Handbook of Applied Spatial Analysis

Software Tools, Methods and Applications

Bearbeitet von
Manfred M Fischer, Arthur Getis

A.1 Spatial Statistics in ArcGIS

Lauren M. Scott and *Mark V. Janikas*

A.1.1 Introduction

With over a million software users worldwide, and installations at over 5,000 universities, Environmental Systems Research Institute, Inc. (ESRI), established in 1969, is a world leader for the design and development of Geographic Information Systems (GIS) software. GIS technology allows the organization, manipulation, analysis, and visualization of spatial data, often uncovering relationships, patterns, and trends. It is an important tool for urban planning (Maantay and Ziegler 2006), public health (Cromley and McLafferty 2002), law enforcement (Chainey and Ratcliffe 2005), ecology (Johnston 1998), transportation (Thill 2000), demographics (Peters and MacDonald 2004), resource management (Pettit et al. 2008), and many other industries (see http://www.esri.com/industries.html). Traditional GIS analysis techniques include spatial queries, map overlay, buffer analysis, interpolation, and proximity calculations (Mitchell 1999). Along with basic cartographic and data management tools, these analytical techniques have long been a foundation for geographic information software. Tools to perform spatial analysis have been extended over the years to include geostatistical techniques (Smith et al. 2006), raster analysis (Tomlin 1990), analytical methods for business (Pick 2008), 3D analysis (Abdul-Rahman et al. 2006), network analytics (Okabe et al. 2006), space-time dynamics (Peuquet 2002), and techniques specific to a variety of industries (e.g., Miller and Shaw 2001). In 2004, a new set of spatial statistics tools designed to describe feature patterns was added to ArcGIS 9. This chapter focuses on the methods and models found in the Spatial Statistics toolbox.

Spatial statistics comprises a set of techniques for describing and modeling spatial data. In many ways they extend what the mind and eyes do, intuitively, to assess spatial patterns, distributions, trends, processes and relationships. Unlike traditional (non-spatial) statistical techniques, *spatial* statistical techniques actually use space – area, length, proximity, orientation, or spatial relationships – directly in their mathematics (Scott and Getis 2008).
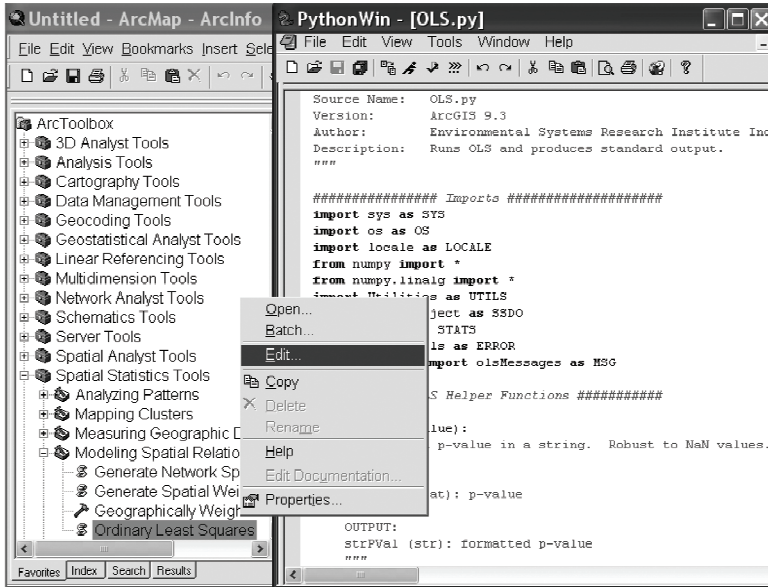
Fig. A.1.1. Right click on a script tool and select *Edit* to see the Python source code

By 2008 the Spatial Statistics toolbox in ArcGIS contained 25 tools. The majority of these were written using the Python scripting language. Consequently, Arc-GIS users have access not only to the analytical methods for these tools, but also to their source code (see Fig. A.1.1).

The Spatial Statistics toolbox includes both statistical functions and general-purpose utilities. With the most recent release of ArcGIS 9.3, statistical functions are grouped into four toolsets: Measuring Geographic Distributions, Analyzing Patterns, Mapping Clusters, and Modeling Spatial Relationships.

A.1.2   Measuring geographic distributions

The tools in the Measuring Geographic Distributions toolset (Table A.1.1) are descriptive in nature; they help summarize the salient characteristics of a spatial distribution. They are useful for answering questions like:

- Which site is most accessible?
- Is there a directional trend to the spatial distribution of the disease outbreak?
- What is the primary wind direction for this region in the winter?
- Where is the population center?
- Which species has the broadest territory?

**Table A.1.1.** Tools in the measuring geographic distributions toolset

Tool	Description
Central feature	Identifies the most centrally located feature in a point, line, or polygon feature class
Directional distribution (standard deviational ellipse)	Measures how concentrated features are around the geographic mean, and whether or not they exhibit a directional trend
Linear directional mean	Identifies the general (mean) direction and mean length for a set of vectors
Mean center	Identifies the geographic center for a set of features
Standard distance	Measures the degree to which features are concentrated or dispersed around the geographic mean center

Even the simplest tool in the Spatial Statistics toolbox can be a powerful communicator of spatial pattern when used with animation. The mean center tool is a measure of central tendency; it computes the geometric center – the average X and average Y coordinate – for a set of geographic features. In Fig. A.1.2, the weighted mean center of population for the counties of California is computed every decade from 1910 to 2000. The center of population is initially located in the northern half of the state near San Francisco. Animation reveals steady movement of the mean center south, every decade, as population growth in Southern California outpaces population growth in the state's northern counties.
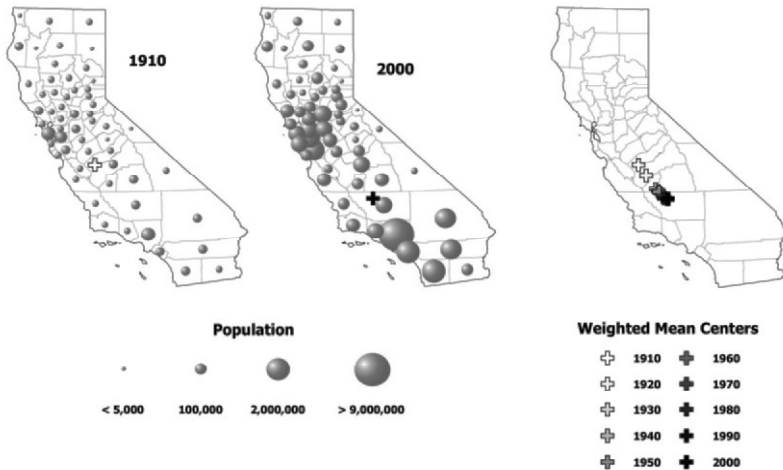
Fig. A.1.2. Weighted mean center of population, by county, 1910 through 2000
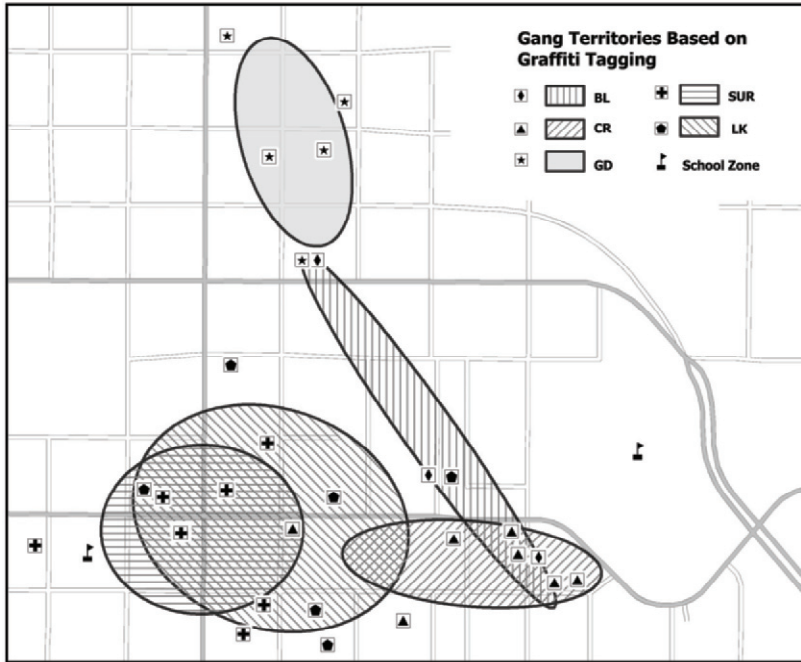
Fig. A.1.3. Core areas for five gangs based on graffiti tagging

The Standard Deviational Ellipse and Standard Distance tools measure the spatial distribution of geographic features around their geometric center, and provide information about feature dispersion and orientation. Gangs often mark their territory with graffiti. In Fig. A.1.3, a standard deviational ellipse is computed, by gang affiliation, for graffiti incidents in a city. The ellipses provide an estimate of the core areas associated with each gang's turf. The potential for increased gang-related conflict and violence is highest in areas where the ellipses overlap. By increasing the presence of uniformed police officers in these overlapping areas and around nearby schools, the community may be able to curtail gang violence. Mitchell (2005), Scott and Warmerdam (2005), Wong (1999) and Levine (1996) provide additional examples of applications for descriptive statistics like mean center, standard distance, and standard deviational ellipse.

A.1.3   Analyzing patterns

The Analyzing Patterns toolset (Table A.1.2) contains methods that are most appropriate for understanding broad spatial patterns and trends (Mitchell 2005). With these tools you can answer questions like:

- Which plant species is most concentrated?
- Does the spatial pattern of the disease mirror the spatial pattern of the population at risk?
- Is there an unexpected spike in pharmaceutical purchases?
- Are new AIDs cases remaining geographically fixed?

Consider the difficulty of trying to measure changes in urban manufacturing patterns for the United States over the past few decades. Certainly broad changes have occurred with globalization and the move from vertical integration to a more flexible and dispersed pattern of production. One approach might be to map manufacturing employment by census tract for a series of years, and then try to visually discern whether or not spatial patterns are becoming more concentrated or more dispersed. Most likely a range of scenarios would emerge. The Global Moran's *I* tool computes a single summary value, a *z*-score, describing the degree of spatial concentration or dispersion for the measured variable (in this case manufacturing employment). Comparing this summary value, year by year, indicates whether or not manufacturing is becoming, overall, more dispersed or more concentrated.

Similarly, viewing thematic maps of per capita incomes (PCR)[1] in New York for a series of years (see Fig. A.1.4), it is difficult to determine whether rich and poor counties are becoming more or less spatially segregated. Plotting the resultant *z-s*cores from the Spatial Autocorrelation (Global Moran's *I*) tool, however, reveals decreasing values indicating that spatial clustering of rich and poor has dissipated between 1969 and 2002.

Table A.1.2. A summary of the tools in the analyzing patterns toolset

Tool	Description
Average nearest neighbor	Calculates the average distance from every feature to its nearest neighbor based on feature centroids
High/low clustering (Getis-Ord general *G*)	Measures concentrations of high or low values for a study area
Spatial autocorrelation (global Moran's *I*)	Measures spatial autocorrelation (clustering or dispersion) based on feature locations and attribute values
Multi-distance spatial cluster analysis (Ripley's *K* function)	Assesses spatial clustering/dispersion for a set of geographic features over a range of distances

The *K* function is a unique tool in that it looks at the spatial clustering or dispersion of points/features at a series of distances or spatial scales. The output from the *K* function is a line graph (see Fig. A.1.5). The dark diagonal line represents the expected pattern, if the features were randomly distributed within the study area. The *X* axis reflects increasing distances. The solid curved line represents the

---

[1] PCR is per capita income relative to the national average.

observed spatial pattern for the features being analyzed. When the curved line goes above the diagonal line, the pattern is more clustered at that distance than we would expect with a random pattern; when the curved line goes below the diagonal line, the pattern is more dispersed than expected. Based on a user-specified number of randomly generated permutations of the input features, the tool also computes a confidence envelope around the expected line. When the curved line is outside the confidence envelope, the clustering or dispersion is statistically significant.
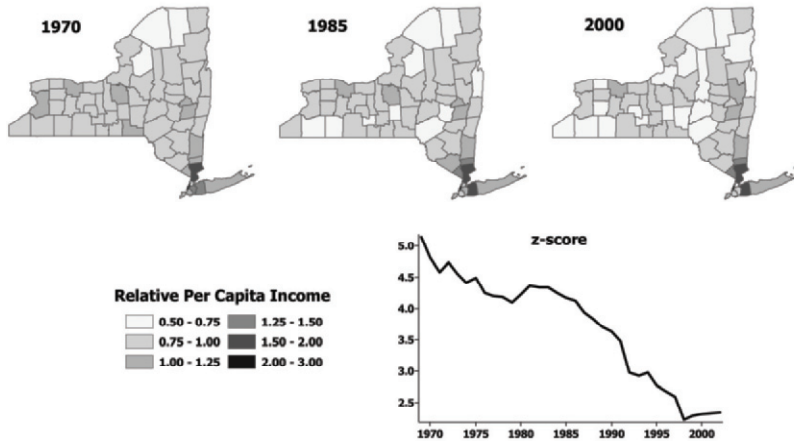

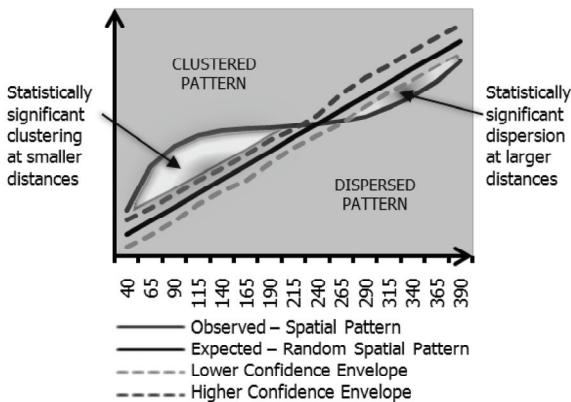**Fig. A.1.4.** Relative per capita income for New York, 1969 to 2002

Fig. A.1.5. Components of the *K* function graphical output

The *K* function is useful for comparing different sets of features within the same study area, such as two strains of a disease or disease cases in relation to population at risk. Similar observed spatial patterns suggest similar factors (similar spatial processes) are at work. A researcher might compare the spatial pattern for a disease outbreak, for example, to the spatial pattern of the population at risk to help determine if factors other than the spatial distribution of population are promoting disease incidents. Wheeler (2007), Levine (1996), Getis and Ord (1992), and Illian et al. (2008) provide examples of additional applications for the tools in the Analyzing Patterns toolset.

A.1.4 Mapping clusters

The tools discussed above in the Analyzing Patterns toolset are global statistics that answer the question: *Is* there statistically significant spatial clustering or dispersion? Tools in the Mapping Clusters toolset (Table A.1.3), on the other hand, identify *where* spatial clustering occurs, and *where* spatial outliers are located:

- Where are their sharp boundaries between affluence and poverty in Ecuador?
- Where do we find anomalous spending patterns in Los Angeles?
- Where do we see unexpectedly high rates of diabetes?

In Fig. A.1.6, the Local Moran's *I* tool is used to analyze poverty in Ecuador. A string of outliers separate clusters of high poverty in the north from clusters of low poverty in the south, indicating a sharp divide in economic status.

Table A.1.3. A summary of the tools in the mapping clusters toolset

Tool	Description
Cluster and outlier analysis (Anselin's local Moran's *I*)	Given a set of weighted features, identifies clusters of high or low values as well as spatial outliers
Hot spot analysis (Getis-Ord G_i^*)	Given a set of weighted features, identifies clusters of features with high values (hot spots) and clusters of features with low values (cold spots)

The Hot Spot Analysis (Getis-Ord G_i^*) tool is applied to vandalism data for Lincoln, Nebraska in Fig. A.1.7. In the first map (left), raw vandalism counts for each census block are analyzed. The picture that emerges would not surprise local police officers. Most vandalism is found where most people and most overall crime are found: downtown and in surrounding high crime areas. Fewer cases of vandalism are associated with the lower density suburbs. In the second map (right), however, vandalism is normalized by overall crime incidents prior to analysis. Running the Hot Spot Analysis tool on this normalized data shows that
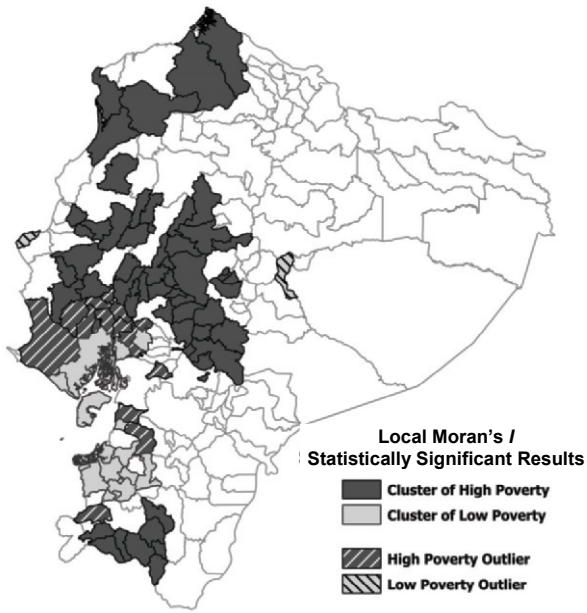
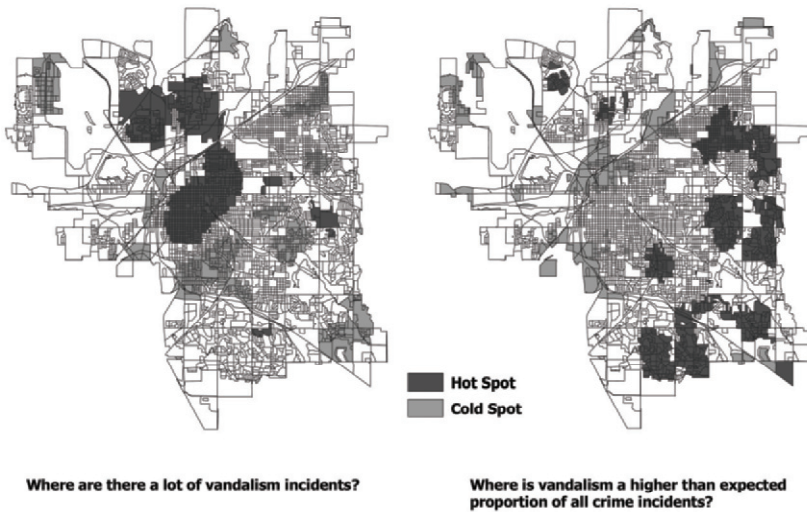**Fig. A.1.6.** An analysis of poverty in Ecuador using local Moran's *I*

Fig. A.1.7. An analysis of vandalism hot spots in Lincoln, Nebraska using G_i^*

while Lincoln may have more incidents of vandalism in downtown areas, vandalism represents a larger proportion of total crime in suburban areas. Zhang et al. (2008), Jacquez and Greiling (2003), Getis and Ord (1992), Ord and Getis (1995), and Anselin (1995) provide additional applications for the tools in the Mapping Clusters toolset.

A.1.5 Modeling spatial relationships

The tools in the Modeling Spatial Relationships toolset (Table A.1.4) fall into two categories. The first category includes tools designed to help the user define a conceptual model of spatial relationships. The conceptual model is an integral component of spatial modeling and should be selected so that it best represents the structure of spatial dependence among the features being analyzed (Getis and Aldstadt 2004).

The options available for modeling spatial relationships include inverse distance, fixed distance, polygon contiguity (Rook's and Queen's case), k nearest neighbors, Delaunay triangulation, travel time and travel distance. Figure A.1.8 illustrates how spatial relationships change when they are based on a real road network, rather than on straight line distances.
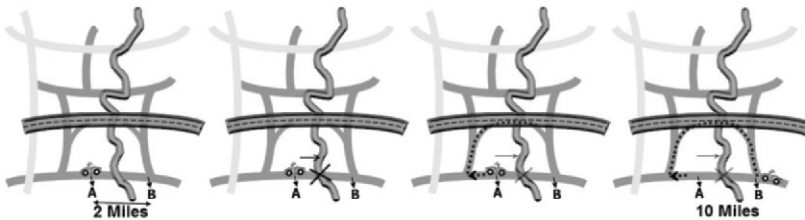
Fig. A.1.8. Traffic conditions or a barrier in the physical landscape can dramatically change actual travel distances, impacting results of spatial analysis

Table A.1.4. A summary of the tools in the modeling spatial relationships toolset

Tool	Description
Generate network spatial weights	Builds a spatial weights matrix file specifying spatial relationships among features in a feature class based on a Network dataset
Generate spatial weights matrix	Builds a spatial weights matrix file specifying spatial relationships among features in a feature class
Geographically weighted regression	A local form of linear regression used to model spatially varying relationships among a set of data variables
Ordinary least squares regression	Performs global linear regression to model the relationships among a set of data variables

Constructing spatial relationships prior to analysis generally results in improved performance, particularly within the context of larger datasets or when applied to multiple attribute fields. The spatial weights matrix files (*.swm) are sharable, reusable and can be directly edited within ArcGIS. Furthermore, options are available to facilitate both importing and exporting spatial weights matrix files from/to other formats (*.gal, *.gwt, or a simple *.dbf table).[2]

The second category of tools in the Modeling Spatial Relationships toolset includes ordinary least squares (OLS) (Woolridge 2003), and geographically weighted regression (GWR) (Fortheringham et al. 2002 and Chapter C.5). These tools can help answer the following types of questions:

- What is the relationship between educational attainment and income?
- Is there a relationship between income and public transportation usage? Is that relationship consistent across the study area?
- What are the key factors contributing to excessive residential water usage?

Regression analysis may be used to model, examine, and explore spatial relationships, in order to better understand the factors behind observed spatial patterns or to predict spatial outcomes. There are a large number of applications for these techniques (Table A.1.5).

$$\beta_0 \quad + \quad \beta_1 \text{ Population} \quad + \quad \beta_2 \text{ Income} \quad = \quad \text{Crime}$$
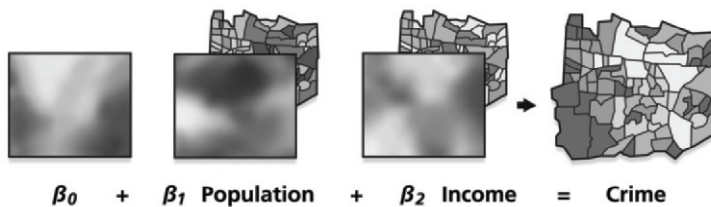
Fig. A.1.9. GWR optionally creates a coefficient surface for each model explanatory variable reflecting variation in modeled relationship

OLS is a global model. It creates a single equation to represent the relationship between what you are trying to model and each of your explanatory variables. Global models, like OLS, are based on the assumption that relationships are static and consistent across the entire study area. When they are not – when the relationships behave differently in separate parts of the study area – the global model becomes less effective. You might find, for example, that people's desire to live and work close, but not too close, to a metro line encourages population growth: the relationship for being fairly close to a metro line is positive while the relation-

---

[2] See http://resources.esri.com/geoprocessing/ for a description and examples of exporting /importing *.swm files to *.gal and *.gwt formats.

ship for being right up next to a metro line is negative. A global model will compute a single coefficient to represent both of these divergent relationships. The result, an average, may not represent either situation very well.

Local models, like GWR, create an equation for every feature in the dataset, calibrating each one using the target feature and its neighbors. Nearby features have a higher weight in the calibration than features that are farther away. What this means is that the relationships you are trying to model are allowed to change over the study area; this variation is reflected in the coefficient surfaces optionally created by the GWR tool (see Fig. A.1.9). If you are trying to predict foreclosures, for example, you might find that an income variable is very important in the northern part of your study area, but very weak or not important at all in the southern part of your study area. GWR accommodates this kind of regional variation in the regression model.

Table A.1.5. A variety of potential applications for regression analysis

Application Area	Analysis Example
Public health	Why are diabetes rates exceptionally high in particular regions of the United States?
Public safety	What environmental factors are associated with an increase in search and rescue event severity?
Transportation	What demographic characteristics contribute to high rates of public transportation usage?
Education	Why are literacy rates so low in particular regions?
Market analysis	What is the predicted annual sales for a proposed store?
Economics	Why do some communities have so many home foreclosures?
Natural resource management	What are the key variables promoting high forest fire frequency?
Ecology	Which environments should be protected to encourage reintroduction of an endangered species?

The default output for both regression tools is a residual map showing the model over- and underpredictions (see Fig. A.1.10). The OLS tool automatically checks for muliticollinearity (redundancy among model explanatory variables), and computes coefficient probabilities, standard errors, and overall model significance indices that are robust to heteroscedasticity. The online help documentation for these tools provides a beginner's guide to regression analysis, suggested step by step instructions for the model building process, a table outlining and carefully explaining the challenges and potential pitfalls associated with using regression analysis with spatial data, and recommendations for how to overcome those potential problems.[3]

---

[3]   See http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Regression_analysis_basics
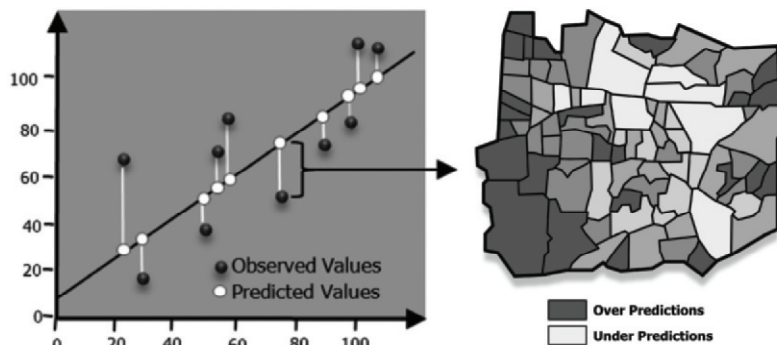
Fig. A.1.10. Default output from the regression tools is a map of model over- and under-predictions

A.1.6   Custom tool development

The tools in the Spatial Statistics toolbox were developed using the same methods and techniques that an ArcGIS user might adopt to create his/her own custom tools. They illustrate the extendibility of ArcGIS, and ESRI's commitment to providing a framework for custom tool development.

The simplest way to create a new tool in the geoprocessing framework is to use Model Builder to string existing tools together. The resultant model tool can then be exported to Python and further extended with custom code. In addition, any third party software package that can be launched from the DOS command line is an excellent custom tool candidate. Simply point to the executable for that software and define the needed tool parameters.

For software developers, the geoprocessing framework offers sophisticated options for custom tool development. Python script tools can be run 'in process', resulting in a cohesive interface that improves both performance and usability. Numerical Python (NumPy) provides an avenue to perform complex mathematical operations (Oliphant 2006), and is currently part of the ArcGIS software installation. Other Python libraries can be added as well. Perhaps the most logical extension is Scientific Python (SciPy),[4] which provides a host of powerful statistical techniques and works directly with NumPy. PySAL (a Python Library for Spatial Analytical Functions, see Chapter A.10), developed in conjunction with GeoDa (see Chapter A.4) and STARS (see Chapter A.5), is a crossplatform library of spatial analysis functions that may also provide opportunities for extending Arc GIS functionality.[5]

---

[4] http://www.scipy.org/

[5] See http://www.sal.uiuc.edu/tools/tools-sum/pysal and http://www.sal.uiuc.edu/tools/tools-sum/pysal

Python works nicely with other programming languages, and this has resulted in several hybrid libraries including Rpy and PyMat, giving users access to the methods in R (see Chapter A.3 for spatial econometric functions in R) and in MatLab, respectively.[6] There are also a number of spatial data analysis add-on packages for R (Bivand and Gebhardt 2000) and a spatial econometrics toolbox for MatLab (LeSage 1999). Sample scripts demonstrating integration of ArcGIS 9.3 with R are available for download from the Geoprocessing Resource Center[7] (see Fig. A.1.11).
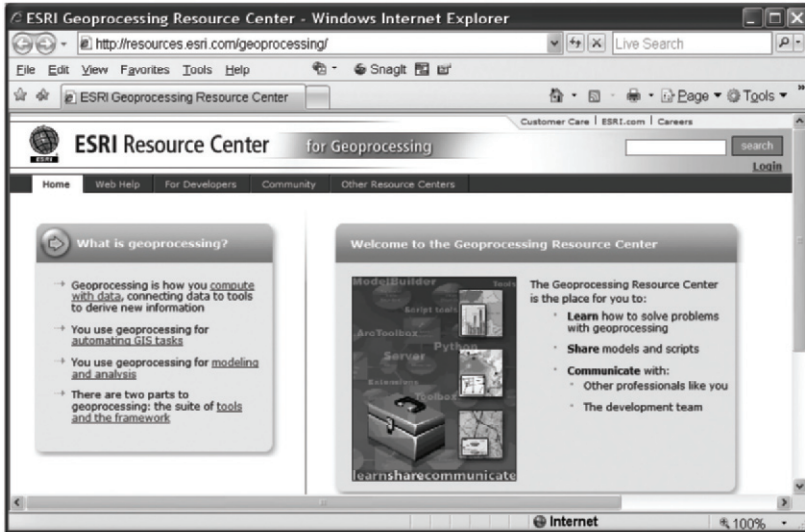

Fig. A.1.11. Geoprocessing Resource Center Web page

A.1.7   Concluding remarks

The Spatial Statistics toolbox provides feature pattern analysis and regression analysis capabilities inside ArcGIS where users can leverage, directly, all of its powerful database management and cartographic functionalities. The source code for these tools is provided inside a geoprocessing framework that encourages development and sharing of custom tools and methods. People and organizations developing custom Python tools can take advantage of existing libraries, documentation, sample scripts, and support from a worldwide community of Python software developers. The Geoprocessing Resouce Center (see Fig. A.1.11),

---

[6] See http://rpy.sourceforge.net/, http://www.r-project.org/, http://www.mathworks.com/, and http://claymore.engineer.gvsu.edu/~steriana/Python/pymat.html.

[7] http://resources.esri.com/geoprocessing/

launched in August of 2008, offers a platform for asking questions and getting answers, for sharing ideas, tools, and methodologies, and for participating in an ongoing conversation about spatial data analysis. The sincere hope is that this conversation will extend beyond the realm of academics, theoreticians, and software developers – that it will embrace the hundreds of thousands of GIS users grappling with real world data and problems – and that, as a consequence, this might foster new tools, new questions, perhaps even new approaches altogether.

References

Abdul-Rahman A, Zlantanaova S, Coors V (2006) Innovations in 3D geo information systems. Springer, Berlin, Heidelberg and New York

Anselin L (1995) Local indicators of spatial association: LISA. Geogr Anal 27(2):93-115

Anselin L, Syabri I, Kho Y (2009) GeoDa: an introduction to spatial data analysis. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.73-89

Bivand RS (2009) Spatial econometrics functions in R. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.53-71

Bivand RS, Gebhardt A (2000) Implementing functions for spatial statistical analysis using the R language. J Geogr Syst 2(3):307-317

Chainey SP, Ratcliffe JH (2005) GIS and crime mapping. Wiley, London

Cromley EK, McLafferty SL (2002) GIS and public health. Guilford, New York

Fotheringham SA, Brunsdon C, Charlton M (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, New York, Chichester, Toronto and Brisbane

Getis A, Aldstadt J (2004) Constructing the spatial weights matrix using a local statistic. Geogr Anal 36(2):90-104

Getis A, Ord JK (1992) The analysis of spatial association by use of distance statistics. Geogr Anal 24(3):189-206

Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical analysis and modeling of spatial point patterns. Wiley, London

Jacquez GM, Greiling DA (2003) Local clustering in breast, lung and colorectal cancer in Long Island, New York. Int J Health Geographics 2:3

Johnston CA (1998) Geographic information systems in ecology. Blackwell Science, Malden [MA]

LeSage JP (1999) Spatial econometrics using MATLAB. www.spatial-econometrics.com

Levine N (1996) Spatial statistics and GIS: software tools to quantify spatial patterns. J Am Plann Assoc 62(3):381-391

Maantay J, Ziegler J (2006) GIS for the urban environment. ESRI, Redlands [CA]

Miller HJ, Shaw S-L (2001) Geographic information systems for transportation: principles and applications. Oxford University Press, Oxford and New York

Mitchell A (1999) The ESRI guide to GIS analysis, volume 1: geographic patterns and relationships. ESRI, Redlands [CA]

Mitchell A (2005) The ESRI guide to GIS analysis, volume 2: spatial measurements and statistics. ESRI, Redlands [CA]

Okabe A, Okunuki K, Shiode S (2006) SANET: a toolbox for spatial analysis on a network. Geogr Anal 38(1):57-66

Oliphant T (2006) Guide to NumPy, Trelgol [USA]

Ord JK, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and an application. Geogr Anal 27(4):287-306

Peters A, MacDonald H (2004) Unlocking the census with GIS. ESRI, Redlands [CA]

Pettit C, Cartwright W, Bishop I, Lowell K, Pullar D, Duncan D (eds) (2008) Landscape analysis and visualization: spatial models for natural resource management and planning. Springer, Berlin, Heidelberg and New York

Pick JB (2008) Geo-Business: GIS in the digital organization. Wiley, New York

Peuquet DJ (2002) Representations of space and time. Guilford, New York

Rey SJ, Anselin L (2009) PySAL: a Python library of spatial analytical methods. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.175-193

Rey SJ, Janikas MV (2009) STARS: Space-time analysis of regional systems. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.91-112

Scott L, Getis A (2008) Spatial statistics. In Kemp K (ed) Encyclopedia of geographic informations. Sage, Thousand Oaks, CA

Scott L, Warmerdam N (2005) Extend crime analysis with ArcGIS spatial statistics tools. ArcUser Magazine, April-June [USA]

Smith MJ, Goodchild MF, Longley PA (2006) Geospatial analysis. Troubador, Leicester

Thill J-C (2000) Geographic information systems in transportation research. Elsevier Science, Oxford

Tomlin DC (1990) Geographic information systems and cartographic modeling. Prentice-Hall, New Jersey

Wheeler D (2007) A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003. Int J Health Geographics 6(1):13

Wheeler D, Paéz A (2009) Geographically Weighted Regression. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.461-486

Wong DWS (1999) Geostatistics as measures of spatial segregation. Urban Geogr 20(7):635-647

Woolridge JM (2003) Introductory econometrics: a modern approach. South-Western, Mason [OH]

Zhang C, Luo L, Xu W, Ledwith V (2008) Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. Sci Total Environ 398 (1-3):212-221