with: (i) support the DAML ontology description language, (ii) better visualization and browsing capabilities, (iii) version control and management and (iv) support for distributed access.

The classification component supports four distinct classifiers: the traditional Maximum Likelihood, the C4.5 decision tree (both examples of supervised classifiers), k-means clustering and the Self Organizing Map (SOM) (both examples of the unsupervised approach). So the maximum likelihood and decision tree require a learning or training phase prior to application, whereas the k-means and SOM require a post-application interpretation of the identified clusters.

The exploratory visualization components comprise a scatterplot matrix, parallel coordinate plot (examples of which appear later in the paper), image viewer, minimum spanning tree and Gabriel graph. These are linked together (Monmonier, 1992) for examining the multivariate

structure of the data, checking for outliers and redundant attributes, and identifying the emergent signatures of the various target categories.

To effectively connect these three tools together we utilize a sophisticated software coordination tool. Linking and brushing have been used for a number of years (Newton, 1978) to explore patterns in data displayed in different tools, here we extend this approach across different levels of abstraction. To date, linking and brushing has largely been limited to exploring how the same data items (records, objects) are manifest in different graphing and mapping displays; hence the cardinality of relationships between the displays are one-to-one—one dot in the scatterplot equates to one line in the parallel coordinate plot and so forth. An example of this kind of connection is shown at the bottom of Figure 4, where the darker blue regions have been selected in one of the scatterplots and the same selection then projected into the parallel coordinate plot, where they appear as heavier lines. Here we must also extend the linking across conceptual hierarchies to link the data to the classifiers and the ontology browser. To do so we need to support hierarchical and many-to-many relationships that connect data to ontologies, via interpretation and analysis activities (in this case, classification).

4 Use Case Scenario

In order to convey a sense of what these tools are like to use, we now present an annotated notebook that captures screenshots from an analysis exercise, with short descriptions of the analyst's thought processes leading to each decision.

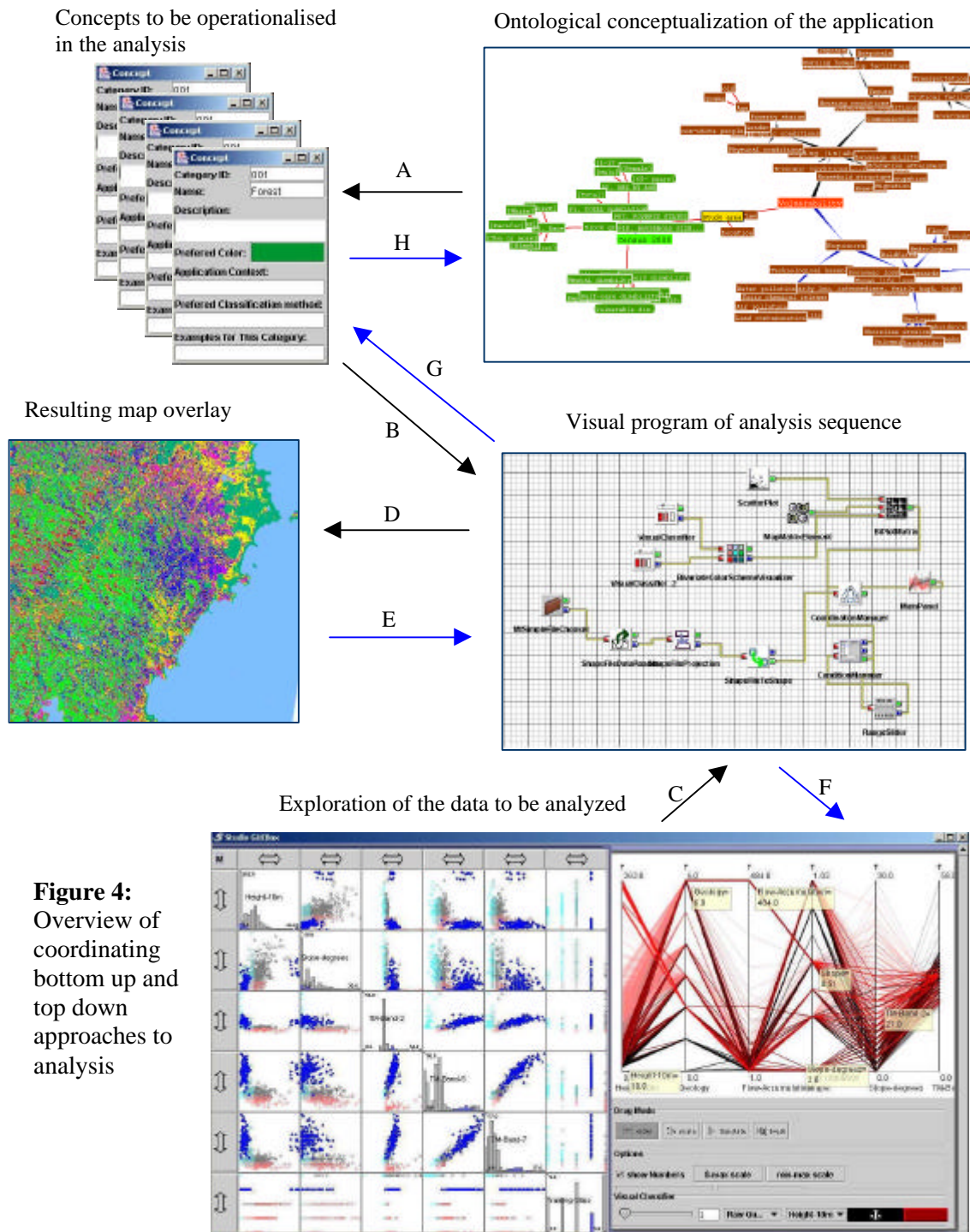
The exercise involves classifying the Kioloa dataset (Lees & Ritman, 1991), the latest version of which is available as a NASA Pathfinder reference site (<http://sres.anu.edu.au/associated/pathfinder/>). The dataset describes a coastal region in New South Wales, Australia. The aim of the classification is to differentiate the various floristic species occurring in a highly variegated landscape. The region contains a small amount of rainforest, intermixed with eucalypt tree types, cleared land and coastline.

The data fields gathered contain a mixture of Landsat TM and ancillary data as follows:

Aspect-N, Aspect-E, Flow-Accumulation, Geology, Height, Shape (surface morphology), Slope, TM-Band-2, TM-Band-4, TM-Band-5, TM-Band-7, Training-Sites.

The target categories (for which field-based training data are available) are initially identified as follows:

- Category 1: Dry Sclerophyll
- Category 2: E. Botryoides
- Category 3: Lower Slope Wet
- Category 4: Wet E. Maculata
- Category 5: Dry E. Maculata
- Category 6: Rainforest Ecotone
- Category 7: Rainforest
- Category 8: Paddocks / Cleared Land
- Category 9: Sea or Lake



KEY

- A. Concepts to be used in an analysis are extracted from the ontology and held in into an experimental notepad.
- B. A design for the experiment is constructed using the visual programming utility in Studio.
- C. The data are analysed for emergent structures and relationships that can be utilised, and for errors and unhelpful attributes that should possibly be removed.
- D. The experiment produces a result set of categories, held intensionally as pieces of a classifier model and extensionally as a map or dataset.
- E. Problems with the result can cause the experimental design to be changed
- F. Problems with the result might lead to a re-exploration of the data
- G. Problems with the result might cause the user to modify the concepts being utilised
- H: Modified concepts can be inserted into the ontology, leading to a modified ontology.

One possible taxonomic view of the categories, based loosely on a moisture gradient (the terrain is hilly with significant local differences in surface water accumulation) is shown below in Figure 5.

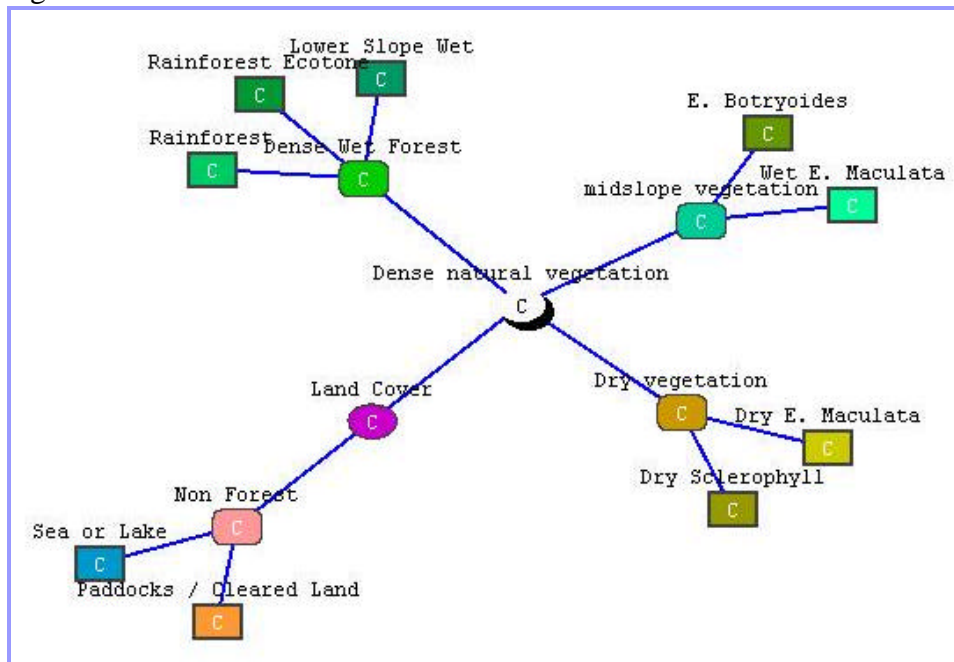


Figure 5. (See text for details.)

Figure 6 shows the results of applying a k-means classifier on the training dataset. Comparing the classification results with the training site, most classes are pretty much mixed with each other—they separate poorly as shown by a jumble of colours in both visualization tools.

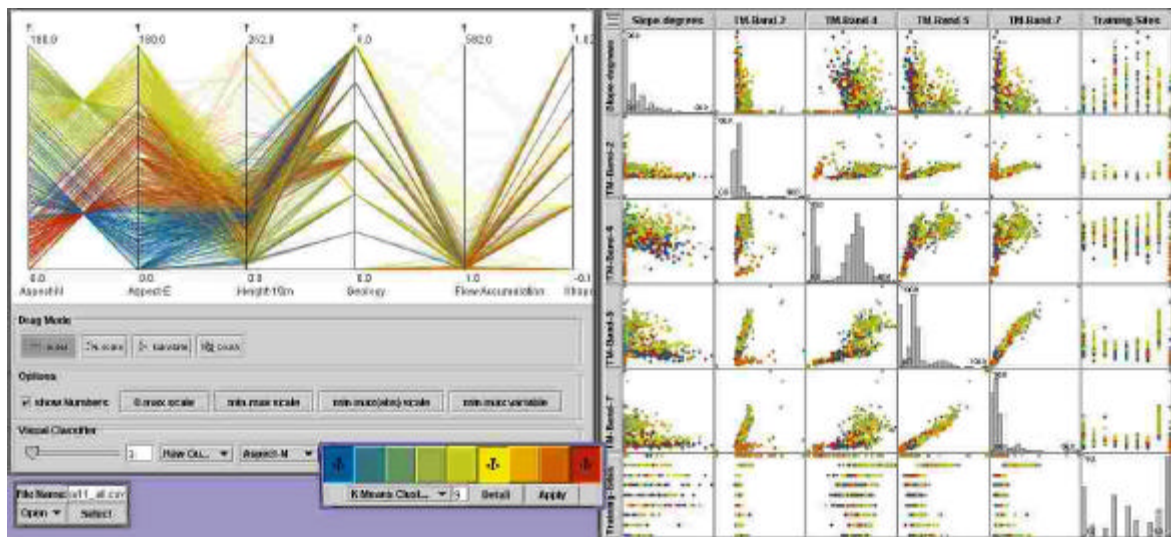


Figure 6. (See text for details.)

The following figure (Figure 7) explores the “Paddocks / cleared land” class in more detail. This class shows a huge spread of values in the parallel coordinate plot on the axes for Aspect, Geology and Shape, but seems to be bivariate (two distinct ranges of values) on the height axis. Characterization is better in the Landsat data shown in the scatterplot.

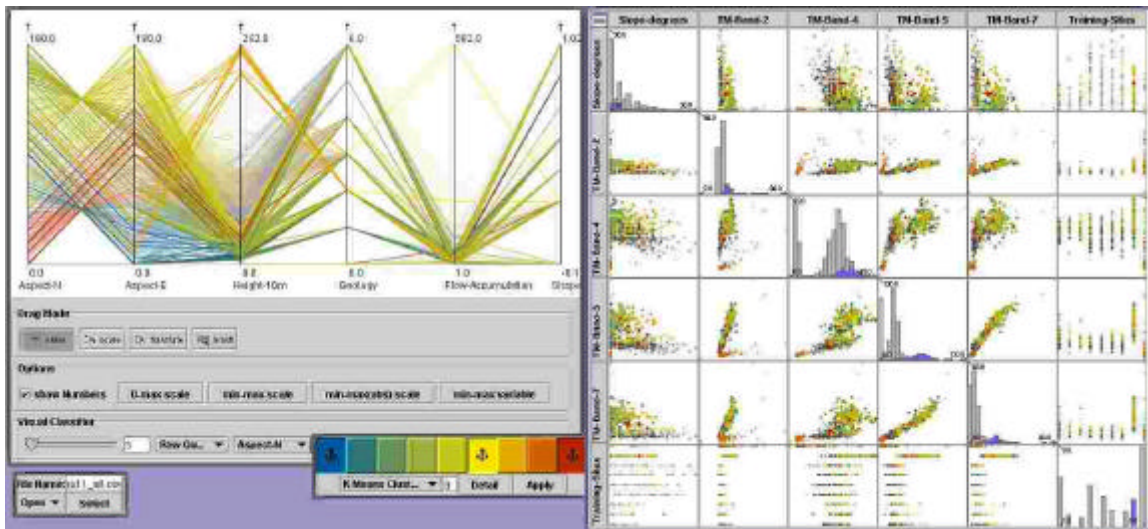


Figure 7. (See text for details.)

The next figure displays the distribution of observations and the resulting Gaussian distribution in each class in terms of every variable in training dataset (Figure 8). Using all of the 11 variables, we trained a maximum likelihood classifier and applied the classifier to the training dataset; the following figure shows the (disappointing) result (Figure 9).

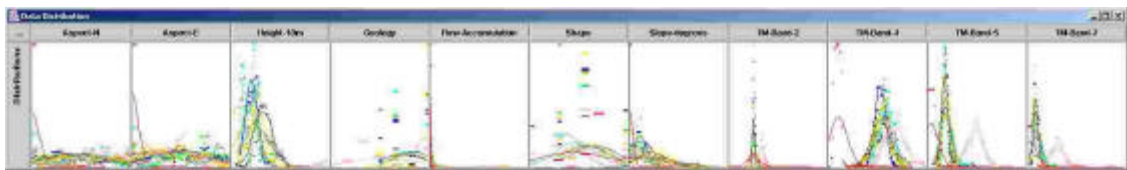


Figure 8. (See text for details.)

The maximum likelihood classifier cannot classify the dataset according to all of the 11 dataset. The failure might be caused by some variables not adhering to a Gaussian distribution, as maximum likelihood classifier assumes.

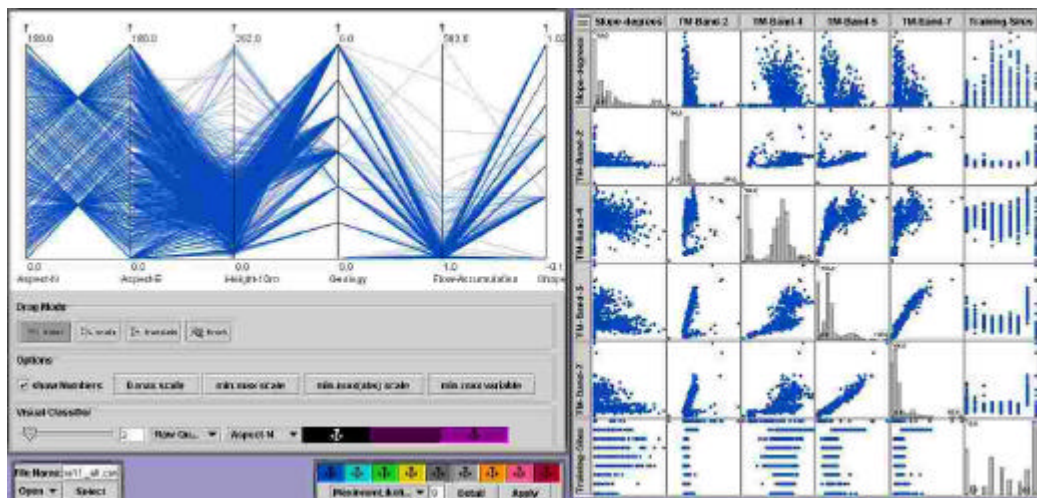


Figure 9. (See text for details.)

According to the distribution of observations, we dynamically dropped two variables, Aspect_N and Aspect_E, and used the rest of the variables to repeat the training process. This time, the maximum likelihood classifier succeeds in classifying the training dataset (Figure 10). Comparing the classification results, indicated by distinct colours, and the training sites, shown in the last column and last row in scatterplot matrix, many pixels have been classified correctly.

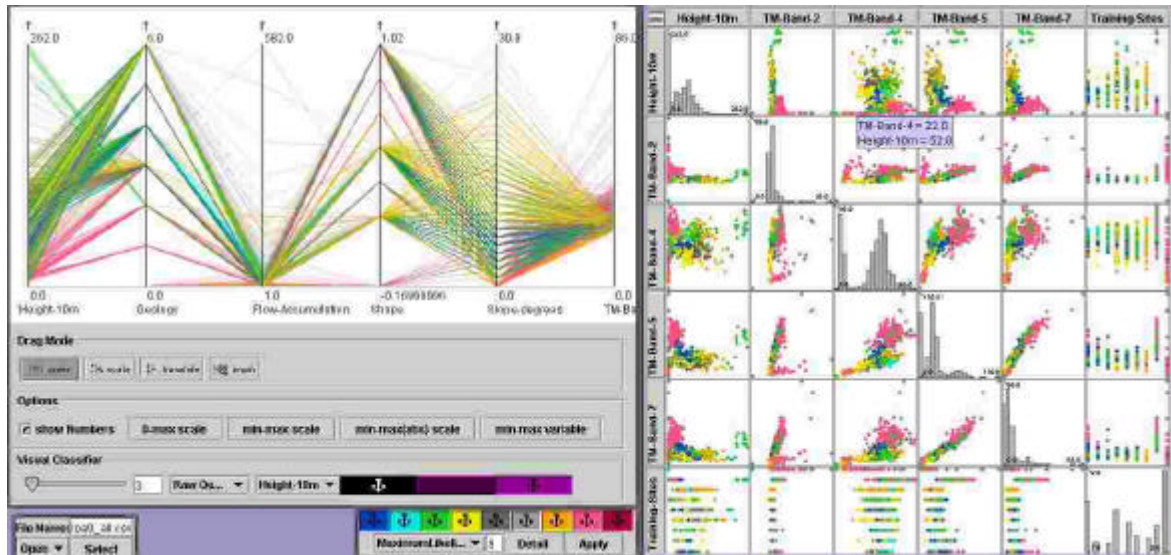


Figure 10. (See text for details.)

Selecting training sites with class “Paddocks / cleared land”, most of them are classified correctly, with colour pink (Figure 11). However, a part of the pixel in this class has been classified as class “Lower slope wet”. They are mostly outliers with different height from the rest of the pixels in this class.

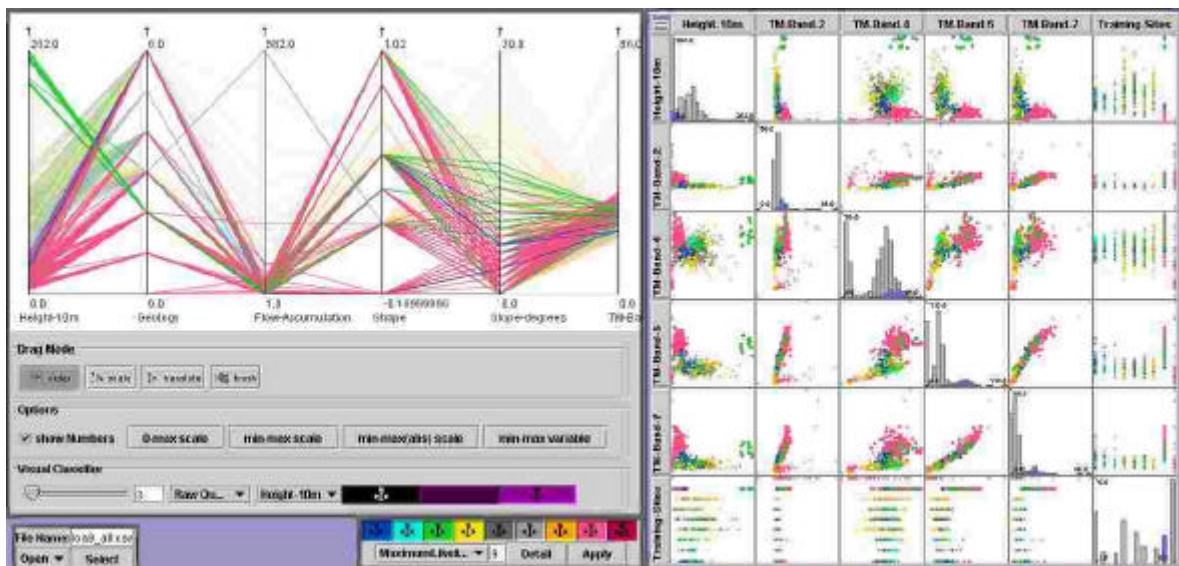


Figure 11. (See text for details.)

Having observed this difficulty in classifying “sea and lake”, we examine the data distributions produced by the classifier to see why (Figure 12).

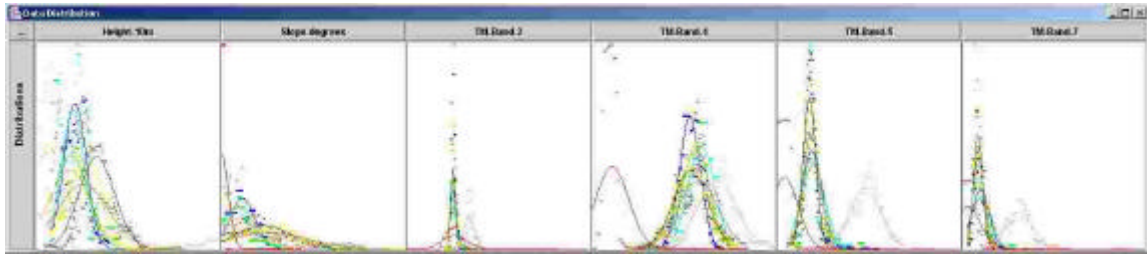


Figure 12. (See text for details.)

We notice that the distribution of class “sea and lake” (indicated as dots and lines in red) of variable height-10m (the left-most distribution in the above figure) is missing. That is caused by the fact that most of the pixels in this class have a height value of zero, and thus cannot fit the Gaussian distribution. Therefore, we drop the data for Height, Flow-accumulation, Geology and Shape, because the observations on these variables do not fall into a Gaussian distribution. The results of re-running the classification are displayed below in Figure 13. The class “sea and lake”, indicated in the red colour, and the class “Paddocks / cleared land”, shown in pink, are mostly classified correctly now.

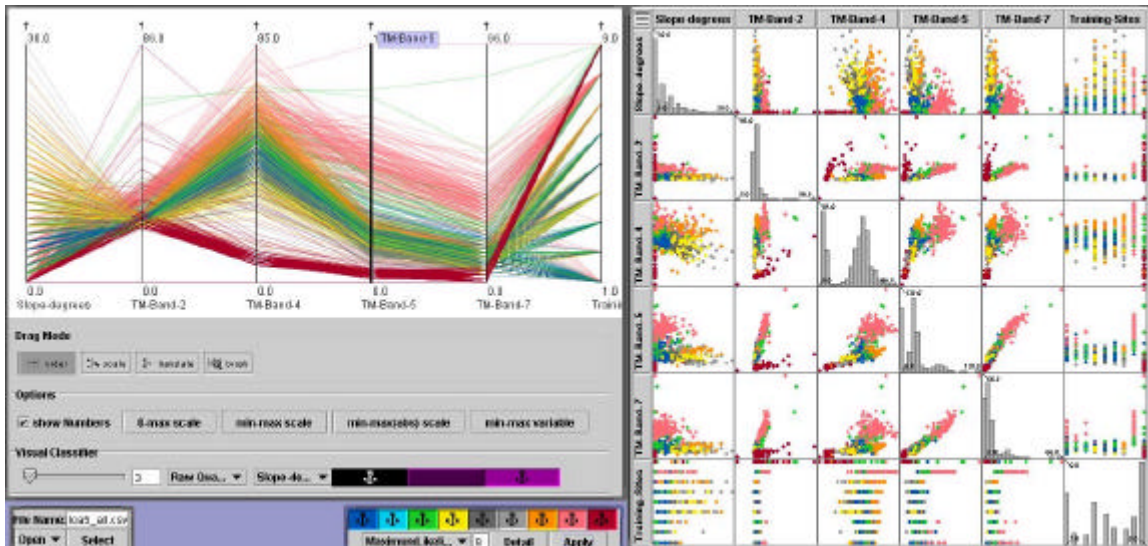


Figure 13. (See text for details.)

Some pixels in the class “Paddocks / cleared land”, however, have been misclassified as class “Lower slope wet”. There is also confusion between Dry Sclerophyll and the Lower Slope Wet and Paddocks / Cleared Land categories.

Rainforest Ecotone and Rainforest are selected in Figure 14. If all are correctly classified, they should be coloured in orange and green. However, the resulting of maximum likelihood classification contains a mixture of class Rainforest Ecotone, Rainforest, Dry Sclerophyll, Lower Slope Wet and Wet E. Maculata.

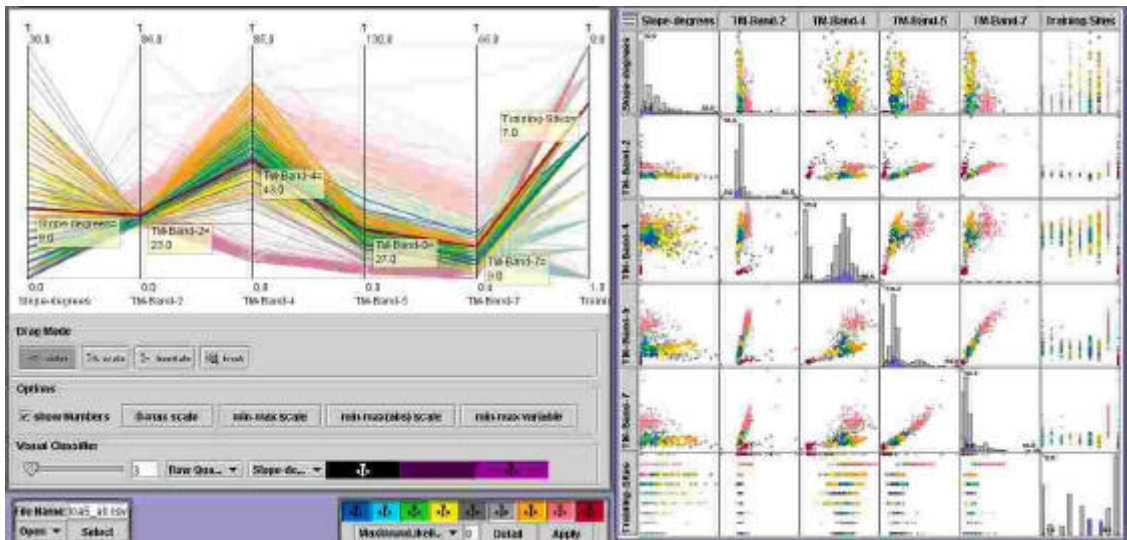


Figure 14. (See text for details.)

The 2 dimensional Gaussian distribution matrix (Figure 15) illustrates the problem in using maximum likelihood classifier to classify this dataset. The distribution of class Sea and Lake, Paddocks / Cleared Land are separable from the rest of the pixels. The other classes are, however, are still largely mixed with each other.

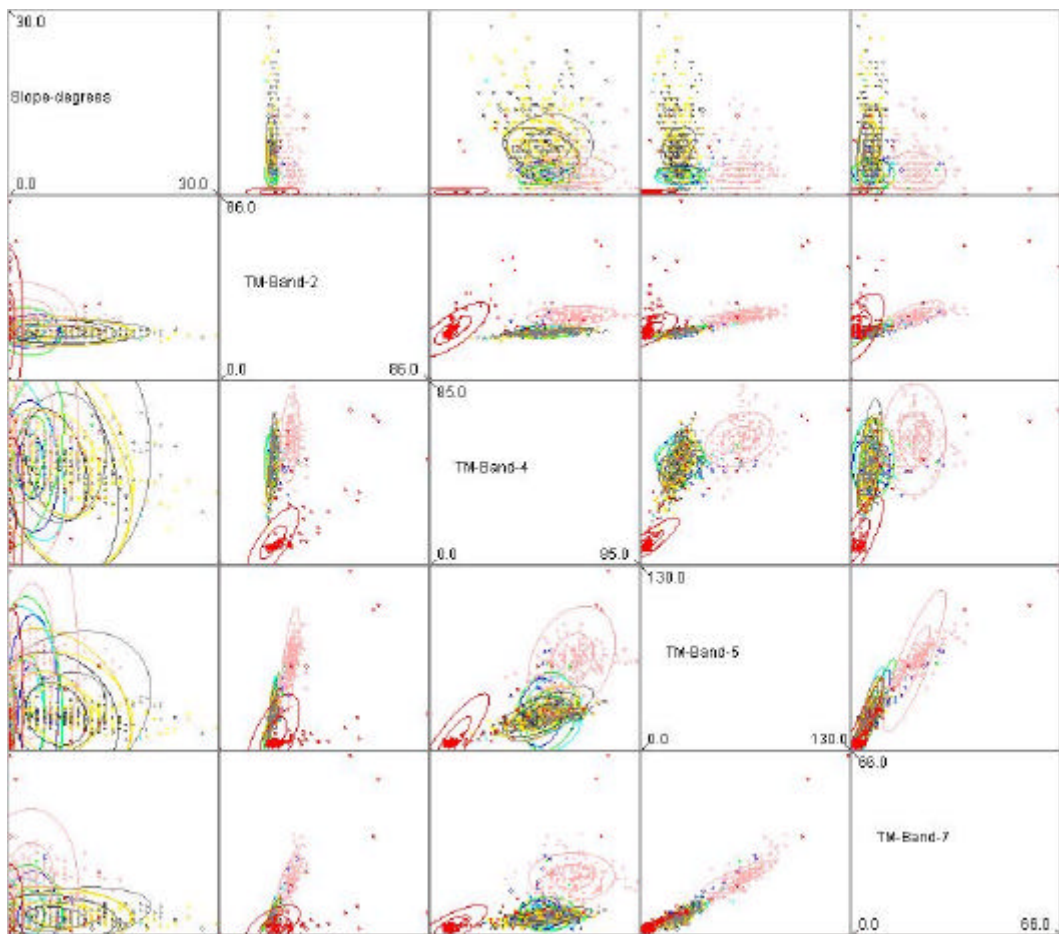


Figure 15. (See text for details.)

Figure 16 shows how a different classifier, based on the C4.5 decision tree would perform on the same dataset. Black lines are decision rules gleaned from the of decision tree training results.

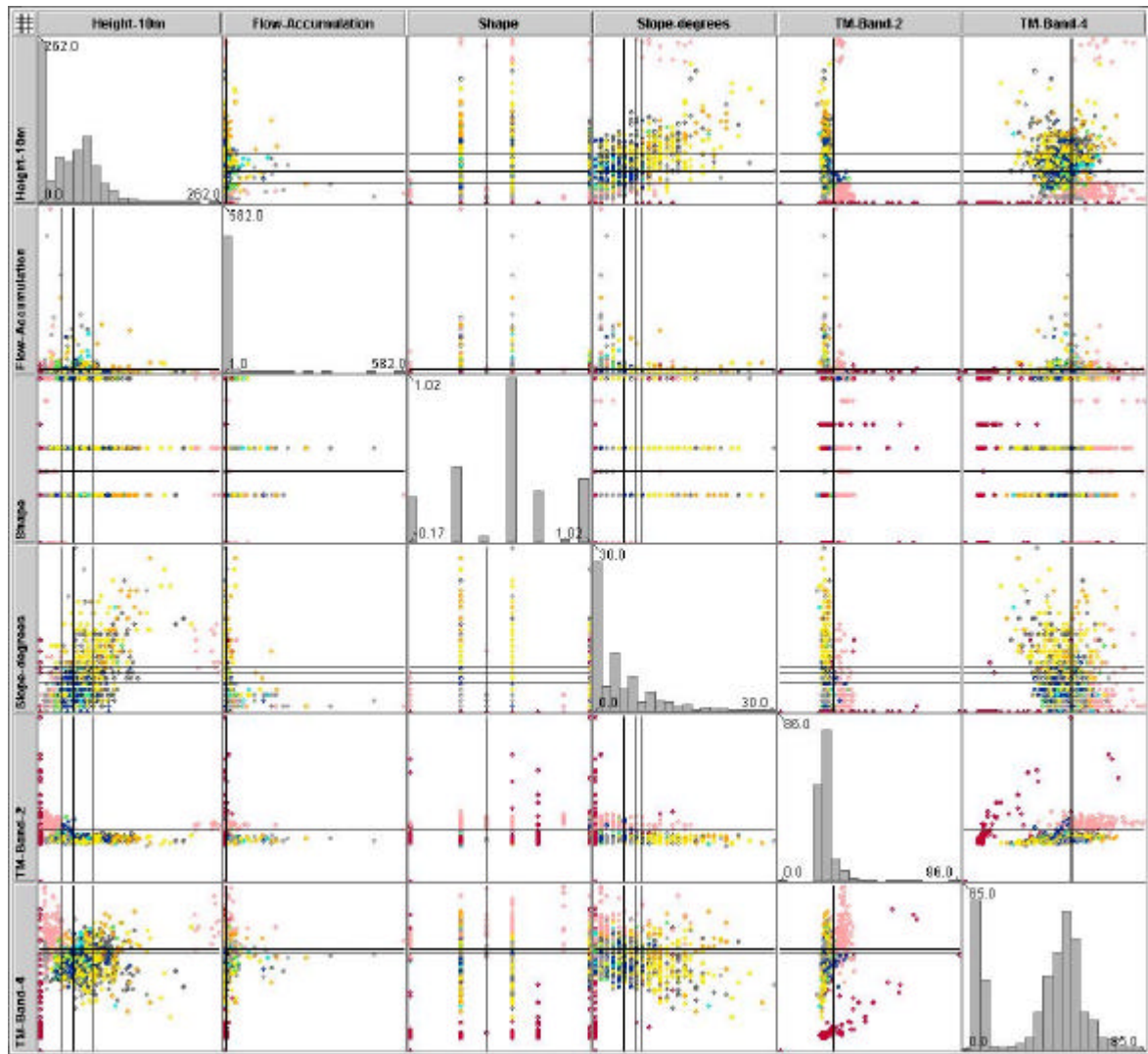


Figure 16. (See text for details.)

Bearing in mind the confusion between the forest classes we can achieve results with less errors by merging the overlapping categories together, in this case by generalizing them within the taxonomy shown earlier. One possible new classification system, using only 5 categories, shown below as a generalized version of the original taxonomy produces a clearer result when visualized in the parallel coordinate plot (Figure 17).

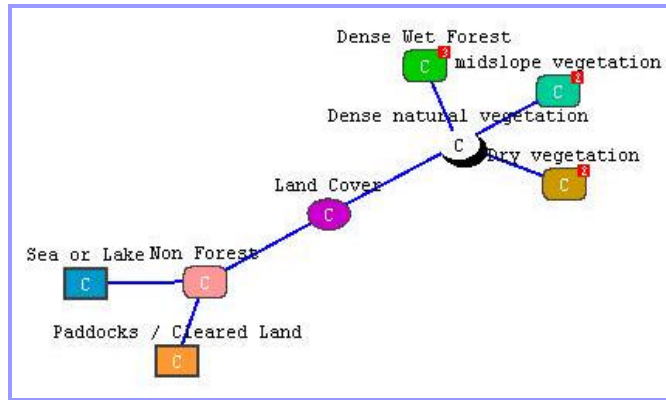


Figure 17. (See text for details.)

Figure 18 then shows how the resulting data spaces appear in the scatterplot matrix and parallel coordinate plot, as before. Although not perfect, the resulting generalized set of categories do show better separation

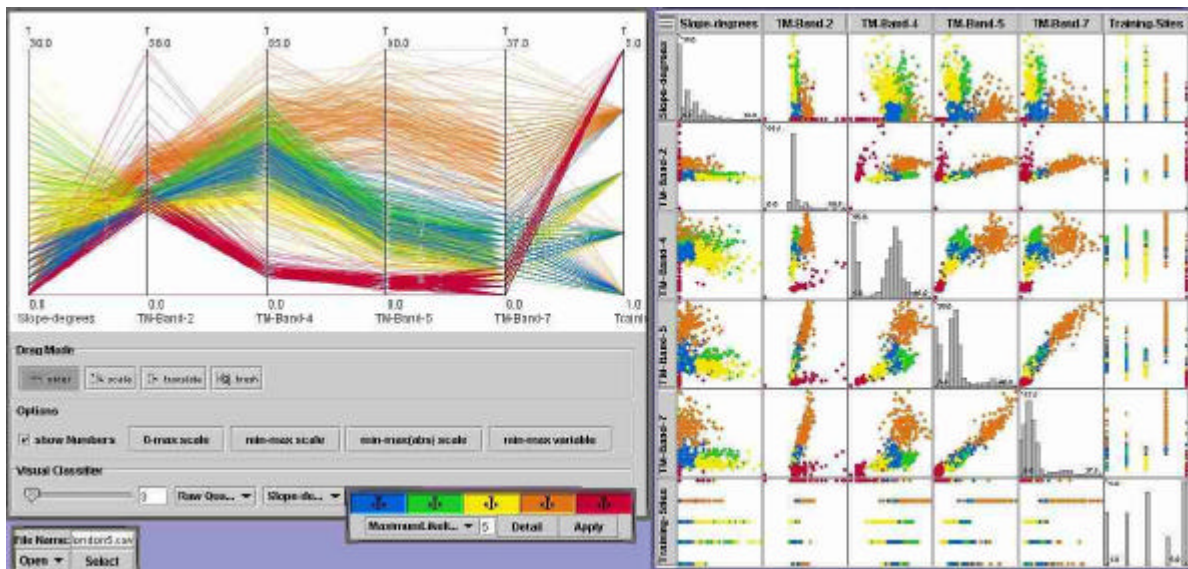


Figure 18. (See text for details.)

Finally, Figure 19 shows the result of classifying the modified (reduced) data space using the modified (generalized) taxonomy of Figure 17. The accuracy is improved, but the price of this is a more general picture of land cover than was originally planned.

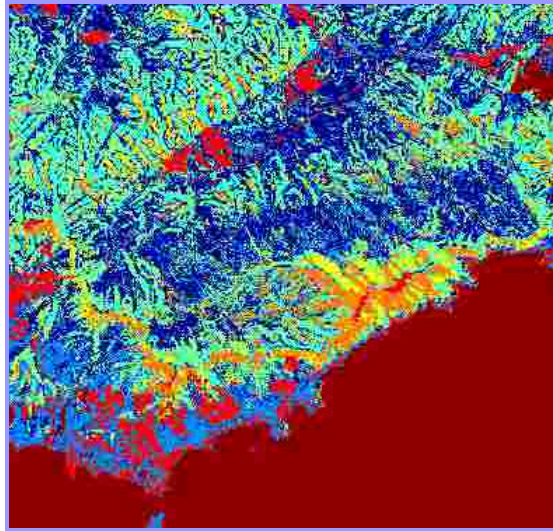


Figure 19. (See text for details.)

5 Summary and Future Work

We have described an approach to geographical analysis that attempts to house all aspects of the problem, from conceptualization to final analysis and verification all within the same computational environment. As a consequence of explicitly modelling the connections across the different computational and visual tools used (refer back to Figures 3 & 4), the analyst can ask questions that span the connections, such as:

- Which training data were used to form this category (between the data and the classifiers)?
- Is this sample typical of its category, or an outlier?
- What effect will the imposition of this classifier have on the data space, and in which regions will it perform well (badly)?
- What instantiations of this concept are available (which classifiers have been used to operationalise this concept)?
- If a classifier is struggling to separate two categories, can these categories be merged or generalized into one (i.e. by moving up the concept hierarchy? (e.g. merging deciduous, mixed and coniferous forests into a single forest category in Figure 1).
- Likewise, if a category appears to contain two or more specific sub-categories (e.g. is bimodal) can it be specialized to better model the data?

In future work we will concentrate on making the interactions between tools stronger, using cognitive principles and theories about the science process as our guide. For readers interested in using these tools for themselves, the entire *GeoVISTA Studio* codebase is downloadable from SourceForge: <http://sourceforge.net/projects/geovistastudio/>.

Acknowledgments

This research was supported in part by NSF grants 9978052, 0219025, 9983451, 9983459, 9983461 and 0225673. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- Anderberg, M. R. 1973, *Cluster Analysis for Applications*, Boston, USA, Academic Press.
- Blenkinsop, S., Fisher, P., Bastin, L. and Wood, J., 2000, Evaluating the perception of uncertainty in alternative visualization strategies. *Cartographica*, **37**(1), 1-13.
- Dykes, J. A. 1996, Dynamic Maps for Spatial Science: A Unified Approach to Cartographic Visualization. In *Innovations in Geographical Information Systems 3* (Ed, Parker, D.) Taylor & Francis, London, pp. 171-181.
- Gahegan, M.. 2001, Exploratory Geographic Visualization: Analysis with Light. In: *Knowledge Discovery with Geographic Information*, (Eds. Miller, H. and Han, J.) Taylor & Francis: London.
- Gahegan, M., Wachowicz, M., Harrower, M. and Rhyne, T., 2001, The integration of geographic visualization with knowledge discovery in databases and geocomputation. *Cartography and Geographic Information Science*, **28** (1), 29-44.
- Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F., 2002, GeoVISTA *Studio*: a geocomputational workbench. *Computers, Environment and Urban Systems*, **26**, 267-292.
- Lees, B. G. and Ritman, K., 1991, Decision tree and rule induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. *Environmental Management*, **15**, 823-831.
- Lucieer, A. and Kraak, M., 2002, Interactive visualization of a fuzzy classification of remotely sensed imagery using dynamically linked views to explore uncertainty. *Proceeding of Accuracy 2002 Symposium*, July 10-12, Melbourne, Australia, 348-356.
- MacEachren, A. M., Wachowicz, M. Edsall, R., Haug, D. and Masters R. 1999, Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science* **13** (4), 311-334.
- Monmonier, M., 1992, Authoring graphic scripts: experiences and principles. *Cartography & Geographic Information Systems* **19** (4), 247-260.
- Newton C. M., 1978, Graphics: From Alpha to Omega in data analysis, *Graphical representation of multivariate data*, ed. P.C.C. Wang, New York: Academic Press, pp. 59-92.
- Rodriguez, M. A. and Egenhofer. M. J., 2003, Determining Semantic Similarity among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, **15** (2): 442-456,
- Smith, B., and Mark, D. M., 1998, Ontology and geographic kinds. In Poiker, T.K., and Chrisman, N., (eds.), *Proceedings, 8th International Symposium on Spatial Data Handling*, pp. 308-320.

Biography

Mark Gahegan is often mistaken for a graduate student because he is scruffy-looking. To his constant amazement he is professor of geography at the Pennsylvania State University, where he holds the faculty medal for the longest singly-linked paper-clip chain. He has worked as an academic on three continents already (Europe, Australasia, North America), so we can extrapolate that he will run out of continents before reaching retirement age (assuming nobody

is silly enough to create a university in Antarctica). His degrees are in computing and GIScience and his research interests include: geovisualisation, computational semantics, spatial analysis, remote sensing and machine learning applied to geography. He is also an acknowledged expert of tetrapylotomy² and has one of the largest collection of garden gnomes in his entire neighbourhood. The others are also unsavoury characters, but at least they do not look so scruffy.

² The art of slicing a hair four ways (from *Foucault's Pendulum*, by Umberto Eco).