Automatic Geospatial Typology Development to Support Decision-Making and Visualization

Introduction

This is a cross-cutting project that seeks to develop a set of web-based geospatial data management, visualization, and decision support tools. We plan to evaluate the effectiveness of these tools on real-world problems such as habitat classification, coastal zone nutrient fluxes, and water management in the High Plains aquifer in Kansas.

The seed for this collaborative project has been the Land-Ocean Interaction in the Coastal Zone [LOICZ] typology group: a loosely organized set of researchers examining and estimating fluxes in basic nutrients in the coastal zone [1] This prior work has resulted in a well-defined process and a set of stand-alone and web-based tools for automatic typology development with geospatial data sets [3][2]. Typologies are useful classifications of data that point out similarities between geographic regions. We are now poised to expand upon this work, giving it broader applicability and direct application to current and future geospatial management issues.

Proposed Objectives and Approach

The primary objective of this project is to develop methods and tools for supporting decision-making with--usually high-dimensional--geospatial data sets. The approach we plan to take is based on automatic typology development. A typology provides information for decision making by 1) identifying important characteristics of geographic regions and 2) grouping areas that share similar characteristics. From data-driven typologies we can then identify the unique characteristics and spatial extent of the different types. We can also use typologies to develop predictive models for biological and chemical processes since the typologies highlight the driving variables that cause differences between geographic locations.

As an example, consider the High Plains Aquifer in Kansas. That state is facing immediate water management issues because the aquifer is being depleted by current usage patterns. The two questions the state needs to answer are:

- What water management strategies are required for different areas of the state?
- What are the geographic regions over which the different strategies should be applied?

The state has a detailed data base of geospatially distributed variables relevant to the health of the aquifer. Our claim is that a typology, developed automatically from the data set, forms a strong basis for making decisions in this situation. In particular, it provides information directly pertinent to the water management issues identified above. As an example, consider Figure 1, which is a visualization of an example typology of the High Plains Aquifer in Kansas.

The questions we can explicitly answer with a typology include:

- What variables are the most influential within each class of the typology?
- What are the most important characteristics of each class?
- How are the classes spatially related to one another?

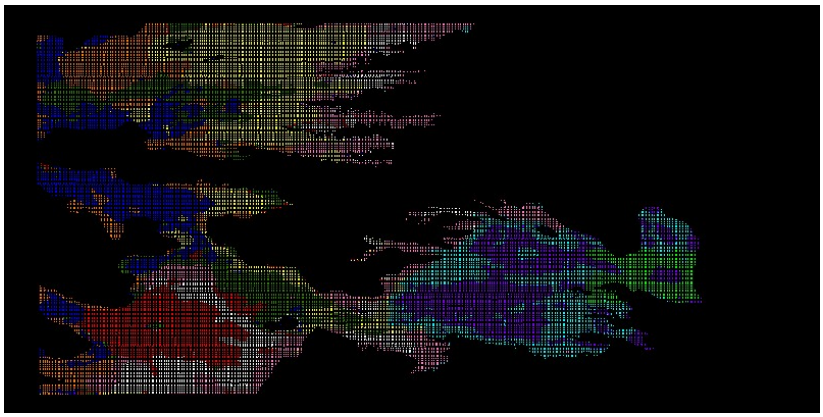

Figure 1 Example 10-class typology of the High Plains Aquifer data set. Each color represents a different class in the typology.

Questions that require further analysis and development to answer are:

- How do you measure similarity in a high-dimensional geospatial data set?
- How many classes naturally exist in the data set for a given similarity measure?
- How do the classes in a typology match up with ecological or environmental models?
- How can we use a typology to derive predictive models of biological/chemical processes?
- How is a typology affected by the quality of the data we are using?
- How can we do all of this in an interactive, collaborative, web-based manner?

Our second objective is to develop methods for answering all of these questions automatically, thus providing support for decision-making. These questions form the basis for our research plan.

Comparing high-dimensional geospatial data points. We propose comparing more traditional statistical distances--Euclidean distance, scaled Euclidean distance, and Mahalanobis distance--with less traditional approaches. One approach that shows promise in a geospatial/ecological context is maximum scaled distance. This distance measure is based on the variable that shows the greatest statistical difference between two multi-variable geospatial data points. In an ecological context this makes since extreme characteristics tend to drive or limit ecological systems.

Discovering appropriate numbers of classes. We propose using information theoretic measures to identify the balance point between model complexity--the number of classes--and the modeling error--how well the classes represent the data set. Rissanen's minimum description length criteria, for example, gives a well-defined measure in the case of Gaussian distributions of variables [4]. An important aspect of this part of the research is developing independent methods of determining appropriate numbers of classes and evaluating the results using expert judgements.

Matching typologies with ecological models. As part of our previous work with typologies we have compared typology outputs with the outputs of expert judgements and model-based analysis. Based on these informal comparisons, it is clear that typologies based on high-dimensional geospatial data sets are capturing a significant amount of the information created by ecological or hydrological models of the same areas. We propose exploring this relationship more explicitly. In particular, we want to explore whether a typology can be used as a proxy for complex ecological models. If so, it would be a more efficient method of getting similar information. It may also permit the projection of the typology analysis to areas where there is insufficient data for the ecological models to function.

Deriving predictive models from typologies. In many geospatial data sets there are certain basic variables--climatalogical or hydrological, for example--that are related to other qualities of the environment such as biological flora and fauna and chemical fluxes. A typology based on the basic variables produces a set of typical environments. Based on sparse measurements of biological or chemical variables in each of these typical environments we argue it should be possible to develop predictive models based on the typology. This hypothesis is verifiable, and suggests a way of extrapolating data collected in one area to a larger set of geospatial locations.

Assessing data quality through typology development. Our previous work has suggested that typology development is a useful exercise for assessing data quality. Classes that appear in a typology that have odd characteristics are easy for a user to see in a visualization and straightforward for a computer to flag and bring to the attention of the user. While there is no guarantee that bad data will end up in its own class, there are steps we can take during the typology development process that will encourage that to happen. We believe there are synergies between initializing the typology development process and identifying bad data. These synergies should be explored.

Collaboration and Web-Based Visualization. Our previous work has resulted in a web-based program for analysis and visualization of geospatial data [5]. The major issues we need to address now include: how to build effective collaboration support into the web framework, how to create intuitive 3D visualizations with geospatial data sets, and how to visualize typologies so that the color relationships encode information. We have already explored the typology coloring issue [2]. The collaboration and visualization issues require a combination of expert users and system developers, which is the real strength of our seed group of LOICZ researchers that combines people with a variety of expertise.

Expected Outcomes

The expected outcome of this project is a set of web-based data management, analysis and visualization tools for high-dimensional geospatial data sets. This web site would be available to researchers around the world and would be setup to support collaborative analysis. In addition to pre-installed data sets, we would also permit users to upload their own data sets for analysis and visualization. We believe this would be a powerful analysis and decision-support tool for high-dimensional geospatial data sets.

References

[1] LOICZ [Land Ocean Interactions in the Coastal Zone], http://www.nioz.nl/loicz, 2000.

[2] B. A. Maxwell, “Visualizing Geographic Classifications Using Color”, to appear in J. of Cartography, 2000-2001.

[3] B. A. Maxwell and R. Buddemeier, “Coastal Typology Development with Heterogeneous Data Sets”, Kansas Geological Survey Open-File Report 2000-53, submitted to J. of Regional Environmental Change, April 2000.

[4] Rissanen, J (1989) *Stochastic Complexity in Statistical Inquiry*, World Scientific Publishing Co. Ptc. Ltd., Singapore.

[5] Web-LoiczView, http://www.palantir.swarthmore.edu/loicz, 2000.