# Experimental Use of an Unsupervised Classification Technique on Historical Land Use Data

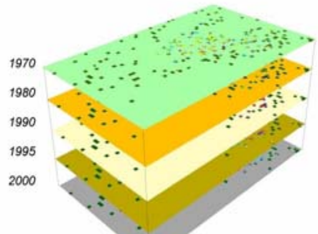
Michael Zoldak, Santiago Lopez, Jana Fry, C. Scott Smith and Charles Redman
CAP-LTER and the Arizona State University Information Technology, Research Support Lab, GIS Services

## INTRODUCTION

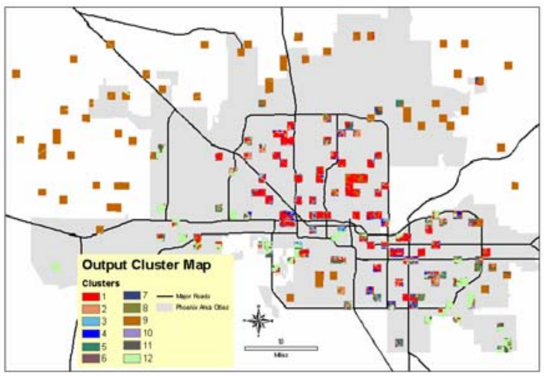
The LTER historical land use database for Greater Phoenix, Arizona has both spatial and temporal components. In the GIS database, individual square mile areas surrounding a stratified random sample of more than 200 study sites are composed of many different land use polygons. In addition, there are different GIS map layers representing land use at particular moments in time. The historical data were derived from air photos and historical records. Within each layer, change within the square mile can be described and analyzed, and, at any given location, change through time can be described and analyzed. The challenge is to do both simultaneously. The work presented here is a first, experimental, attempt to use an image analysis technique, unsupervised classification, to simultaneously describe both spatial and temporal changes.

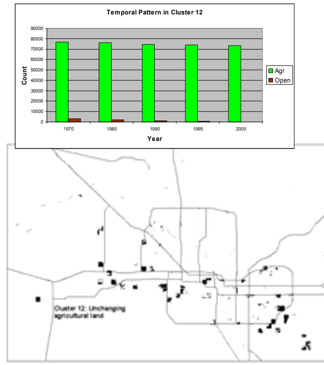
## METHODS

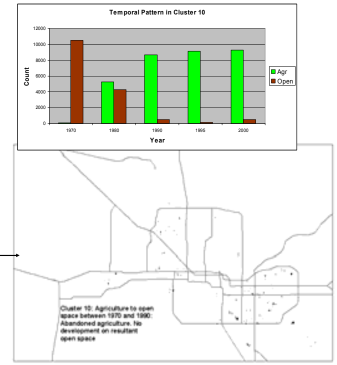
The ISODATA (Iterative Self-Organizing Data Analysis Technique) clustering algorithm, widely used for unsupervised classification of raster data, attempts to find clusters of pixel values in attribute-space and refines the resultant output clusters based on the spatial proximity of pixels within attribute-space clusters and the user-specified maximum number of clusters. It is deemed unsupervised because the output clusters are not defined beforehand.

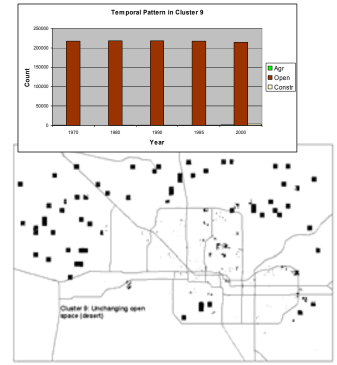



Several processing steps were required before applying this technique. Using Arc/Info, the LTER land use was generalized from 45 to nine broad Land use classes for the five GIS map layers representing the years 1970, 1980, 1990, 1995, and 2000. Each layer was then converted to a raster grid with 100-foot resolution (this is safely below the resolution of the polygon layers, many of which were created from air photos with sub-meter resolution). Using ERDAS Imagine, the raster layers were converted and combined into a multi-spectral image. The unsupervised classification routine packaged with Imagine was used in multiple configurations until twelve output clusters were settled on as optimal. Runs with more clusters created small groupings that were difficult to interpret. Runs with fewer clusters lumped together spatial and temporal patterns and severely limited the method's usefulness as an analytical tool. Along with the output cluster map, a tabular summary of the area (literally, raster cell counts) of each land use class within each cluster for all five years was created and used in interpreting the results.



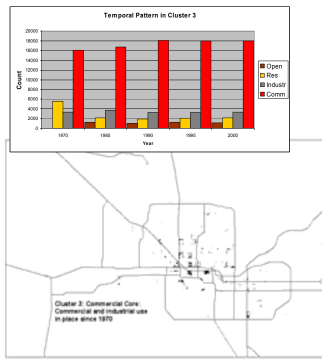



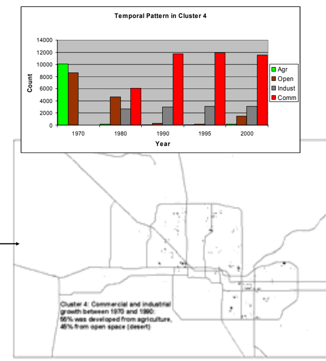



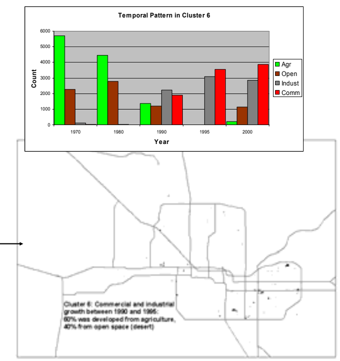

## INTERPRETATION

Due to the scale and fragmentary nature of the dataset, interpretation of the output clusters map was problematic. Each of the twelve clusters displayed distinct spatial and temporal patterns that became more obvious when viewed individually. The spatial pattern of each cluster and the temporal changes in land use within each cluster were used to classify them. Further interpretation revealed that some of the clusters were linked temporally. Many showed similar changeovers in land use type, but occurring at different times.

The results of the interpretation of the twelve clusters are arranged according to their temporal linkages. The maps have graphs associated with them that demonstrate the temporal change in land use within each cluster that was used in the interpretation.

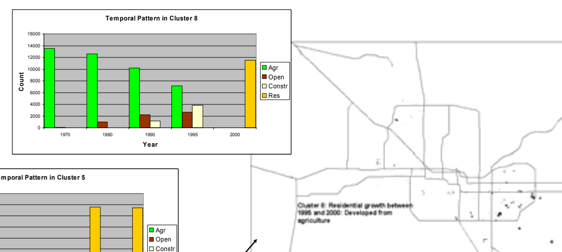


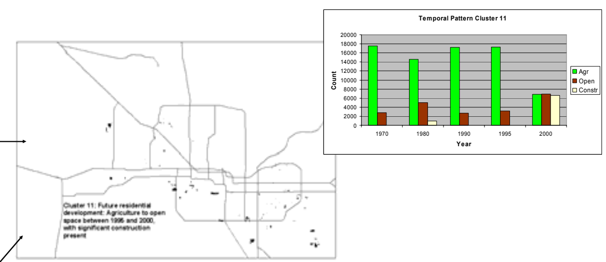



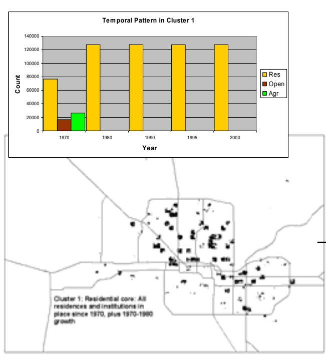



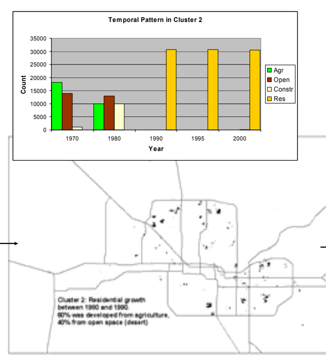

## FUTURE WORK

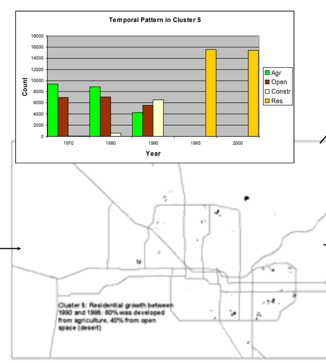
Future research will focus on overcoming some of the more obvious deficiencies with the technique presented here. There are two major deficiencies. The first is that the land use classes are categorical, yet need to be represented as ratio data in order to perform the unsupervised classification. It is likely that some types of land use and land use change patterns were clustered together based purely on the relative closeness of the numbers assigned. Current remedies for this are in the development stage. These include the incorporation of logit modeling and related techniques that account for the limitations of categorical data and limited dependency among variables. The second deficiency is that the selection of the number of output clusters and their interpretation is open to improvement. Refinement of the temporal component of the dataset and the use of sub-classification schemes are being researched.

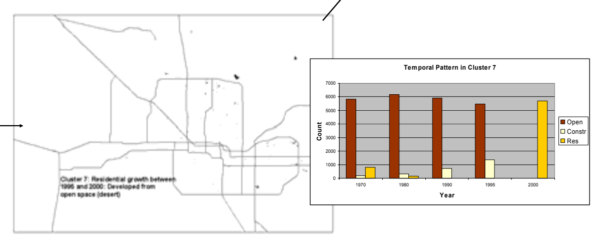



ASU Arizona State University    CAP LTER

September 27, 2001