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

(Extended Abstract: GeoComputation 2003)
Paper presentation

Mark Gahegan, Xiping Dai, James Macgill, Sachin Oswal and William Pike

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

Biography

Mark Gahegan is often mistaken for a graduate student because he is scruffy-looking. Despite that he is Professor of Geography, at the Pennsylvania State University. 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. His degrees are in computing and GIScience and his research interests include: visualisation, spatial analysis, remote sensing and machine learning applied to geography.

Problem Setting

Within the geographic domain there has been vast progress recently in the design, implementation and use of two distinct sets of tools, those that encode and depict conceptual structures such as ontologies and those that support exploration and knowledge discovery activities from rich geospatial datasets. These two types of tools represent both ends of a continuum from mentally-held concepts[1] and their relationships (a top-down view of the world) through to the actual data to be analysed (a bottom-up view of the world). Such tools are currently separated from each other with no means of interaction. But in reality, activities at either end of this continuum are not isolated, in fact they are intimately connected. 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 is a compromise between the user's conceptualisation of a problem and the structure and 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).

It is common in exploratory cartography / visualization, GIS and remote sensing for users to construct categories that help them better discover and comprehend the structure of complex

[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.

data. These categories may be simply 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).

For example, consider the case of land cover classification. Ontological tools that describe hierarchies of concepts (such as might be drawn from the Anderson landcover classification taxonomy) can offer sets of candidate categories from which a classifier might be trained, or conversely exploring the clustering of sample points in attribute space might lead one to hypothesize suitable mental concepts to represent these points.

Our Solution

We present a suite of tools, developed in GeoVISTA *Studio* (www.geovistastudio.psu.edu; Gahegan *et al*., 2002) that facilitate and connect together the processes of (i) specifying and browsing concepts ontologically, (ii) selecting concepts to use in a specific analysis exercise, (iii) operationalising the concepts with classifiers (iv) exploration to help formulate concepts from emergent structures in the data (v) modifying the concepts, 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 1 shows the actual tools developed in *Studio* to facilitate these activities, with arrows being used to indicate schematically some of their interactions. We 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.

References

Anderberg, M. R. (1973). *Cluster Analysis for Applications*, Boston, USA, Academic Press.

Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F. (2002). GeoVISTA *Studio*: a geocomputational workbench. *Computers, Environment and Urban Systems*, Vol. 26, pp. 267-292.

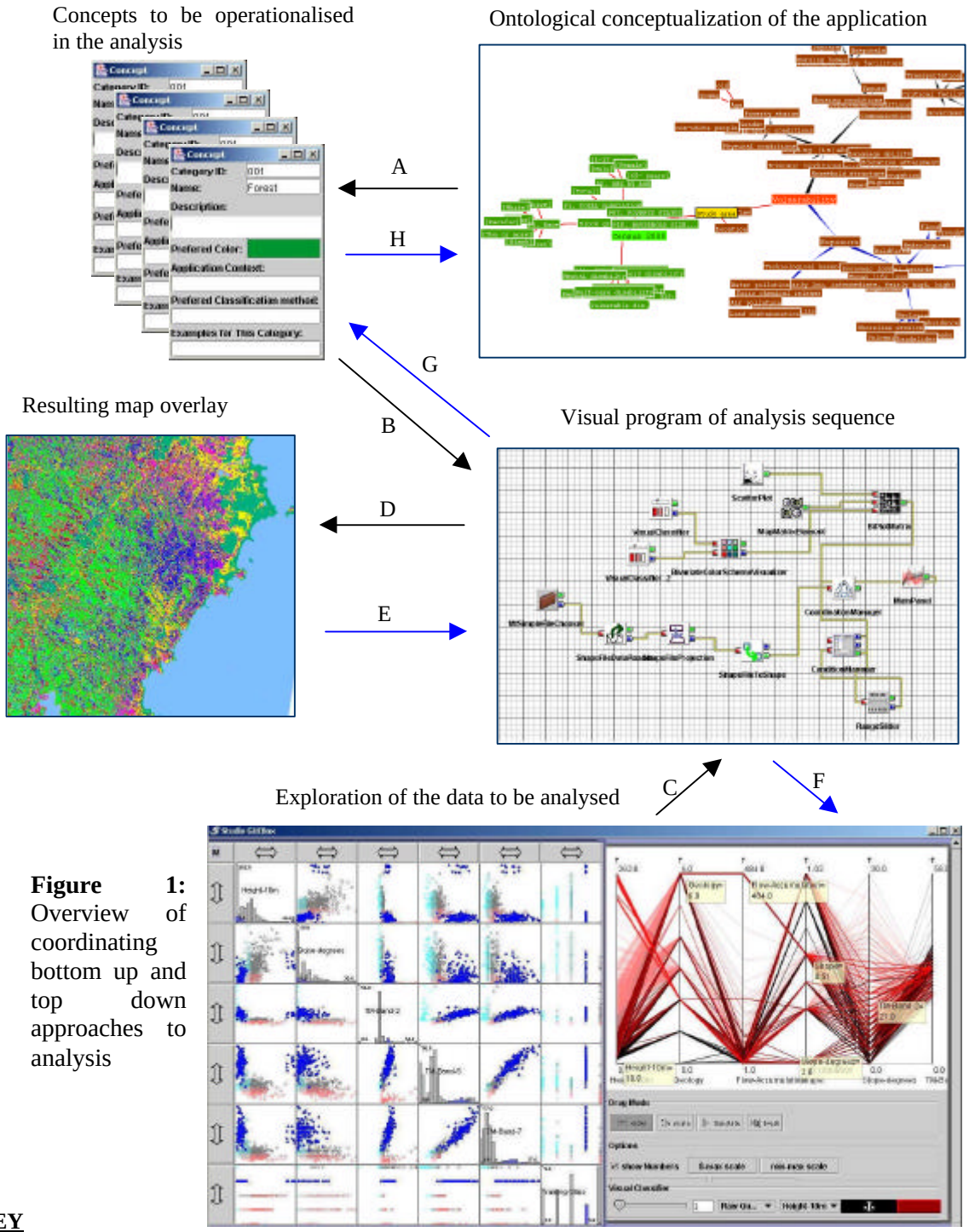

Figure 1: Overview of coordinating bottom up and top down approaches to analysis

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.