

An Exploration into the Definition, Operationalization and Evaluation of Geographical Categories

Mark Gahegan; Masa Takatsuka; Xiping Dai

*GeoVISTA Center, Department of Geography, The Pennsylvania State University, University Park, PA 16802,
USA.e-mail: mark@geog.psu.edu*

Abstract. Categories are essential to human reasoning, yet the tools we have at our disposal to *discover, propose, apply* and *evaluate* categories are still primitive. Current GIS often assume, *a-priori*, that the required categories for a specific exercise have been constructed by some external process. However, within the field sciences, it is often the case that categories must be proposed and refined *in-situ*, and through use. A further complication is the need to compromise between the needs of the analyst—who wants to be able to study some phenomena using existing categories defined by existing taxonomic knowledge, and the capabilities of the data gathered and algorithms used—that must be able to differentiate the required categories reliably. In this paper, we describe our latest progress towards developing an environment that supports the development and evaluation of categories. We show how a 'mixed initiative system', driven both by domain expertise (theory) and by data-oriented analysis tools (visualization and machine learning), can support the creation of categories that are both robust and useful. Theory is injected via a visual interface to the human expert, category separation is explored using visual and machine-based inductive analysis.

1. INTRODUCTION

Categories are at the very heart of how we conceptualise and reason. Many of the concepts we utilise in geographic analysis (such as susceptibility to flooding) we operationalise as a series of categories (high risk, low risk, no risk). Classification therefore plays a major role because it is the mechanism by which we apply these categories, by taking a complex dataset and extracting from it meaningful information in a summarised form. As such, a classifier can be defined as a method for mapping between the domain of numbers to the domain of concepts (represented by category labels), thus transforming data from one conceptual model to another (Gahegan, 1996).

Category representation and classification have both received a great deal of attention from many research communities, including machine learning and remote sensing as well as geography. However, classification and categories do tend to be studied in isolation; the mechanisms by which categories might be applied are often divorced from how they are conceptualised. Interest in classification has stemmed from the real need to reduce large data volumes and complexities into a manageable form that is expressible in the language of the analyst (e.g. Jenks, 1977; Mason *et al.*, 1988; Foody *et al.*, 1995; Cromley, 1996; Mitchell, 1997). Interest in categories, on the other hand, has focused on the mental processes by which they are conceptualised, learned and shared (Rosch, 1973, 1978; Smith and Medin, 1981; Lloyd *et al.*, 1996; Sowa, 1999). This paper forms part of a larger project whose aims are to capture some of the reasoning that underpins the development, use and communication of geographic categories that are, in part at least, defined by human experts. In doing so, we will

connect categorical understanding and theory applied by experts to the machine learning and statistical tools developed for classification, mixing the best of what humans and machines can offer.

1.1 Where do categories come from?

Categories are fundamental to human thought, so it is not surprising that they are also crucial to the structuring of information within Geographic Information Systems (GIS). The classical approach to mental categories dates back to the time of Aristotle, and assumes that categories exist in the world and that we can *discover* them (Sutcliffe, 1993). These natural categories are taken to be clearly defined and mutually exclusive. Realistically, this is the model we currently use in GIS, albeit implicitly. Categories are assumed to have a shared, consistent meaning among all users, that is somehow intrinsic to their nature, and so not represented formally (Nyerges, 1991; Usery, 1993). There is evidently some merit with such an approach; if it were not so, GIS would fail. However, a large body of theoretical work has demonstrated that this assumption contains many fallacies and, in its place, more advanced conceptual frameworks have been proposed. Typically these newer frameworks address the complexities of mental categorization as a *process* through which we *impose* meaning on the world (e.g., Barsalou, 1991; Lakoff, 1987; Lloyd *et al.*, 1996; MacEachren, 1995; Rosch, 1975). Such human-centred frameworks account for, in theory at least, differences among individuals, the use of "situated" knowledge, and the notion of categorical uncertainty (e.g. Clancey, 1997; Schuurman, 1999).

Currently, the primary role of the category within a GIS is to define collections of objects (instances), usually in

terms of the common properties that they share. The category is thus a template for constructing an instance (object) of a given type. Fundamental to the exchange of meaning among humans and machines is the reconciliation of the categories used. Indeed, categories play a major role in the ontologies used to formally describe and interoperate database and information system schemata (Guarino, 1997;1998; Fonseca *et al.*, 2000).

However, these categories are little more than empty shells; they convey no real substance, except in terms of syntax. Instead they rely heavily on the user to supply their meaning, their relevance and their reliability. In some cases perhaps this is acceptable, users may share a common understanding of a *post code* or a *lake*, for example, at least in terms of their general properties. In other circumstances, a category might represent a concept that is changing, is uncertain or is open to many possible interpretations. It is difficult to share this kind of knowledge, so when data is moved away from its originator(s), such details may be lost, increasing the chances for misinterpretation.

The problem of constructing appropriate categories seems to be one that is assumed to happen prior to data entry. GIS are simply not equipped to 1) uncover categories, 2) evaluate their utility, 3) represent their meaning, or 4) allow them to evolve. But all four of these aspects are fundamental to categories as they apply in the geographic sciences, and the way in which we reason about them (e.g. Smith & Samuelson, 1997). For example, the concepts of 'poverty' or 'gentrification' might prove useful in understanding an urban region, but they are difficult ideas to define and communicate, and even their very definitions do not seem to be fixed in time. Furthermore, in some cases it is unrealistic to expect categories to be known at the outset, especially when dealing with complex or unfamiliar datasets, they must instead be uncovered by exploration. Finally, to assume that categories can be defined statically seems to be too strong of an assumption. Categories may change as a direct consequence of use, indeed it might be very valuable to refine the description of a category based on feedback from attempts to apply it analytically.

1.2 Categories as Compromise

To make the matters yet more complex, the categories employed to analyse geospatial information are often a compromise between the analyst's needs—to provide a detailed and reliable understanding, and the supporting evidence in the data—which may be inconclusive, ambiguous and contain errors or uncertainties. In remote sensing there is a well-developed sense of these two forms of categories when applied to image classification (e.g. Jensen, 1999). Information classes are defined as the categories by which a human might understand a landscape, whereas spectral classes refer to the segmentation of a feature space into differentiable regions. The former approach is top-down, and might be

described as ontological, with a structure derived from concepts known to the expert. The latter is bottom-up and entirely data driven, with structure derived from the data values recorded, with no higher-level knowledge applied (Wisniewski and Medin, 1994).

1.3 Supporting the Development of Categories

Human expertise is typically used to arbitrate a compromise between these two approaches. It is obviously problematic to define concepts only from the data that make no sense to the analyst. Conversely, there is no point whatsoever in trying to operationalise any high level concepts that are not supported by the data; the classifier would simply fail to recognise them.

However, this division of effort between the human and the computer has become problematic. Software and methodologies have developed that focus interpretation and knowledge construction activities into distinct and separate processes with little co-ordination or feedback between them. These activities are confined to perhaps a less structured set of exploratory tools, which are in turn divorced from the systems that will ultimately apply the knowledge gained (categories in this case). To continue the example of image classification, current visualization and statistical methods that might support the development of useful categories are not integrated with the software that applies the categories (e.g. a GIS) or evaluates their utility. Such a situation stifles creativity and ignores the considerable experience that a human might bring, because it cannot be gathered and represented conveniently in a machine-accessible form.

In contrast, we regard *knowledge discovery* or *knowledge construction* within the geospatial sphere as a *developmental process*, with meaning being progressively constructed and refined through a series of pre-processing and interpretative steps (e.g. Ankerherst *et al.*, 1999; MacEachren, *et al.*, 1999; Valdez-Perez, 1999; Wachowicz, *in press*). Current systems lack adequate tools for supporting this process.

1.4 Our Vision

What we envisage instead is a single environment where a user can move seamlessly between deriving categories (knowledge construction), applying them operationally (analysis) and assessing their performance (evaluation). This has led us to construct a software environment, GeoVISTA *Studio*, which aims to encompass this entire spectrum of activities in an integrated manner. For the early stages of analysis, *Studio* makes use of inductive learning (Mitchell, 1997; Gahegan, 2000) and geovisualization (Kraak and MacEachren, 1999); methods that are inherently well suited to dealing with uncertainty and ambiguity. Together, these methods provide flexible mechanisms for category construction that are a better match to human thought processes. They furthermore allow users to explore the meaning and development of a category, by examining the data, methods and actions by which it was formed, as the later

sections of this paper detail. This has advantages for communication and sharing of understanding, since the categories can be represented and explored visually as well as statistically.

1.5 Some further goals

A further advantage of this integration is that it becomes possible to study aspects of the development of knowledge (e.g. Hao et al., 1999). For instance, we can pose questions such as: “What combinations of visual and computational methods are most effective for category construction within a specific application or dataset.?” Or: “How can evaluation of the use of categories in analysis lead to the refinement of their description and hence to increasing accuracy?” Of particular interest here are issues of data quality and uncertainty that accompany category construction. Our goal here is to provide visual tools to help the expert user to ascertain (and communicate) the viability of the information classes that proposed categories represent. For example, a category might be constructed from data that are noisy or contain errors, or from few examples, so may not be robust.

The approach described below is highly interactive, allowing direct manipulation of data and parameters with immediate feedback, engaging human abilities to rapidly process and rank visual information (Buja *et al.*, 1996). This interactive analysis also provides a richer experience for the analyst, via a more complete understanding of the relationship of categories to localised features and multivariate relationships in the data.

Finally, categories often overlap, but the overlap is not conveyed visually to the analyst and is often only available *a-posteriori* via confusion measures. But the visualization methods described here enable overlapping classes to be explored both in feature space, via a parallel coordinate plot, and within the decision space of the classifier, via a rendering of the hidden layer in a Kohonen Self Organising Map (see later).

2. GEOVISTA STUDIO

A technical description of *Studio* has been previously reported (Takatsuka & Gahegan, 2001) so will not be repeated here. In short, it is a visual programming environment, that allows users to quickly design, test and refine strategies to explore and analyse geospatial data. Functionality is encapsulated in JavaBeans that support a range of activities, from visualising high dimensional feature spaces, applying neural networks and traditional statistical analysis tools through to mapping outcomes. The specific tools used in this paper are the parallel coordinate plot (PCP) (Inselberg, 1985; 1997), the 2D (geographic) map and the Kohonen self organising map (SOM) (Kohonen, 1995), a type of neural network. The use of this particular PCP for exploration of highly multivariate data has been previously described (Gahegan, 2000) and the SOM and

related visualization tools to examine the break-up of feature space are discussed in a companion paper to this one (Takatsuka, 2001).

Figure 1 shows a typical design for the complex analysis described below, that co-ordinates a number of analytical and visual tools to explore geographical data. The real complexity in this design is introduced by the assignment of data properties to the visual variables supported by the renderer (shown down the centre of the design layout in the figure). Unlike typical GIS displays, our renderer can transform any data value into any of the visual variables that the selected geometry can support, in a similar manner to scientific visualization environments such as AVS and IBM's Data Explorer (DX).

All of the images included below are direct screen shots captured during exploratory analysis sessions using *Studio*.

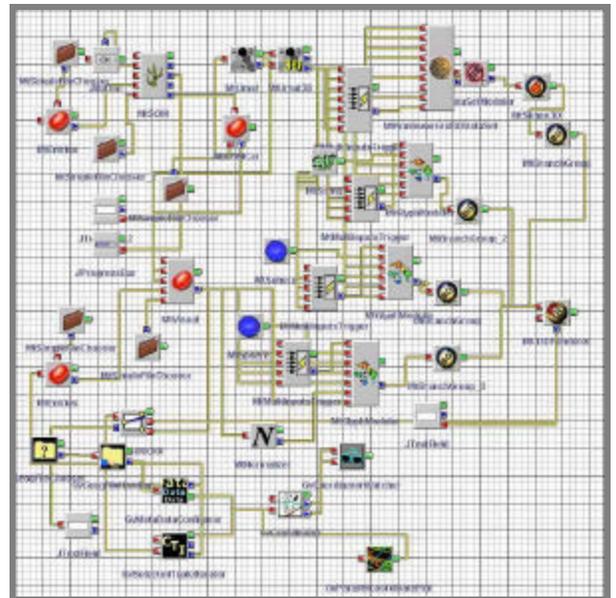


Figure 1: The *Studio* design canvas. For complex analysis tasks such as the one outlined here, it can take on the appearance of a printed circuit board! It is still a lot easier that programming though ;-)

3. EXPERIMENTS

Two examples are described below. In the first, the goal is to help the expert assess the viability of proposed categories by exploring their multi-dimensional nature and then examining the capabilities of a classifier (the SOM in this case) to construct the necessary decision boundaries that separate out the categories in the data. Direct interaction between the SOM and PCP helps to enable this task; the user can highlight problematic regions in the SOM and explore the data samples giving rise to them via the PCP. Alternatively, samples or regions in the PCP that the expert suspects of being problematic can be selected and the nodes in the SOM that are activated (fired) by these samples are also highlighted. This is an example of the use of an established exploratory technique called brushing and

linking (Buja *et al.*, 1991) but what is new here is that visualization tools are coordinating with sophisticated analytical methods.

The second example introduces a 2D map into the mix, coordinating in a similar manner with the PCP and SOM.

3.1 Experiment 1

The first experiment builds on previous work (Gahegan *et al.*, 2000) reported in the GeoComputation 2000, so details here are kept deliberately short. The data samples are drawn from a complex, ten dimensional feature space, describing a highly fragmented coastal landscape that is difficult to categorize due to the large number of different processes and disturbances (development, agriculture, fire) that have shaped it. Defining useful categories, that work in the spectral and informational sense described above, is difficult and requires expertise and careful study. One reason for this is the vast size of the feature space itself. Given that the individual data values in each dimension are represented within the range of one byte (0-255), the number of

possible values in this space is 10^{255} , a massive number. As is typically the case, most of this space is empty, but nevertheless an automated classifier must work through this space to partition the data. By lining up the dimensional axes vertically, the PCP creates a mapping through this space that allows the individual samples to be compared and visually grouped, and effectively removes the empty regions from consideration.

Figure 2 shows one view from the PCP where the strings have been colored to reflect five possible categories whose viability is to be investigated. These categories show as the rightmost axis in the figure. Note that the blue and green strings appear to cluster well and are visibly separate from other strings in many places. This is a good sign that they will be separable by the classifier. Note also that the yellow strings seem to group within the pink strings. This indicates that the yellow class is, in terms of the data at least, a subclass of the pink. Clearly, this will lead to separation problems. Finally, the red strings are all but obscured, indicating more potential problems.

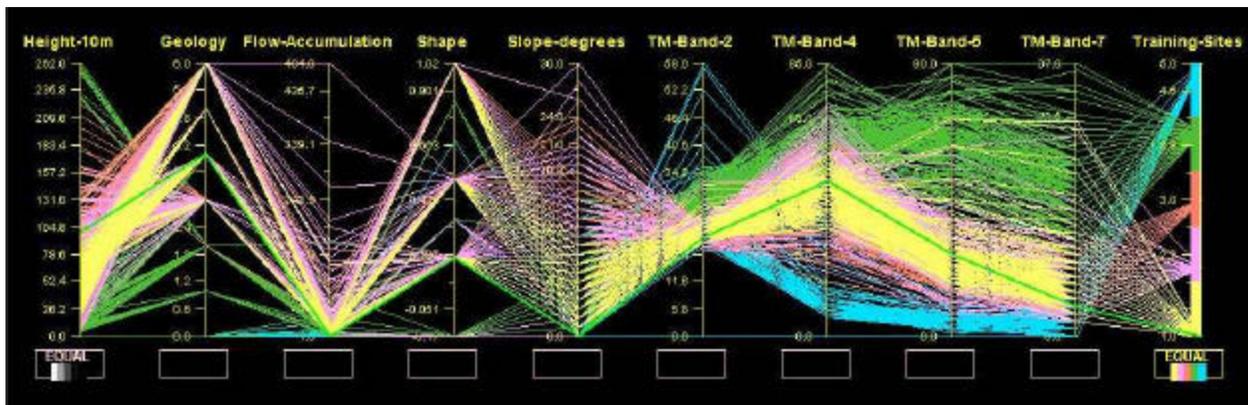


Figure 2: The parallel coordinate plot (PCP) displaying a complex feature space and 1703 data samples within it. Samples are shown as strings and are color-coded according to a proposed categorization.

Figures 3-7 show each of these five classes, in the order that they appear on the rightmost axis in Figure 2. Each image at the left side of the figure now shows just one proposed category in the PCP, and the right side depicts the neurons in the SOM that respond to this class after training using learning vector quantisation (Kohonen, 1995). Activated neurons are shown as larger spheres draped on a landscape-style surface. The topography of this surface is defined according to the distance in feature space between adjacent neurons on the grid, so any kind of relief between two neurons shows that they represent noticeably different data artifacts (and possibly concepts). Neurons that do not activate are shown as smaller gray dots. The color of each neuron indicates a measure of its internal error in representing the data, mapped continuously from green to red. So, green neurons show negligible error, whereas red neurons are under considerable ‘tension’ and therefore indicate a large degree of uncertainty. (see Takatsuka,

this volume, for full technical details of this implementation).

Examining each image in turn we see that the first class (water) is the best clustered within the SOM. By reading the topography here, we can also determine that this class has the strongest (furthest) separation from the other classes because it nestles neatly in a valley structure and is isolated from other regions of the U-Matrix. This represents ideal class behaviour; the concept is well-defined in the information space of the analyst *and* the feature space of the classifier. There is a small amount of class confusion, represented by a few outlying neurons that cluster on the ridgeline that separates this class from others. When examined in detail by “clicking” on those neurons (Figure 8) we see that they are activated by a few observations that show a high reflectance in the Landsat infra-red bands (4, 5 and 7). We would not normally expect this response from water, so they are atypical, and probably due to

poor image registration along the coastal plain. These infrared dimensions are normally the most discriminative of water from other classes, so the SOM struggles to deal with this apparent conflict, and this in turn

suggests to the analyst that further investigation is required.

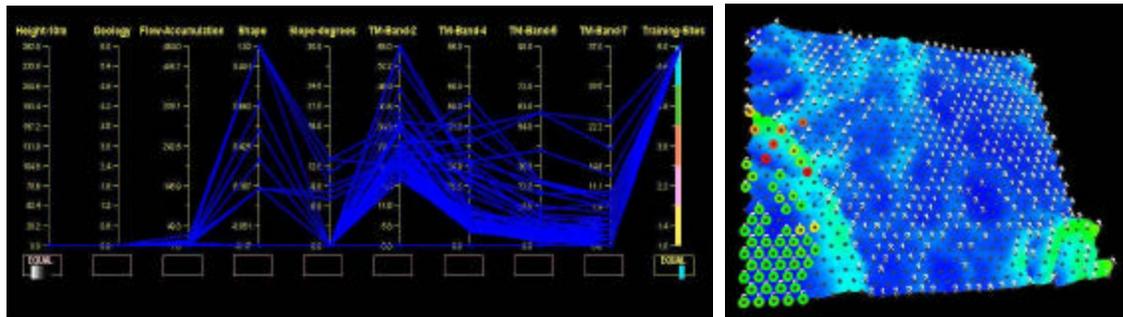


Figure 3: Proposed 'water' category in PCP and SOM. See text for details.

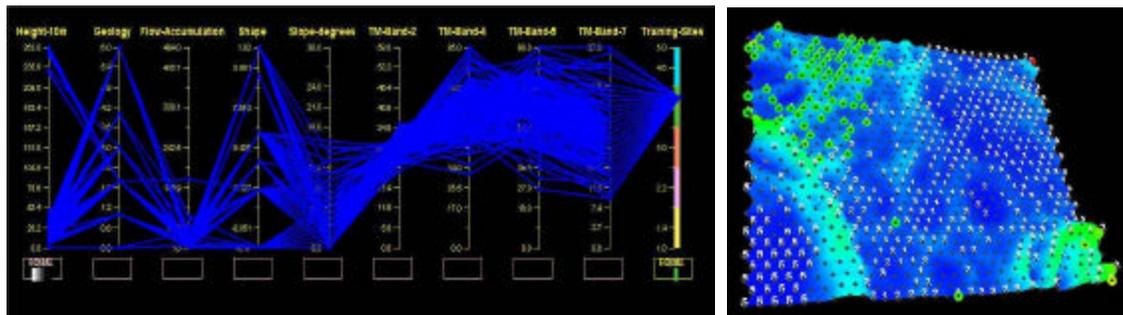


Figure 4: Proposed 'agriculture' category in PCP and SOM. See text for details.

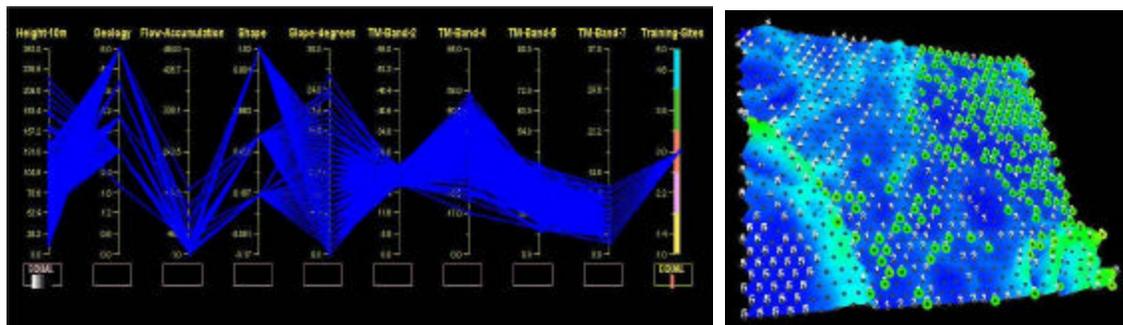


Figure 5: Proposed 'Upper-Slope-Forest' category in PCP and SOM. See text for details.

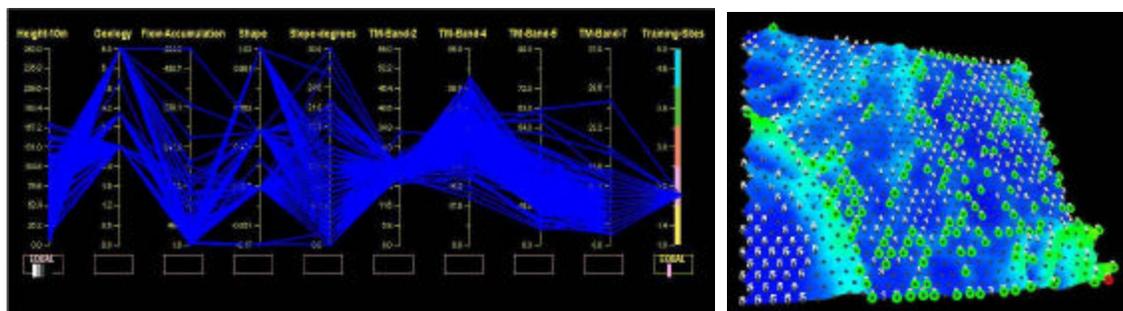


Figure 6: Proposed 'Mid-Slope-Forest' category in PCP and SOM. See text for details.

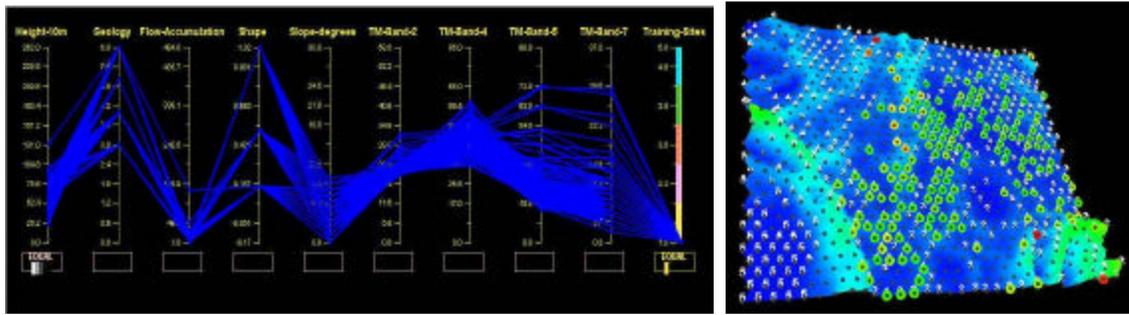


Figure 7: Proposed 'Lower-Slope-Forest' category in PCP and SOM. See text for details.

Notice in the above example how high level theory is used to form a hypothesis to explain the observed pattern, an example of the abductive power of human reasoning stimulated by the visual and computational tools used.

The second class, shown in Figure 4 above is agricultural (cleared) land and shows similar behaviour. Neurons activated by this class cluster well in the top left of the U-Matrix with very few neurons displaying any significant error.

The remaining three classes represent different types of forest habitat. They are more problematic to separate since they occupy a wide and intermingled swath within the SOM. Classes one and three do have some significant clusters evident, but not enough to allow separation of these concepts in the data. As was hinted at by the PCP in Figure 2, the SOM confirms that the pink category, that of mid-slope-forest, subsumes the other two. The needs of the analyst and the properties of the data are in conflict here, and will lead to a lowering of classification accuracy if not addressed.

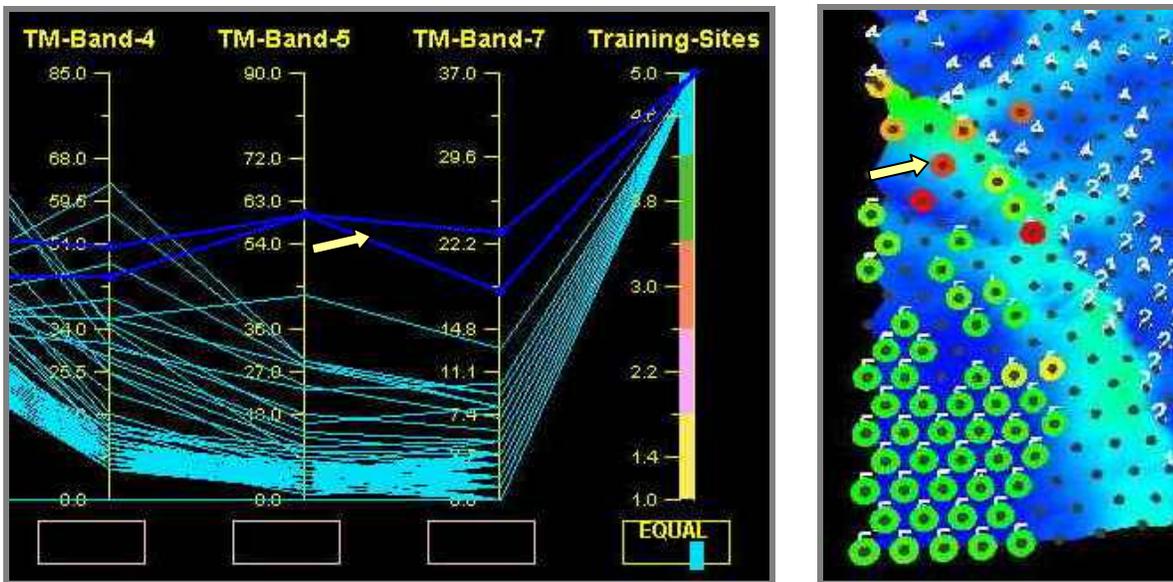


Figure 8: Close-up of outliers in the PCP (left) for the 'water' class identified by clicking on a red neuron in the SOM (shown by light yellow arrow).

3.2 Exploring State Demographics

The next example illustrates the usefulness of these tools in highlighting exceptional places or examples within a dataset. The PCP in Figure 9 depicts 23 census variables describing the 48 contiguous states in the USA. The highlighted string (shown in light green) represents California.

The PCP shows that California is often isolated from the other states in terms of its attribute values and indeed provides the lowest or highest endpoint for several data dimensions. In this example, color is provided in the PCP by visually classifying according to median rent values of residential properties (the leftmost axis).

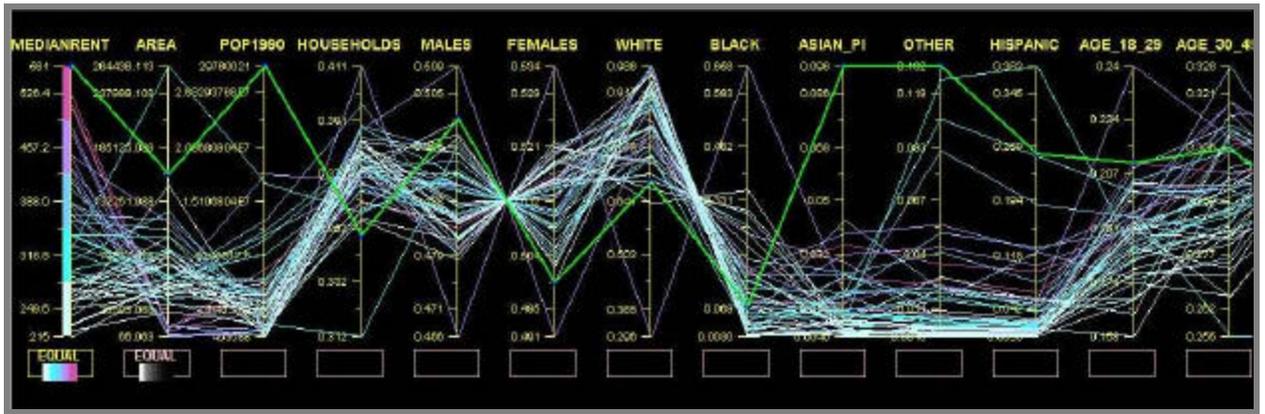


Figure 9: PCP showing 13 census variables for the 48 contiguous states of the USA. See text for details.

Figure 10 shows the linked map view, again with California highlighted in light green, via the coordination mechanism in *Studio*. Other states appear colored according to median rent, as in the PCP. Note the availability of much less expensive rental properties throughout the mid-west, and the more expensive eastern seaboard. Using the PCP we can also see that there are visible correlations between this variable and the population age profile and also with ethnic diversity.

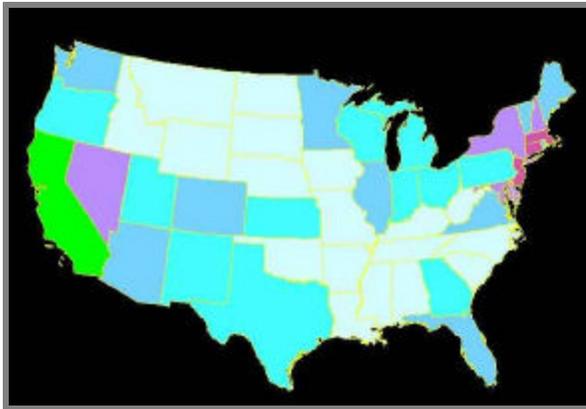


Figure 10: Corresponding choropleth map: median residential rental costs across the 48 contiguous states.

Finally, Figure 11 shows a quite alarming displacement in the SOM, which was trained in unsupervised mode on the census variables shown above. The striking red topographic feature shows just how distant in feature space California is from states to its immediate right in the SOM, and somewhat distinct from New York and Texas, its nearest neighbors vertically. There is an important caveat to be added here: when collapsing data from higher dimensional spaces into 2 or 3 dimensional form for visualization or analysis, regions close in the reduced space are not necessarily close in feature space; although they may appear so as a consequence of the mapping. In this example, California is about as unlike Virginia as it is possible to be (within the USA, at least). Dimensional compression techniques often attempt to preserve topology, but this is not always possible.

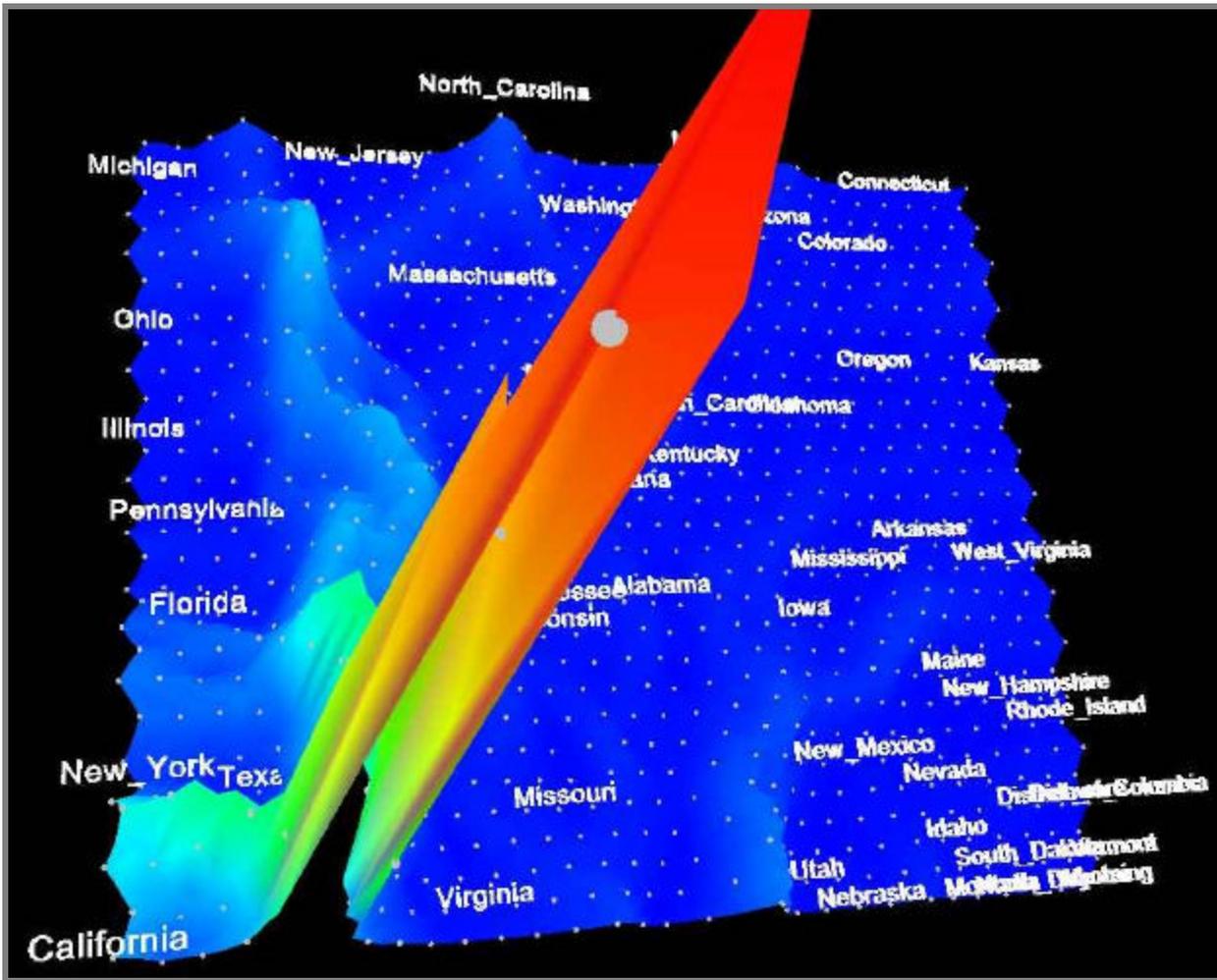


Figure 11: A visualization of the U-Matrix after the SOM is applied to state-level census data

4. CONCLUSIONS

This paper has shown, by worked examples, the application of both human expertise, and computational techniques to improve the quality and understanding of the categories by which geospatial data is organized and summarized. Notice that the computational tools are not used to replace the expert, but to harness her abilities and insights more directly. This expertise is brought to bear by a series of linked, coordinating visual tools that together aid the exploration of datasets that are both complex in terms of the structure contained, and deep in terms of the number of dimensions represented. As the above examples illustrate, human understanding directs the process of investigation, but is also deepened by it, in a synergy that appears to be rewarding for the analyst.

Future work will continue to explore such linkages, across a wider range of visual and computational methods, and will aim to embed the categorical descriptions produced directly into suitable representational structures (e.g. Sowa, 1999) for a more formal presentation and interoperation of their meaning. We will also continue to probe at deeper questions concerning the effectiveness of particular combinations

of machine and human expertise and the interfaces that empower them.

Acknowledgements

The support of NSF for this work (under grant EIA-9983445) is gratefully acknowledged.

REFERENCES

- Ankerst, M., C. Elsen, M. Ester, and H.-P. Kriegel. 1999. Visual classification: An interactive approach to decision tree construction. In: *KDD'99 Proc., Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, ACM Press, . pp. 392-6.
- Buja, A., J. A. McDonald, J. Michalak, and W. Stuetzle. 1991. Interactive data visualization focusing and linking. Proc. *IEEE Conference on Visualization (Visualization '91)*, San Diego, California. IEEE Computer Society. pp. 156-63.
- Buja, A., D. Cook, and D. Swayne. 1996. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics* 5(1): 78-99.
- Clancey, W.J. (1997). *Situated Cognition: on human knowledge and computer representations*. Cambridge University Press, New York.

- Cromley, R. G. (1996). A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International Journal of Geographical Information Systems*, **10** (4): 405-424.
- Fonseca, F. T., Egenhofer, M. J., Clodoveu, A. D. Jr., and Borges, K. A. V. (2000). Ontologies and Knowledge Sharing in Urban GIS. *Computers, Environment, and Urban Systems*, **24**(3), 251-272.
- Foody, G. M., McCulloch, M. B. and Yates, W. B. (1995). Classification of remotely sensed data by an artificial neural network: issues relating to training data characteristics. *Photogrammetric Engineering and Remote Sensing*, **61**(4), 391-401.
- Gahegan, M. (2000). On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, **32**(2), 113-139.
- Gahegan, M., M. Takatsuka, M. Wheeler, and F. Hardisty. 2000. GeoVISTA Studio: A geocomputational workbench. Proc., 4th Annual Conference on GeoComputation, UK, August 2000. [<http://www.geog.psu.edu/~mark/geocomp2000a/gc018.htm>].
- Guarino, N. (1997). Understanding, building, and using ontologies. *International Journal of Human-Computer Studies*, **46**, 293-310.
- Guarino, N. (1998). Formal Ontology in Information Systems. In N. Guarino (ed.) *Formal Ontology in Information Systems*. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. IOS Press, Amsterdam: 3-15.
- Hao, M., U. Dayal, M. Hsu, J. Baker, and R. D'Eletto. 1999. A Java-based visual mining infrastructure and applications. In: Proc., *InfoVis'99*, October 24-29, San Francisco, California. pp124-7.
- Inselberg, A. 1985. The plane with parallel coordinates. *The Visual Computer* 1: 69-97.
- Inselberg, A. (1997). Multidimensional detective. Proc. *IEEE conference on Visualization (Visualization '97)*, Los Alamitos, CA: IEEE Computer Society, pp. 100-107
- Jensen, J. R. (1999) *Introductory Digital Image Processing: A Remote Sensing Perspective* (Prentice Hall), 2nd edition.
- Jenks, G. F. (1977). Optimal data classification for choropleth maps, *Occasional paper No. 2*. Lawrence, Kansas: University of Kansas, Department of Geography.
- Kohonen, T. (1995). *Self-Organising Maps*. Springer-Verlag, Berlin, Germany.
- Kraak, M.-J., and A. M. MacEachren (Eds). 1999. *International Journal of Geographic Information Science* (special issue on exploratory cartographic visualization) **13**(4).
- Lakoff, J. R. R. (1978). *Wizards, Fire and Dangerous Rings*. Rivendell: Middle Earth Press.
- Lloyd, R., Patton, D., and Cammack, R. (1996). Basic-level geographic categories. *Professional Geographer*, **48**, 181-194.
- MacEachren, A. M. (1995) *How Maps Work*. Guilford Press, NY, USA.
- MacEachren, A. M., Wachowitz, M., Edsall, R., Haug, D. and Masters, R. (1999). Constructing knowledge from multivariate spatio-temporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographic Information Science*, **13**(4), 311-334.
- Mason, D. C., Corr, D. G., Cross, A., Hogg, D. C., Lawrence, D. H., Petrou, M. and Taylor, A. (1988). The use of digital map data in the segmentation and classification of remotely-sensed images. *International Journal of Geographical Information Systems*, **2**(3) 195-215.
- Mitchell, T. M. (1997). *Machine Learning*, New York, USA, McGraw Hill.
- Rosch, E. (1973). Natural categories. *Cognitive Psychology*, **4**, 328-350.
- Rosch, E. (1975). Cognitive representations of semantic concepts. *Journal of Experimental Psychology*, Vol. 104, No. 3, pp. 192-233.
- Rosch, E. (1978) Principles of categorization. In Rosche, E. and Lloyd, B., (Eds), *Cognition and Categorization*. Hillsdale, Erlbaum, p.27-48.
- Schuurman, N. (1999). Critical GIS: Theorising an Emerging Science. *Cartographica*, **36**(4).
- Smith, L. B. and Medin, D. L. (1981). *Categories and Concepts*. Cambridge, Harvard University Press.
- Smith, L. B. and Samuelson, L. K., (1997). Perceiving and remembering: Category stability, variability and development. In Lamberts K. and Shanks, D. (Eds.), 1997. *Knowledge, Concepts, and Categories*. MIT Press, Cambridge, MA, 161-196.
- Sutcliffe, J. P., (1993) Concept, class, and category in the tradition of Aristotle. In Mechelen, I. V., Hampton, J., Michalski, R. S., and Theuns, P. (Eds.) *Categories and Concepts: theoretical views and inductive data analysis*. Academic, New York, p.35-66.
- Takatsuka, M. and Gahegan, M., forthcoming. GeoVISTA Studio: a codeless visual programming environment for geoscientific data analysis and visualization. pre-print can be downloaded from <http://www.geog.psu.edu/~mark/>.
- Takatsuka, M. (2001). An application of the Self-Organizing Map and interactive 3-D visualization to geospatial data. This volume.
- Valdez-Perez, R. E. (1999). Principles of human computer collaboration for knowledge discovery in science. *Artificial Intelligence*, Vol. 107, No. 2, pp. 335-346.
- Wachowicz, M.. In Press. GeoInsight: An approach for developing a knowledge construction process based on the integration of GVis and KDD methods. To appear in: Miller, H. J., and J. Han, J. (Eds), *Geographic knowledge discovery and spatial data mining*. London, U.K.: Taylor & Francis.

Wisniewski, E. J. and Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, 18, 221-281.