A Visual Approach to Data Mining Spatial and Temporal Change

Grant Fraley[1], Piotr Jankowski[1], & Cristiano Giovando[1]

[1] Department of Geography, San Diego State University
fraley@rohan.sdsu.edu

INTRODUCTION

This research explores over twelve million records from the United States Border Patrol (USBP) ENFORCE Integrated Database (EID) describing illegal immigrant apprehensions, each with a geographic origin and destination, as well as a temporal reference. Methods for data mining and visualization facilitate such an exploratory analysis, aiding the emergence of hypotheses and prescriptions for further analytical inquiry (Andrienko and Andrienko, 2006). This research uses spatial autocorrelation statistics to illuminate outliers and patterns (Ord and Getis, 2001) in space and time of a statistic of immigration propensity change. Spatial autocorrelation statistics are commonly used to detect spatial clusters of the magnitude of an attribute at a point in time, a cross-section. However, this study reveals spatial clusters of change in the magnitude of an attribute between time points, a temporal lag.

Information is derived from the database by detecting spatial clusters of immigration propensity change. Yet, a large amount of information is produced when considering several origin-destination paths over all time lags. A software tool was designed to organize and visualize immigration propensity change maps, global spatial autocorrelation statistics, and local spatial autocorrelation maps. The tool guides the analyst through exploration of change using linked visualization of statistics and maps.

CLUSTERING CHANGE VALUES

An index of immigration propensity was calculated for each Mexican municipio with respect to each southwesterly USBP sector using EID data from fiscal year 1999 through 2006. The migration propensity index (MPI) (Weeks et al, under review) is a ratio of the percentage of apprehended Mexicans aged 20-34 apprehended in a destination sector (*D*) originating from a municipio (*i*), to the percentage of the Mexican population aged 20-34 (*P*) that reside in municipio *i*. The MPI is calculated as:

- $$MPI = \frac{D_i / \sum D_i}{P_i / \sum P_i}$$

The MPI was calculated with yearly apprehension aggregates. Values of MPI greater than one indicate higher than expected immigration to a sector from a municipio given its share of the population of Mexico. MPI values less than one indicate lower than expected immigration propensity to the sector. Changes over time in immigration propensity were calculated by taking the absolute difference in municipio MPI values to a sector between each possible pairing of years.

Data mining methods for spatial cluster detection were adopted to guide the analyst in gleaning information from the many MPI change maps. The methodology involves calculating a weighted Ripley's K (Getis, 1982) to assess the distance at which MPI change reaches maximum spatial autocorrelation. Ripley's K values then guide the choice of threshold distances for the calculation of the Getis-Ord General G (Getis and Ord, 1992), and Moran's I (Moran, 1950) global spatial autocorrelation statistics on MPI change. Moran's I analysis produces a statistic and z-score for each temporal lag measuring the degree of spatial clustering of similar or dissimilar municipio MPI change in Mexico. Similarly, Getis-Ord G analysis produces a statistic and z-score describing the degree of clustering of high or low values of MPI change. Local Moran's I (Anselin, 1995) and Getis-Ord Gi* (Ord and Getis, 1995) local indicators of spatial autocorrelation were calculated to enable identification and mapping of clusters of MPI change.

VISUAL DATA MINING

A visual approach to data mining using spatial autocorrelation statistics was adopted in this research. A software tool was developed to organize the Ripley's K, Global Moran's I, and Getis-Ord G statistics in matrices where each element corresponds to a temporal lag. Global statistic matrices may be displayed for any of the destination sectors (Figure 1).

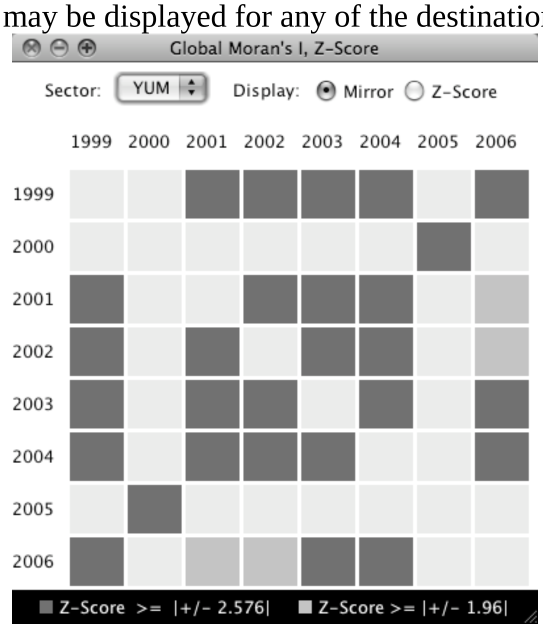

Fig. 1: The Global Moran's I matrix is displayed for Yuma sector.

Elements of global statistic matrices are clicked in the tool to bring up a linked map display of the corresponding local statistic. So, an analyst might notice the indication of clustering in high values of MPI change at the one-year lag from 2005 to 2006 for Tecate sector. Clicking this element in the Getis-Ord G matrix displays the Gi* map for Tecate sector that

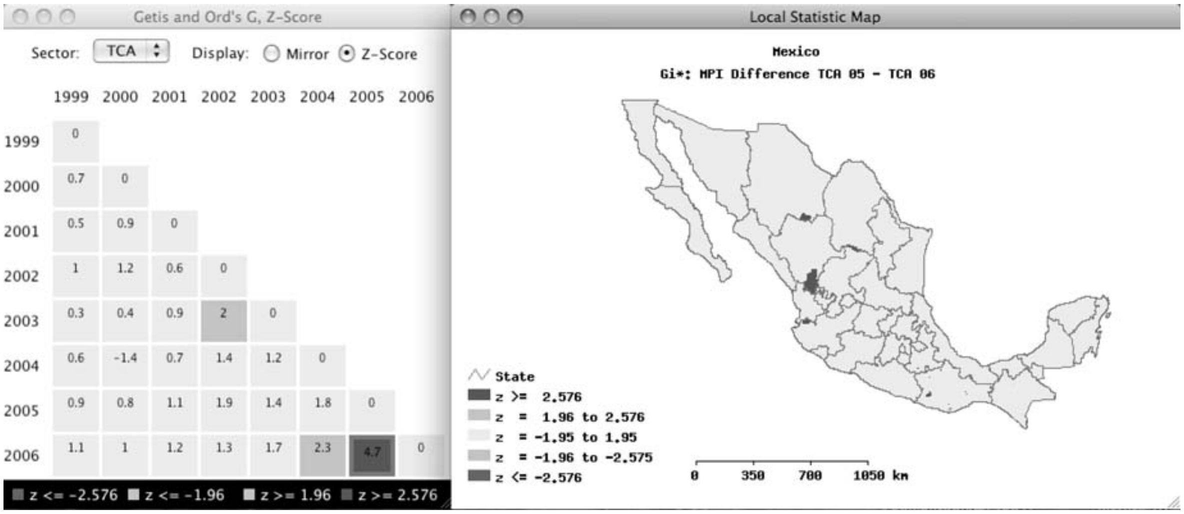

Fig. 2: Getis-Ord G matrix and linked 2005-2006 Gi* map for Tecate sector.

reveals a statistically significant cluster of municipios with high MPI change in the state of Durango, Mexico (Figure 2). Given the realization of this cluster, an analyst may then refer to the map of MPI change from 2005 to 2006 for Tecate. In such a manner, a flexible and iterative exploration of MPI change may occur.

CONCLUSION

A measure of immigration propensity, MPI, was explored with spatial clustering statistics through the magnitude of its change at various temporal lags to reveal temporal features of its geographic distribution. A tool for visualization was developed to organize the statistics so one can recognize features of global clustering to guide investigation of local cluster maps (Ord and Getis, 2001), then subsequently reference maps of MPI change values. Thus, the visualization tool avoids tedious examination of calculated statistics and maps while enhancing information extraction.

REFERENCES

Andrienko, N. and G. Andrienko (2006). Exploratory Analysis of Spatial and Temporal Data, A Systematic Approach. Heidelberg, Springer-Verlag.

Anselin, L. (1995) “Local indicators of spatial association - LISA.” Geographical Analysis 27: 93-115.

Getis, A. (1982) “Interaction Modeling Using Second-order Analysis.” Environment and Planning A 16: 173-183.

Getis, A. and J.K. Ord (1992) “The Analysis of Spatial Association by Use of Distance Statistics.” Geographical Analysis 24: 189-206.

Moran, P.A.P. (1950) "Notes on continuous stochastic phenomena." Biometrika 37:17-23.

Ord, J.K. and A. Getis (1995) "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application." Geographical Analysis 27: 286-306.

Ord, J.K., and A. Getis (2001) "Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation" Journal of Regional Science 41 (3): 411–432.

Weeks, J., J. Stoler, and P. Jankowski. "Migrant Roots in Mexico: Origins of Detainees at the U.S.-Mexico Border". Under review.