Using Spatial Statistics In GIS

K. Krivoruchko[a] and C.A. Gotway[b]

[a]Environmental Systems Research Institute, 380 New York Street, Redlands, CA 92373-8100, USA

[b]Centers for Disease Control and Prevention; 1600 Clifton Road NE Atlanta, GA 30333, USA

Abstract: A Geographical Information System (GIS) provides a powerful collection of tools for the management and visualization of spatial data. These tools can be even more powerful when they are integrated with methods for spatial data analysis. In this context, we provide several examples that show the power of exploratory spatial data analysis (ESDA) within a GIS and how this can provide the foundation for more sophisticated probabilistic modeling. While the ESRI's ArcGIS software now facilitates the integration of spatial data analysis and GIS functionality, more tools are needed for comprehensive spatial data analysis. We suggest how to implement additional spatial statistical methods within a GIS, including methods for using non-Euclidean distances in the analysis of geostatistical, lattice, and point pattern data.

Keywords: *Non-Euclidean distance, spatial correlation, GIS, geostatistics, lattice, marked point pattern, exploratory spatial data analysis*

1. INTRODUCTION

Statistical analysis within a commercial GIS (Geographical Information System) is rapidly becoming an impressive suite of tools. Until recently, statistical analysis was limited to visualization and exploratory data analysis, while statistical modeling was considered problematic for implementation within a GIS. However, probabilistic reasoning and statistical modeling are now important components of GIS science, and users of commercial GIS software are beginning to want more sophisticated statistical tools for spatial analysis.

We discuss our ideas for data exploration and modeling within the ESRI GIS, trying to balance a variety of user needs with a software developer's perspective. We believe that GIS provides a practical approach to data exploration and this helps to identify areas where statistical modeling could be most useful. In this context, we suggest new methods for future development and implementation in GIS.

2. SPATIAL ANALYSIS AND SPATIAL DATA ANALYSIS IN GIS SOFTWARE

Bailey and Gatrell (1995) distinguish between *spatial analysis,* the study of spatial phenomena using the basic GIS operations such as spatial query, join, buffering, and layering, and *spatial data analysis*, the application of statistical theory and techniques to the modeling of spatially-referenced data, which is the discipline of spatial statistics. ESRI's GIS software includes modules that address both tasks, namely the Spatial Analyst and the Geostatistical Analyst extensions to ArcGIS. While there are some similarities between these two extensions, there are also some key differences. Spatial Analyst functions allow the user to construct maps of where things are and how they change, find what is inside or nearby, and identify the largest and smallest values in the area under investigation. Simple descriptive statistics and statistical graphics such as means, standard deviations and pie charts are often enough to quantify the variability in the data and the results. Zonal addition, proportional allocation, and buffering are often sufficient for combining data from different spatial units. Figure 1 shows the main menu of the Spatial Analyst program that provides a comprehensive suite of deterministic functions for spatial analysis.

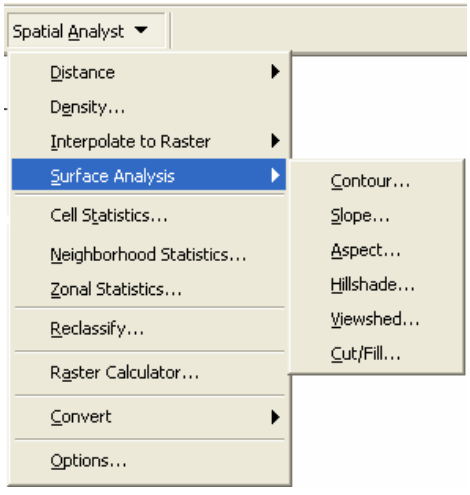

Figure 1. Main menu of the Spatial Analyst.

For many GIS users, these analyses provide more than enough capability for most of the applications of interest. However, the analyses ignore uncertainties in the data and the results and produce new surfaces without taking into account the errors that propagate with each operation on the data. Unfortunately for some researchers, this will not give an adequate analysis and many users require much more sophisticated methods for spatial data analysis. These necessarily require inferential spatial statistics: estimation, prediction, and hypothesis testing. ESRI's Geostatistical Analyst extension to ArcGIS provides a greater suite of both qualitative and quantitative statistical tools for spatially continuous data. Figure 2 shows the main menu of this extension with a list of the exploratory spatial data analysis (ESDA) tools that are included. When these tools are implemented, graphical dialogs link the results to maps and data tables, allowing the user to quickly and effectively assess the variability, distribution, correlations and cross-correlations, and large-scale trends in the data.

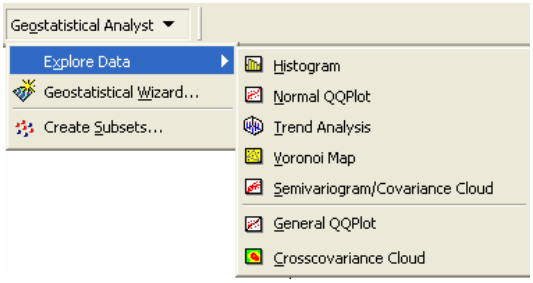

Figure 2. Main menu of the Geostatistical Analyst.

For modeling the user can select from a variety of kriging models with output in the form of predictions, prediction standard errors, quantile and probability maps (Figure 3).

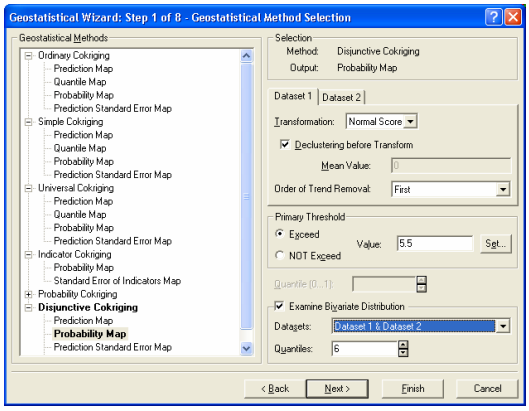

Figure 3. The geostatistical method selection dialog.

Additional tools include several approaches for data transformation and detrending. The user can rely on the default parameters or make more specific choices using the graphical dialogs that facilitate covariance modeling, selection of search neighborhood, and provide validation and cross-validation diagnostics. Many options for mapping and post-processing are available.

Geostatistical Analyst has several thousands of users and they have very different backgrounds and interests. Unfortunately, many researchers use the software only to make maps. Many statisticians do not understand or appreciate the full utility of GIS for spatial data analysis. Other researchers are not educated in spatial statistics, so they are unaware of techniques for modeling uncertainty, even though they realize that measuring and modeling without errors is impossible. At the same time, they readily use automatic "geoprocessing" tools that arithmetically add or average raster data values in the cells without taking into account the impact of error propagation on the results. After several such geoprocessing steps, the resulting data structure can be completely random and, consequently, decisions made from these results may be wrong. Still others use the software for analyses for which it was not designed. This problem often arises when users implement Geostatistical Analyst with aggregated data that are associated with spatially discrete units. For example, we have watched users discuss which interpolator, inverse squared distance weighting or splines, is better for mapping of proportions of females in burial populations. What is the best prediction of the proportion of females outside cemeteries? One would hope it is zero. There are other methods in spatial statistics that are more suitable for this type of data. For example, a marked point pattern analysis that incorporates an attribute value recorded at each location could be used. The mark would be the occurrence (or not) of a female burial at each location, and a marked point pattern model could then be used to estimate and map the intensity of females in burial populations, assuming that data point locations are given by nature and not selected by the user.

Hopefully, all of these problems can be solved by education. Case studies can help to show users how GIS can and should be used for more sophisticated statistical analysis and modeling.

3. IMPROVED SPATIAL DATA ANALYSIS USING GIS

Good data visualization is important both for data understanding and for representing the results of statistical analysis. Without a GIS, users may struggle to create meaningful visualization tools. For instance, consider the graph in Figure 4 that displays land sales with different characteristics.

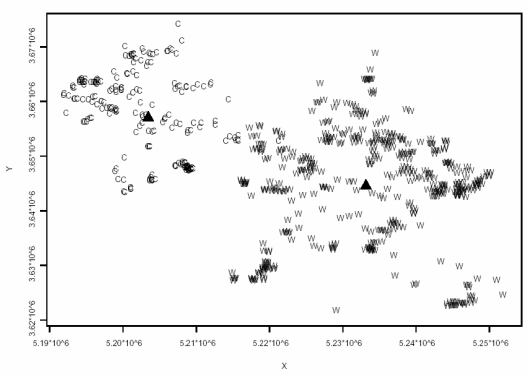

Figure 4. Typical data representation without GIS usage.

Graphs like this are typical and plentiful in many of today's journals, but the utility of this graph is very low. In many papers on spatial statistics, the authors do not even use graphics but only one-dimensional graphs and tables with estimated parameters and diagnostics. Land cost depends on many factors and many of them are readily available for GIS users (street networks, school and shopping locations, etc). Visualizing these factors with a map could be a very valuable tool for understanding land costs and the results on any statistical analysis.

Good visualization should be the prelude to sophisticated modeling. For example, consider the contour map in Figure 5 that is typical of maps displayed in statistics journals. Such maps are often the sole visualization tool used to support the development of sophisticated statistical models. This illustration is not very helpful for understanding the spatial variation in the data, or for investigating hypothesis about the reasons for such variations. Unfortunately, maps like this are plentiful in the scientific literature.

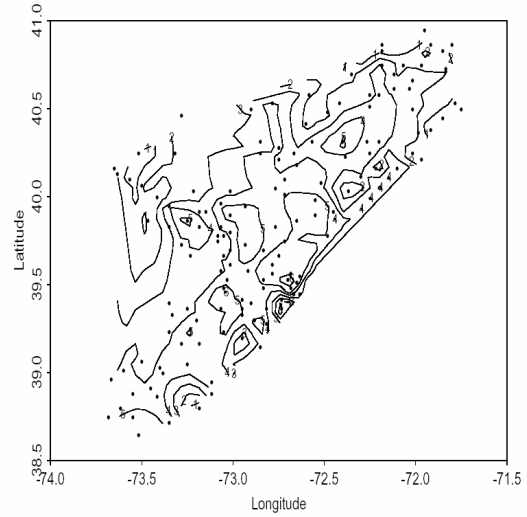

Figure 5. The result of predictions is visualized without GIS.

Consider instead the type of map that can be produced using GIS. Figure 6 shows a map of ozone measurements in 1999 in Southern California. The city of Los Angeles lies in a coastal plain, surrounded by mountains that separate a desert from the coastal climate. In the summer, pollutants in the lower layer of air (smog) move from the city to the east, but are blocked by the mountains. So, to obtain a better understanding of how pollution might move, the topography is also displayed in the same map. One of the major sources of air pollution is exhaust from motor vehicles, so the major road networks are also displayed on the map. Good data visualization can help us to construct a model for air quality. For example, we can now easily see a large-scale east-west trend in the measurements and the barrier to movement that the mountains provide. The larger values of ozone tend to be close to the mountains in the east, and the ozone concentration declines toward the coast. Thus, any model we select should account for this trend. This can be easily investigated further using Geostatistical Analyst's ESDA tools and the results can then be used to choose an appropriate geostatistical model.

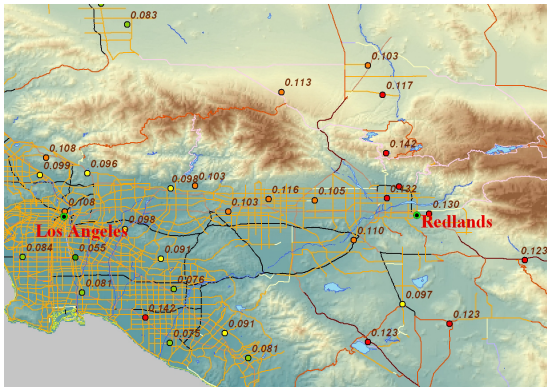

Figure 6. Ozone measurements, elevation, and major road network visualization using ArcGIS, Los Angeles area.

There are other useful GIS features which can also improve statistical data analysis. With advances in GIS, the ability to find and use necessary data has grown tremendously. For example, ArcGIS allows users to find and immediately use relevant data associated with a selected region on the map by displaying a list of all data sources available for this area on the Internet servers. The user can also do a search for a specific type of data. Even if the user did not find exactly what is necessary, similar data can be used for comparison with data under investigation.

Modern GIS software allows the management of very large datasets. This is particularly important

in the environmental sciences because large, remote-sensing images are easily obtained for minimal cost. As another example, California daily measurements of air quality are available for many cities for the last 22 years and queries within a GIS can be very helpful for exploring such large data sets. With so much spatial data available over the Internet, we can easily obtain elevation measurements, census and epidemiological data, land use classifications, and meteorological information. Often these data are collected using different coordinate systems, and a GIS is very helpful in changing projection.

4. SPATIAL STATISTICS TOOLS AND MODELS FOR IMPLEMENTATION IN GIS IN THE NEAR FUTURE

One interesting consequence of developing statistical software for a large audience is the possibility of learning what "typical" users want. We found that the ideal world where data are accurately measured and normally or log-normally distributed is not common in most user applications and the spatial coordinates of the data are often not known exactly. In this section, we will discuss two important options to be implemented in the spatial data analysis software in the near future: the use of non-Euclidean distances and methods for adjusting for measurement and locational errors.

Non-Euclidean distances

Many GIS users are analyzing data in the environment with natural and artificial barriers. For many applications, the map in Figure 6 that uses meteorological, elevation, and traffic data in addition to the pollution measurements, together with some basic results from typical GIS functions like buffering, will be enough to understand the spatial distribution of ozone. However, many applications require more sophisticated analytical methods. Continuing with the example above, Figure 7 shows a 3D view of the area near Los Angeles shown in Figure 6.

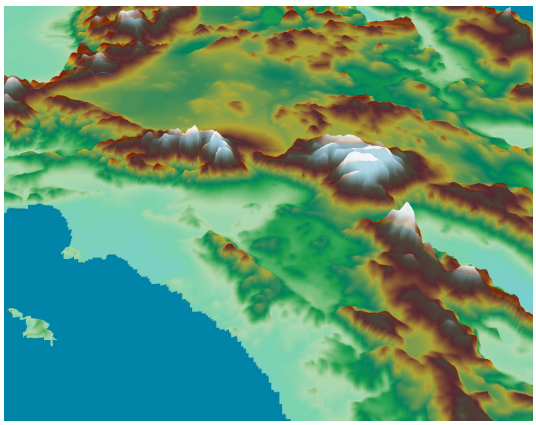

Figure 7. 3D view of the area near Los Angeles.

Intuitively, distances between objects on this map should not be "as the crow flies." Figure 8 shows the result of ozone concentration predictions over the elevation map. In this analysis, the distance between locations was calculated using a non-Euclidean distance metric that incorporates the mountainous barrier to ozone movement. This metric constructs a cost surface based on information on altitude (Krivoruchko and Gribov, 2002).

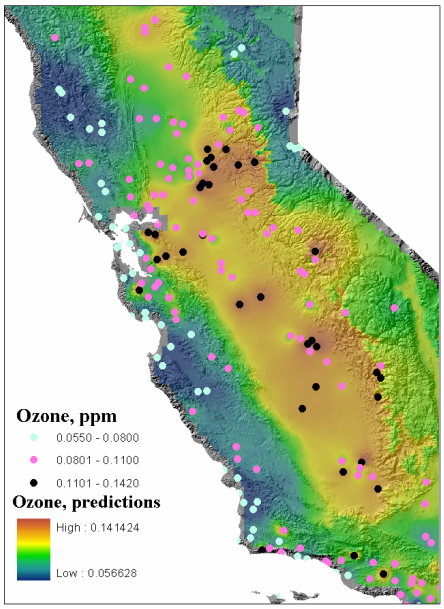

Figure 8. Visualization of the result of predictions using GIS.

For some variables of interest, such as air pollution, we need to factor in the mountainous barriers to movement. For other variables such as ocean temperature and fish abundance, we need to take into account the shape of the coastline. Another example is provided by halibut abundance near the western Canada coastline (Figure 9. Data provided by the International Pacific Halibut Commission). Here the small islands are barriers to aquatic movement. Thus, we would want to use a distance metric that reflects how the fish swim.

Traditional analysis of geostatistical and spatial point processes are based upon straight-line distance. In practice, environmental and artificial barriers due to rivers, roads, soil types variability, and other natural boundaries always exist and it is necessary to account for them to create a meaningful analysis (e.g., as part of a process model approach to spatial analysis as described in Laffan, 2002). For example, weights to the neighbors in lattice data analysis can be naturally based on travel or economic distance between

objects in addition to the length of the common border and distance between polygon centroids.

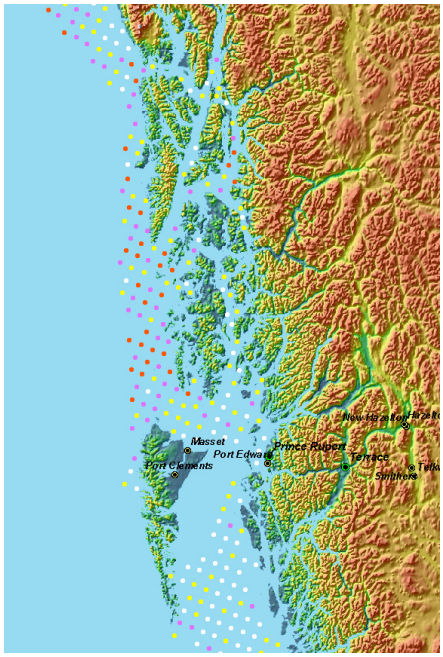

Figure 9. Halibut abundance near the western Canada coastline. Circles are observational locations.

Figure 10 displays crime events data over a street network over a six-month period in a medium sized city. Crime tends to be close to the road and the places where people live or work. Analyzing the spatial distribution of crime dictates the usage of a more sophisticated distance metric than a simple straight-line distance metric.

The next version of the Geostatistical Analyst software will use a general and flexible approach to the problem of using non-Euclidean distance metrics in spatial data analysis. This approach is based on a cost weighted distance, a common raster function in GIS that calculates the cost of travel from one cell of a grid to the next. Specifying high costs for travel between certain cells effectively prevents movement between these cells. The determination of the cost value at each location, calculation of distances between sampled locations and unsampled ones, and choice of covariance function in the case of geostatistics is discussed in Krivoruchko and Gribov, 2002.

Measurement and locational errors

Two main sources of error are data collection and data analysis. Errors propagate as a result of data manipulations: the errors in maps are modified and usually lead to increased output map uncertainty and may lead to wrong conclusions. Even in a well-designed experiment, errors often arise from imperfection in the experimental setup and the researcher's inconsistencies. For most experiments with

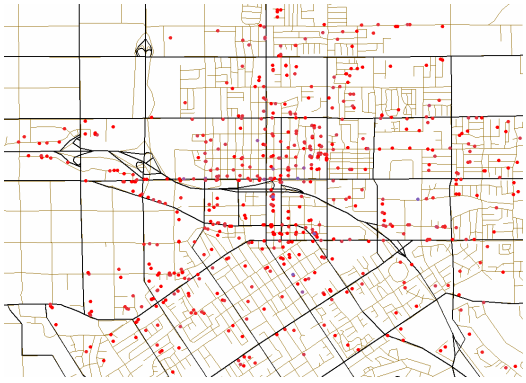

Figure 10. Example of crime events displayed over street network.

measured outcomes, if an experiment is repeated many times, measurement will be slightly different each time. Measurement errors can be attributed to error in the measurement device, human recording error, changes in the measurement conditions, data integration, and faulty sampling techniques. Therefore, the inevitable errors of measurement are something science has to live with and the true signal we are interested in has to be extracted from the noisy data. Such errors also arise in determining locations. Examples of locational errors include measurements collected for territories (polygons) for which the area of which is unknown; use of centroids to measure the location of a polygon; truncated coordinates; and coordinates distorted by map projection.

Measurement error in geostatistics theory was developed in the form of the filtered kriging from the very beginning (Gandin, 1963). Locational errors were discussed from time to time, but detailed theory has appeared only recently (Cressie and Kornak, 2002). The Geostatistical Analyst now allows filtered kriging and future releases will allow for adjustment of locational errors.

Consider an example of lightning strikes in Boulder County, Colorado, on September 1, 2000 (Figure 11). The red polygons reflect a radius of uncertainty in the location of the strikes. In addition to the approximate data location, there is information on the polarity and strength of the lightning strikes. Positive polarity is more likely to ignite wildfire. We could use the Geostatistical Analyst to create a probability map of positive polarity lightning strikes and use this as an indication of the risk of a wildfire. However, this does not account for the spatial distribution in the lightning strikes: there may be far more strikes in other areas, so even if they have negative polarity,

there might be increased fire risk simply due to the number of strikes.

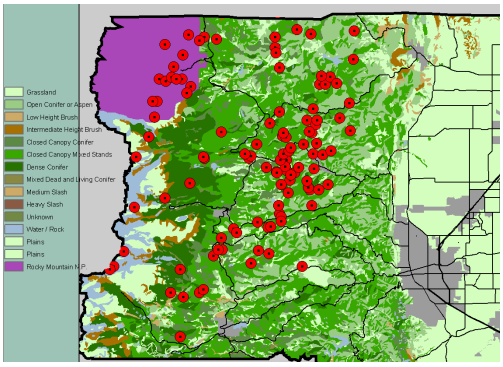

Figure 11. Lightning strikes in Boulder County, Colorado, on September 1, 2000

From a spatial statistics viewpoint, these data would be considered as a marked point pattern process, a process that is doubly stochastic: one process locates the lightning strikes (a point process) and another process controls the strength and polarity associated with strikes at the recorded locations (random field process).

To account for errors in identifying the locations of the lightening strikes, we could define much larger polygons that partition the data according to soil and forest types and then count the number of lightning strikes and estimate polarity distribution in the potentially overlapping polygons. Any polygonal data analysis on such aggregated data would require a sophisticated distance metric.

This data discussion raises several questions. First, we need a marked point pattern model with locational error. Second, we need software for geostatistical, point and marked point pattern analysis with an option to use non-Euclidean distances. Software that allows flexible definition of weights for lattice data modeling, including weights defined using cost surface, is required. After all, real datasets are often a combination of continuous, polygonal, and point data, and practitioners will appreciate tools for interacting between different types of data.

To the best of our knowledge, commercial marked point pattern analysis software simply does not exist. Although several implementations of lattice data analysis are available on the market, a flexible set of visualization tools for polygonal data analysis is also not available.

There are many other important additions to the spatial data analysis arsenal, including aggregation, disaggregation, and integration of spatial data obtained at different spatial scales (Krivoruchko and Gotway, 2002).

5. CONCLUSION

Although the requirements of GIS users are varied, they all have two needs in common: interactive visualization and the ability to use statistical methods and models for spatial data. Many users may not be knowledgeable about statistical theory, so they will appreciate a software that helps them make good choices for their data and application. Others more knowledegable in statistics will appreciate the visualization tools in GIS that are lacking in statistical software packages. They will also appreciate the ability to choose to among methods and models that comes with a comprehensive and flexibile set of tools for interactive spatial data analysis.

We think that it is much easier to incorporate modern spatial statistics into GIS environment than it is to implement modern GIS functionality in statistical software. At the same time, a user friendly and understandable implementation of statistical models into the GIS core is the most efficient way to involve more people in inferential spatial data analysis.

6. REFERENCES

Bailey, T. C. and Gatrell, A. C. 1995. Interactive Spatial Data Analysis. Essex: Addison Wesley Longman Limited.

Cressie, N. and Kornak, J. 2002. Spatial statistics in the presence of location error with an application to remote sensing of the environment. Department of Statistics Preprint No. 701, The Ohio State University. Available by request at http://www.stat.ohio-state.edu/~sses/papers.html.

Gandin, L.S. 1963. Objective Analysis of Meteorological Fields. Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad (translated by Israel Program for Scientific Translations, Jerusalem, 1965).

Krivoruchko K. and Gotway C.A. 2002. Expanding the "S" in GIS: Incorporating Spatial Statistics in GIS. Available from ESRI online at http://www.esri.com/software/arcgis/arcgisxtensions/geostatistical/research_papers.html.

Krivoruchko K. and Gribov A. 2002. Geostatistical Interpolation in The Presence of Barriers. Available from ESRI online at http://www.esri.com/software/arcgis/arcgisxtensions/geostatistical/research_papers.html.

Laffan, S.W. 2002. Using process models to improve spatial analysis. *International Journal of Geographic Information Science*, 16(3): 245-257.