Exploratory Spatial Data Analysis

66

Jürgen Symanzik

Contents

Abstract

In this chapter, we discuss key concepts for exploratory spatial data analysis (ESDA). We start with its close relationship to exploratory data analysis (EDA) and introduce different types of spatial data. Then, we discuss how to explore spatial data via different types of maps and via linking and brushing. A key technique for ESDA is local indicators of spatial association (LISA). ESDA needs to be supported by software. We discuss two main lines of software developments: GIS-based solutions and stand-alone solutions.

J. Symanzik
Department of Mathematics and Statistics, Utah State University, Logan, UT, USA
e-mail: juergen.symanzik@usu.edu; symanzik@math.usu.edu

M.M. Fischer, P. Nijkamp (eds.), *Handbook of Regional Science*,
DOI 10.1007/978-3-642-23430-9_76, © Springer-Verlag Berlin Heidelberg 2014

66.1 Introduction

In his groundbreaking book from 1977 on exploratory data analysis (EDA), Tukey (1977) made several statements that are still relevant today, more than 35 years after the publication of this book:

- "The greatest value of a picture is when it *forces* us to notice what we never expected to see." (p. vi)
- "Today, exploratory and confirmatory can — and should — proceed side by side." (p. vii)
- "Exploratory data analysis is detective work — numerical detective work — or counting detective work — or graphical detective work." (p. 1)
- "Unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider." (p. 3)
- "Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone — as the first step." (p. 3)

Tukey's expectations (and limitations) on EDA can easily be extended to exploratory spatial data analysis (ESDA), that is, the exploratory analysis of data with a spatial (geographic) component. As early as 1981, Ripley (1981) followed the distinction between exploratory and confirmatory data analyses in the preface of his book on spatial statistics: "The techniques presented are designed for both of John Tukey's divisions of exploratory and confirmatory data analysis" (p. vi).

The exact definition of ESDA slightly differs from source to source, but all agree that ESDA is an obvious extension of EDA. Commonly found topics that are covered by ESDA include the visualization and exploration of data in a spatial (geographic) framework. ESDA utilizes many methods, tools, and software components from the field of interactive and dynamic statistical graphics, such as brushing and linked views/linked windows. Typically, one or more map views are linked with one or more statistical displays of the data. Modifications to one of the views will result in modifications of all linked views. Questions of interest that ESDA can answer may be whether a cluster of points that can be seen in a scatterplot is related to nearby spatial locations or whether a particular geographic region (say, the coastal region of a country) exhibits different characteristics than the mountainous region that can be seen in a linked statistical view. Moreover, ESDA can help to create new hypotheses about the underlying spatial data that can later be investigated in more detail in a follow-up study. Also, ESDA methods should be applied before any advanced modeling and testing of statistical hypotheses. Anscombe (1973) has provided some striking examples which could happen when a linear regression line is blindly fitted to some unsuitable data set. The same is the case if some methods from spatial statistics are blindly applied to some spatial data set when no prior exploration took place. We should keep in mind that spatial data often are large and diverse data sets, rather than homogeneous data sets. A large number of different methods usually could, and should, be used, including, simple numerical summary statistics. Coming back to Tukey, the goal or expected outcome of the exploration usually is unknown in advance. Moreover, it should be noted that ESDA is more than just an extension of EDA as additional techniques and methods are needed that incorporate the specific spatial structure of the data. A frequent goal of ESDA is the exploration of spatial autocorrelation. We speak of positive spatial autocorrelation when nearby observations on average are more similar than what a random assignment would yield and of negative spatial autocorrelation when nearby observations on average are more distinct than what a random assignment would yield.

In the next section, we will discuss the main types of spatial data and following that, consider basic visualization and exploration techniques via maps. We then discuss two of the key concepts of ESDA: exploration via linking and brushing and local indicators of spatial association (LISA). A section on software for ESDA follows. We then finish with a brief conclusion and outlook on possible future work.

66.2 Types of Spatial Data

There exist four main types of spatial data:

a. In spatial point patterns, the location of an event is of interest itself. Point patterns can be the locations where a patient died from a particular disease or where some specific animal species has been observed. A question of interest might be to explore the spatial patterns of the deaths, for example, at which locations deaths have been due to disease A and at which locations deaths have been due to disease B.
b. Lattice data, sometimes also called area, areal, or grid data, are data that have been aggregated over some small geographic area. Often, a distinction is made between regular lattices (such as encountered for remote sensing data) or irregular lattices (such as states, counties, or health service areas). In a scenario where different economic regions are compared, a question of interest might be to explore how variables such as educational level, age, and racial composition of the population relate to unemployment in that region.
c. Geostatistical data, sometimes also called spatially continuous data, are data that could, at least theoretically, be observed at any spatial location. However, cost and time determine at how many locations such data actually are collected. Examples of this type of data range from precipitation and temperature measurements to air pollution measurements and readings of minerals in the earth. A question of interest might be to visualize the distribution of nitrates in the soil in a specific region before fitting a smooth surface to the data.
d. Origin–destination flow data, sometimes also called link or spatial interaction data, are data that consist of measurements, each of which is associated with a pair of point locations or a pair of areas. Examples for this type of data are home address and workplace address for inhabitants of a particular city or originating and destination airports for airline travel. A question of interest might be to explore from which originating airports most passengers, most flights, or most cargo arrives at a particular destination airport.

Cressie (1993) addressed the first three types of spatial data, both from a theoretical as well as from an applied perspective. When an additional temporal component is available, that is, spatial data are collected over time; we speak of spatiotemporal data. Origin–destination flow data are discussed in detail in Fischer and Wang (2011, Part II). To a considerable extent, the underlying type of spatial data set determines which ESDA techniques are most suitable. Many of the ESDA techniques discussed in this chapter are suitable for more than one type of spatial data.

66.3 Basic Visualization and Exploration Techniques via Maps

The first step to explore spatial data often is to display the data on a map and then to produce several variations of the initial map. Credit needs to be given to John Snow (1813–1858), a British anesthesiologist, who was the first who mapped disease data. His investigation of the 1854 cholera outbreak in London pioneered the field of epidemiology. Nowadays, some consider him the *"father of epidemiology"* but the name *"grandfather of ESDA"* might suit him equally well. The 1854 London cholera outbreak started on August 19, 1854. It lasted about 6 weeks and resulted in more than 575 deaths. Snow (1936) observed: "... Mortality in this limited area probably equals any that was ever caused in this country, even by the plague." Snow's hypotheses were that cholera was transmitted from person to person via a fecal–oral route and that the drinking water of the Broad Street pump was the cause of the cholera outbreak. Snow utilized his map and empirical evidence to convince the Board of Guardians to remove the handle of the Broad Street pump. A mere 48 fatal attacks occurred, following the removal of the handle of the Broad Street pump, indicative that the water feeding the Broad Street pump could indeed be the source of the cholera epidemic.

As demonstrated by Snow, the visualization of spatial locations, that is, spatial point patterns, can provide valuable insights into such a data set. Moreover, if additional information is available for the locations such as age, gender, and case/control, this information can be displayed via different colors, symbols, and symbol sizes in the map display.

66.3.1 Choropleth Maps

For lattice data and geospatial data, several types of map displays exist and can be used for exploration. Best known, and most widely used, are choropleth maps. However, choropleth maps highly depend on choices made by the map creator. Even if the geographic boundaries are fixed as is the case for lattice data, Monmonier (1996, Chaps. 4 & 10) worked out different visual effects depending on whether the data are split into equal–interval classes or into quartile (or other quantile) classes. The same choices that affect histograms, that is, the starting point of a class interval and the width of each class interval, also affect choropleth maps.

Moreover, color choices in choropleth maps have a considerable effect on our perception. A small dark area in an overall bright map may (or may not) be perceived as well as a small white area in an overall dark map. Excellent options for color choices for maps and statistical plots can be obtained from the *ColorBrewer* software tool (Harrower and Brewer 2003), accessible at http://colorbrewer2.org.

Finally, if geographic boundaries are not fixed in advance, choropleth maps can be easily affected by the modifiable areal unit problem (MAUP) (Openshaw 1984). Depending on the boundaries that are used for the aggregation (such as summation or averaging), rather different results for the sums, percentages, or averages may be obtained. Monmonier (1996, Chap. 10) demonstrated the MAUP for the locations of Snow's cholera data set. Therefore, it is necessary to explore what happens when spatial data get aggregated in different ways.

Given these different sources for biases when looking at choropleth maps, it is necessary to create and explore a variety of these maps to explore and understand the underlying spatial patterns. A single choropleth map that is the result of some default setting in a map-producing software package rarely will reveal all details of the underlying spatial data set. Andrienko et al. (2001) discussed how to conduct an exploratory analysis of spatial data via a combination of interactive maps and data mining.

66.3.2 Linked Micromap Plots

Linked micromap (LM) plots (Symanzik and Carr 2008; Carr and Pickle 2010) were introduced as an alternative to choropleth maps, especially to overcome some of the limitations of choropleth maps. The basic idea behind LM plots is to link geographic region names and their statistical values with their locations that are shown in a sequence of small maps, called micromaps. A typical LM plot (see Fig. 66.1) consists of three to five columns. The first column usually shows the maps, the second column lists some identifier (such as country or state names), and the third to the fifth columns contain statistical plots. Each small map highlights a few locations, typically five in a single map. The data are sorted according to some statistical criteria, for example, from highest to lowest (or vice versa), or from highest increase to lowest increase between years 1 and 2. Thus, the topmost map shows locations with the five largest (or smallest) observations according to the sorting criteria, the next map shows the five locations with the next largest (or smallest) observations, and so on. In case of any spatial association, locations with high or low observations tend to be plotted on the same map or on neighboring maps. The columns with the statistical plots may contain dot plots for each location, confidence intervals, time series plots, or box plots that are based on data for each particular location.

Micromaps have been used in print for applications as diverse as for comparisons of changing population density and population growth by state and for the visualization and interpretation of birth defects data in Utah and the United States. Typically, a published micromap is the result of many iterations where the authors experimented with different sortings and arrangements of the data panels and multiple possible layouts of the map panel.

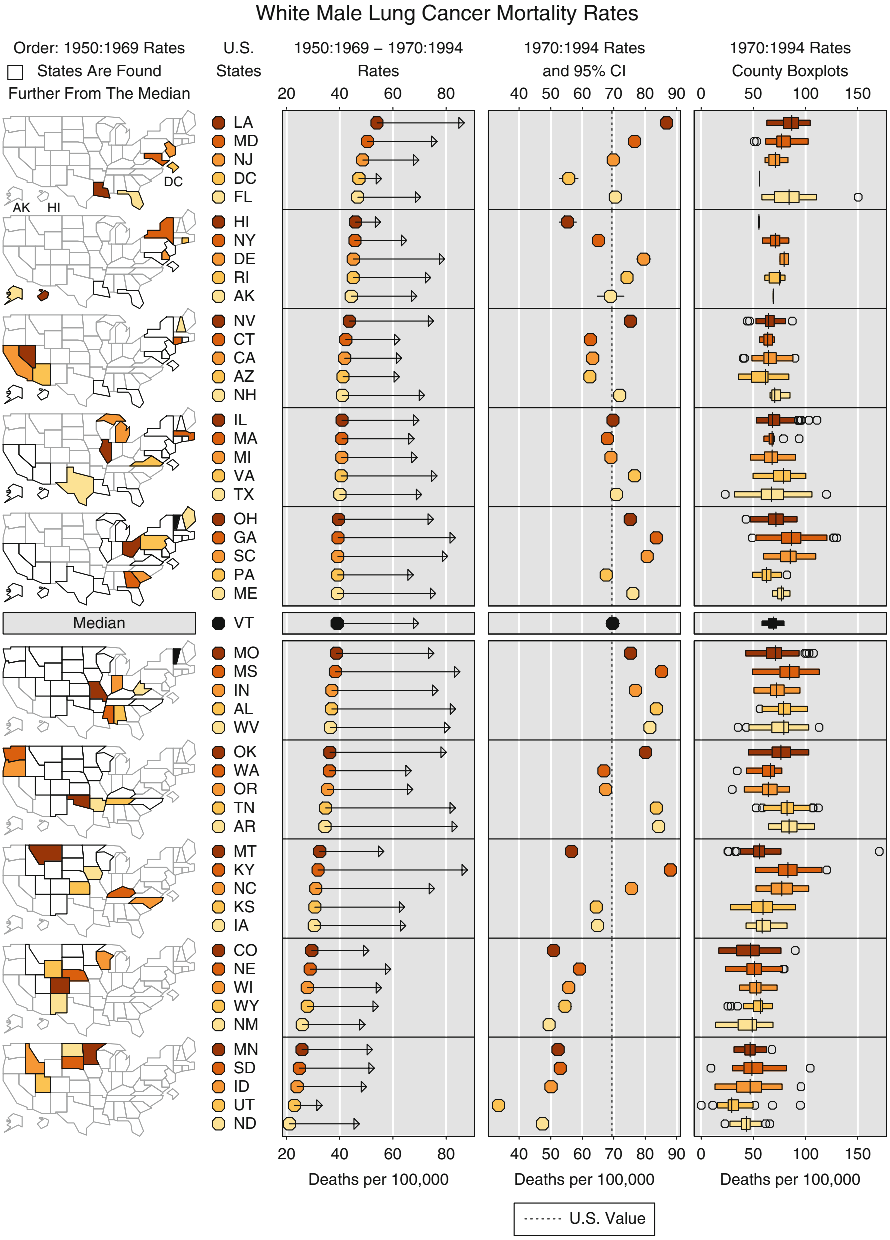


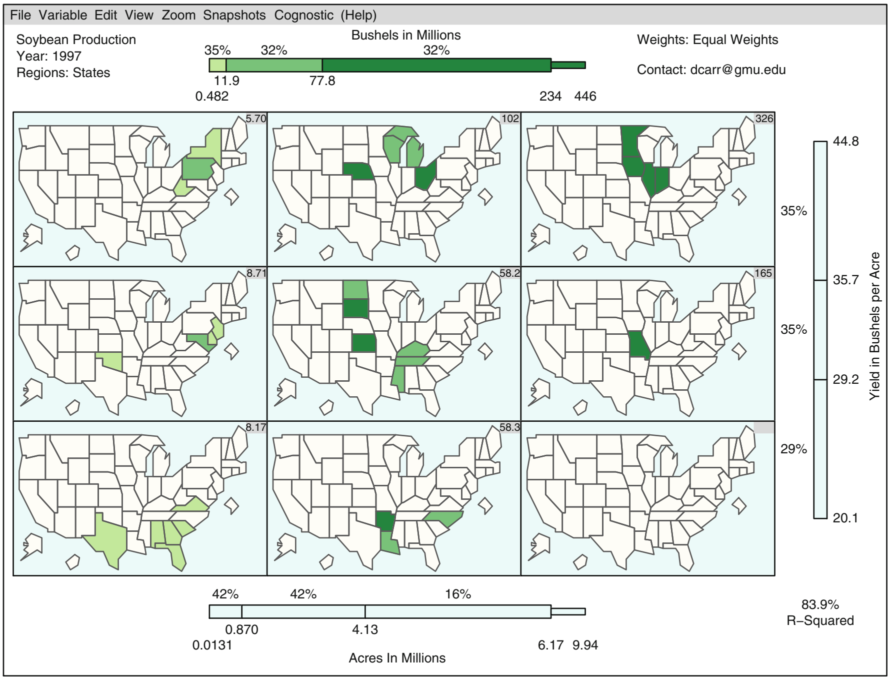
Fig. 66.1 LM plots, based on data from the NCI Web page, showing summary values for white male lung cancer mortality rates in the United States for the years 1950–1969 and for the years 1970–1994 in the *left data panel*, rates and 95 % confidence intervals in the *middle data panel*, and box plots for each of the counties of each state in the *right data panel* (Previously published as Fig. 1.6 in Symanzik and Carr (2008, p. 285))

While LM plots initially were constructed only for a static representation of the underlying data on paper or a computer screen, interactive versions may be introduced to allow an exploration of the underlying data from multiple perspectives. The US Department of Agriculture (USDA) – National Agricultural Statistics Service (NASS) Research and Development Division released an interactive micromap Web site (http://www.nass.usda.gov/research/sumpant.htm) in September 1999 for the display of data from the 1997 Census of Agriculture. This Web site still is accessible today. The National Cancer Institute (NCI) released an interactive micromap Web site (http://www.statecancerprofiles.cancer.gov/micromaps) in April 2003 for accessing their cancer data (Wang et al. 2002; Carr et al. 2002). This Web site is still accessible today and it is permanently updated with new data. While printed (static) LM plots are most suitable when the number of geographic regions ranges from about 10 to about 100, interactive LM plots may be suitable for several hundred geographic regions. Micromaps at the county level for the 254 counties of Texas at the NCI micromap Web site can reveal some very strong patterns, based on the data selection.

Figure 66.1 shows a static LM plot with three statistical columns based on data derived from the NCI Web site. The rows in the figure are sorted according to the 1950–1969 white male lung cancer rates in the United States (US) that reveal some strong geographic pattern with high rates in the eastern, southern, and western United States. A next step could be to resort the rows with respect to highest 1970–1994 rates, then with respect to highest absolute increases from the 1950–1969 to the 1970–1994 rates, and finally with respect to highest relative increases from the 1950–1969 to the 1970–1994 rates. Moreover, the second and third data column might be used to display data from possible confounding variables at the state level, such as smoking rates, gender composition, or educational level. After the exploration of several such LM plots, a researcher likely will have observed many known facts about the spatial distribution of male lung cancer, but, hopefully, some unexpected patterns and relationships also will have emerged.

66.3.3 Conditioned Choropleth Maps

Conditioned choropleth maps (CCmaps) (Carr et al. 2000; Carr and Pickle 2010) were introduced as a tool for the exploration of spatial data that consist of geographic locations, one dependent variable, and two independent variables. Via sliders, a researcher can interactively partition each of the two independent variables and the dependent variable into three different intervals each. A 3 × 3 set of panels containing nine partial maps shows the color-coded level (high, medium, low) of the dependent variable for those geographic locations that relate to high values of variable one and high values of variable two in map one, for those geographic locations that relate to high values of variable one and medium values of variable two in map two, and so on. For example, in an agricultural setting, variable one might be the amount of fertilizer, variable two might be the amount of



Soybean Production
Year: 1997
Regions: States

Weights: Equal Weights

Contact: dcarr@gmu.edu

Fig. 66.2 CCmaps, based on data from the USDA–NASS Web page, related to soybean production in the United States. The plot shows the dependent variable production (*top slider*) that is conditioned on the two independent variables acreage (*bottom slider*) and yield (*right slider*) (Previously published as Fig. 1.7 in Symanzik and Carr (2008, p. 289))

precipitation, and the dependent variable might be the yield of a crop. One might expect that high values for fertilizer and precipitation result in a large yield. The nine maps show the relationship among the three variables in a geographic framework, thus allowing the consideration of the underlying spatial structure of the data and not only the statistical relationships. Cutoff values can be changed interactively, thus allowing the investigation of many possible settings. CCmaps are useful tools for the interactive generation of statistical hypotheses for medical, epidemiological, and environmental applications.

In Fig. 66.2, the 1997 soybean production in the United States is conditioned on acreage and yield. In an interactive environment, slider settings can be further modified to identify geographic areas of interest on the nine maps.

66.4 ESDA via Linking and Brushing

While in the previous section map views were interactively manipulated in a rather direct way, we will discuss in this section how map views and associated statistical displays can be interactively manipulated via linked views and brushing. This is commonly understood as the classical idea of EDA and ESDA. For a detailed discussion of concepts for interactive graphics, such as brushing, linked brushing, linked views, focusing, zooming, panning, slicing, rescaling, reformatting, rotations, projections, and the grand tour, the reader is referred to Symanzik (2004, Sect. 10.3). Main statistical plot types that can be frequently found as components of linked views include histograms, scatterplots, scatterplot matrices, the grand tour, parallel coordinate plots, bar charts, pie charts, spine plots, mosaic plots, ray–glyph plots, and cumulative curves (such as the Lorenz curve). Most of these plot types also were discussed in Symanzik (2004, Sect. 10.3). Figure 66.3 shows one map view that has been linked with two scatterplots. In addition, plots for spatial data, such as variogram–cloud plots (see Fig. 66.4) and spatially lagged scatterplots, can be components of the linked views.

The overall idea of brushing is to mark different subsets of the data in a particular plot with different colors, symbols, sizes, or point or line styles. This is usually done based on the visual appearance of patterns in a specific plot, for example, outliers that seem to be far away from the remaining points in a histogram or scatterplot, or clusters that seem to be well separated from each other. In the next plot that is being produced, the original assessment will be reevaluated, additional points may be marked, or points may be marked differently.

In the framework of linked brushing and linked views, the brushing information is carried over from one plot to the next. For example, outliers that are marked in a histogram or scatterplot will be marked in a similar way (with the same colors, symbols, sizes, or point or line styles) in all related plots, in particular on a map view as well. Monmonier (1989) introduced the term geographic brushing in reference to interacting with the map view of geographically referenced data.

In Fig. 66.3, the US cities with the highest index for education have been brushed in the left scatterplot, and cities with a high crime index have been identified by name in the same scatterplot. The map view shows the locations of these cities with the same color and symbols as in the scatterplot. Moreover, the scatterplot on the right reveals that a high crime index is associated with a high recreation index while a high education index is associated with a medium recreation index. Nothing striking is noticeable when comparing the brushed values for education and crime with the arts index. Extensions of brushing for spatial data, such as moving statistics, or brushing, applied to origin–destination flow data (Liu and Marble 1997), exist.

In advanced software environments, brushing can take place in any of the linked views, including the map view. So, when locations in a specific geographic region are marked, the other statistical views will reveal whether there is some possible statistical relationship among the data from the selected locations as well, for example, whether the statistical values are similar to each other or whether the statistical values span the entire range of the underlying data distribution.

Linked brushing is not always one-to-one between the different displays. In a variogram–cloud plot, the absolute difference (or a related measure) of a variable of interest is calculated for all pairs of spatial locations, and this measure

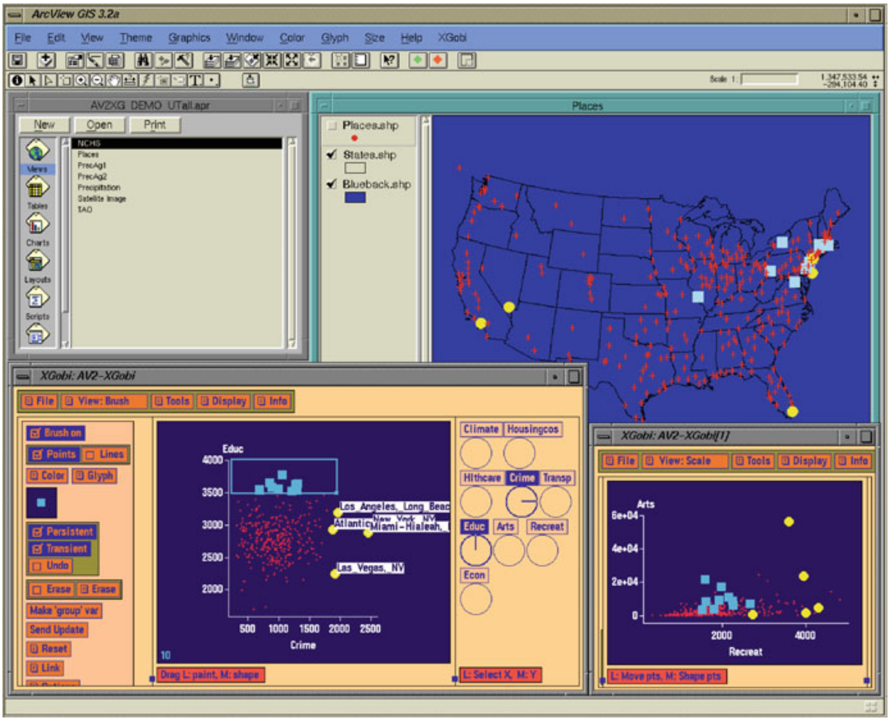


Fig. 66.3 Screenshot of the "Places" data in ArcView/XGobi. A map view of 329 cities in the United States is displayed in ArcView at the top. The two XGobi windows at the bottom are showing scatterplots of crime (*horizontal*) versus education (*vertical*) (*left*) and recreation (*horizontal*) versus arts (*vertical*) (*right*). Locations of high crime have been brushed and identified, representing some of the big cities in the United States. Also, locations of high education (above 3,500) have been brushed, mostly representing locations in the northeastern United States. All displays have been linked (Previously published as Fig. 10.1 in Symanzik (2004, p. 299))

is plotted against the Euclidean distance between the two associated points (up to a cutoff distance chosen by the researcher). Thus, when brushing one point in a variogram–cloud plot, this needs to be translated to a pair of spatial locations that are brushed in the map view.

Figure 66.4 shows such a link for precipitation measurements in the northeastern United States. In this figure, the highest values in the variogram–cloud plot (up to the cutoff distance) have been brushed. The map view shows two points, that is, spatial locations, that are connected to several other spatial locations. The location in the northeast likely is a spatial outlier as its precipitation measurements are considerably different (either higher or lower) than those from all nearby locations. A next step would be to explore additional variables for these locations, starting with elevation. The location in the southwest likely is not a spatial outlier; rather, there is some considerable local variation happening in this region as this location is only connected to some, but by far not all, locations in its neighborhood.

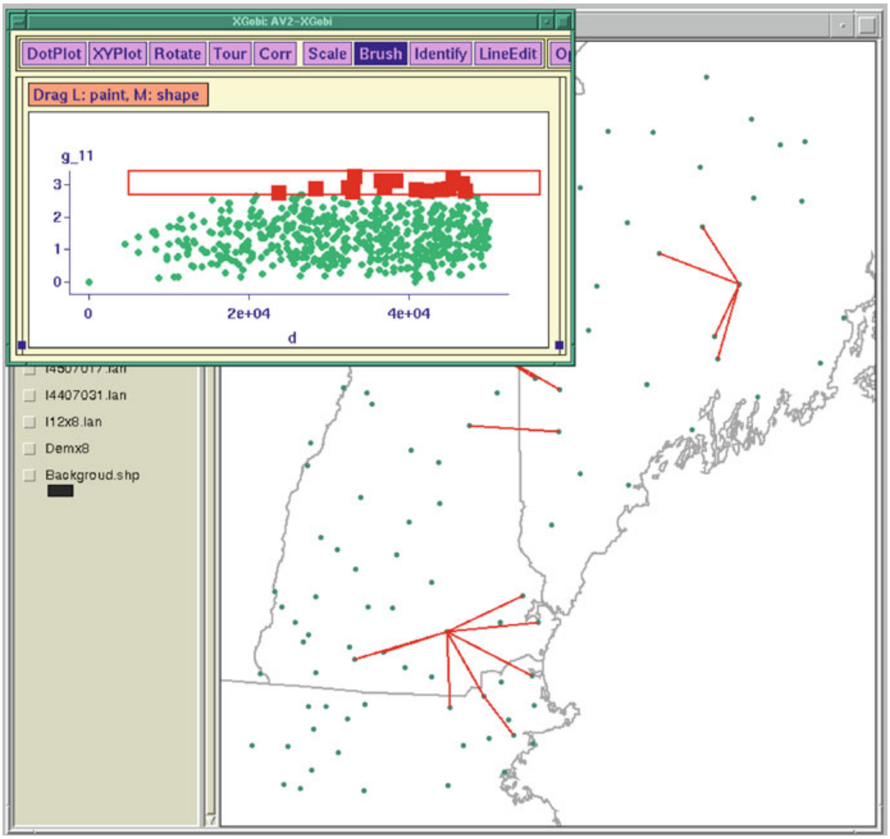


Fig. 66.4 Example of a variogram–cloud plot that is linked to a map view, based on precipitation data for the northeastern United States. In the *upper left* XGobi window, we have brushed (using a *solid rectangle*) the highest values in the variogram-cloud plot. In the *lower right* ArcView map view, each pair of locations, related to a point that has been brushed in XGobi, has been connected by a line (Previously published as Fig. 2 in Symanzik et al. (2000, p. 477). Reprinted with permission from the *Journal of Computational and Graphical Statistics*. Copyright 2000 by the American Statistical Association. All rights reserved)

66.5 Local Indicators of Spatial Association (LISA)

Moran's I statistic is a well-known measure for spatial autocorrelation at the global level for lattice data. Anselin (1995) introduced a local Moran statistic, a local Gamma statistic, a local Geary statistic, a Moran scatterplot, and other LISA statistics to assess the spatial association at a location *i*. The LISA statistics allow to identify local spatial clusters and to assess local instability. Moreover, the LISA statistics allow to assess the influence of a single location on the corresponding global statistic, a feature that is as important as being able to identify influential points in a regression framework.

LISA statistics are probably the most frequently applied ESDA technique, with applications in areas as diverse as regional sciences, spatial econometrics, epidemiology, social sciences, and criminology. Despite its wide use, one should keep in mind that LISA statistics are exploratory in nature, and, usually, additional steps are required to confirm the initial results derived from LISA statistics.

66.6 Software for ESDA

ESDA is highly dependent on software that supports various types of statistical displays, map views, and that allows linked brushing. Two main approaches have been developed during the last 25 years: Conducting ESDA in software environments where a geographic information system (GIS) is linked to statistical software packages and in stand-alone statistical software solutions. A more detailed overview of various software solutions for ESDA has been provided in Symanzik (2004, Sect. 10.6.1).

66.6.1 ESDA and GIS

Fotheringham (1992) pointed out that it is not necessary to conduct an exploratory spatial data analysis within a GIS, but that in many circumstances, using a GIS to do so might simplify the exploration of the data and provide insights that otherwise might be missed. Therefore, over the next decade, several researchers developed software that linked GIS with statistical software, or they added statistical features to existing GIS.

In Anselin (1994), a series of ESDA techniques were discussed in the context of a GIS, with the primary focus on exploring the spatial nature of the underlying data. These techniques could be classified as techniques based on the neighborhood view of spatial association (such as Moran scatterplots and LISA statistics) and as techniques based on the distance view of spatial association (such as spatially lagged scatterplots and variogram–cloud plots). Various software links between GIS such as Arc/Info, ArcView, and Grassland and one or more statistical software packages implemented several of these techniques. Some of the links that were developed and maintained over a longer time period were links between Arc/Info, respectively, ArcView, and SpaceStat (Anselin et al. 1993; Bao and Anselin 1997) and links between ArcView, XGobi, and XploRe (Cook et al. 1996; Symanzik et al. 2000). One major limitation of such software links is that whenever one of the individual software packages is modified with respect to the functionality of the link, the other software packages have to be modified accordingly.

66.6.2 Stand-Alone Software for ESDA

In contrast to linking GIS and statistical software, several software developers focused on the development of stand-alone statistical software that also support map views of the spatial locations that are linked with statistical displays. Some of the best known examples are Spider (Haslett et al. 1990), REGARD (Unwin et al. 1990; Unwin 1994), and, more recently, GeoDa (Anselin et al. 2006). One major limitation of stand-alone software for ESDA is that the functionality that is usually available in a GIS has to be reimplemented in a statistical software package.

In recent years, R (R Development Core Team 2011) has become the *lingua franca* of statistics. Since its appearance around 1996 (Ihaka and Gentleman 1996), R has been further advanced by thousands of creators of contributed packages (almost 4,000 in May 2012) that provide all kinds of additional functionality beyond the original R base functionality. This includes packages for maps, color selections, EDA and ESDA, and advanced statistical functionality for spatial data, such as the following:

- *maptools* (http://cran.r-project.org/web/packages/maptools/index.html) that allows to read and manipulate geographic data, in particular ESRI shapefiles
- *maps* (http://cran.r-project.org/web/packages/maps/index.html) that provides access to a variety of maps
- *RgoogleMaps* (http://cran.r-project.org/web/packages/RgoogleMaps/index.html) that allows to query the Google server for static maps and to use one of the Google maps as a background image to overlay statistical plots from within R
- *RColorBrewer* (http://cran.r-project.org/web/packages/RColorBrewer/index.html), the R implementation of http://colorbrewer2.org, for good color choices for maps and other plots
- *iplots* (http://cran.r-project.org/web/packages/iplots/index.html), an R package in the spirit of Spider and REGARD, for interactive plots in R, including maps
- *splancs* (http://cran.r-project.org/web/packages/splancs/index.html) for the exploration and analysis of spatial and space–time point patterns,
- *spatstat* (http://cran.r-project.org/web/packages/spatstat/index.html) for the exploration and analysis of spatial data, mainly spatial point patterns
- *spdep* (http://cran.r-project.org/web/packages/spdep/index.html) for the analysis of spatial dependence at a local and global scale, including Moran and LISA statistics
- *geoR* (http://cran.r-project.org/web/packages/geoR/index.html) for the exploration and analysis of geostatistical data
- *gstat* (http://cran.r-project.org/web/packages/gstat/index.html) for modeling, prediction, and simulation of spatial and spatiotemporal geostatistical data
- *spgrw* (http://cran.r-project.org/web/packages/spgwr/index.html) for computing geographically weighted regression

While the Web pages listed above provide detailed user guides and information how to use each of these packages, Bivand (2010) demonstrated how many of the ESDA techniques described in this chapter can be performed in R. An extended overview of additional R packages for the reading, exploration, visualization, and analysis of spatial data can be found at http://cran.r-project.org/web/views/Spatial.html.

66.7 Conclusions

In this chapter, we have provided an overview of techniques, methods, and software solutions for ESDA. Most of the developments took place during the last 25–30 years. Due to the rapid development of computer hardware, including high-quality graphic displays, over the last few decades, ESDA techniques are nowadays easily accessible for many researchers on a wide variety of hardware platforms.

A current hotspot for ongoing development of ESDA techniques is R and its thousands of contributed packages. For a few decades, software packages for exploratory data analysis were relatively weak for confirmatory data analysis (using John Tukey's terms here), and vice versa. However, R is continuously getting stronger for both types of data analyses, and it is able to handle a large variety of GIS data formats. It can be expected that in the near future, exploratory and confirmatory data analyses will be conducted almost simultaneously in R or some similar software environment. Once a researcher detects something of interest in a spatial data set via ESDA, a confirmatory analysis can immediately follow, and once a confirmatory analysis has been conducted, ESDA can be used to further explore the spatial fit of the fitted model, its residuals, and so on.

A trend in recent years has been to provide access to spatial data for everyone via Web interfaces. This includes the previously introduced Web sites for interactive micromaps (http://www.nass.usda.gov/research/sumpant.htm and http://www.statecancerprofiles.cancer.gov/micromaps), but, even more, Web-based software such as *gapminder* (Rosling and Johansson 2009), accessible at http://www.gapminder.org/. The Google version, called *Google Public Data Explorer*, accessible at http://www.google.com/publicdata/directory, might become a tool that provides easy and fast access to EDA and ESDA techniques for millions of Web users.

66.8 Cross-References

▶ Spatial Clustering and Autocorrelation in Health Events

References

Andrienko N, Andrienko G, Savinov A, Voss H, Wettschereck D (2001) Exploratory analysis of spatial data using interactive maps and data mining. Cartogr Geogr Inform Sci 28(3):151–165

Anscombe FJ (1973) Graphs in statistical analysis. Am Statistician 27(1):17–21

Anselin L (1994) Exploratory spatial data analysis and geographic information systems. In: Painho M (ed) New tools for spatial analysis. Eurostat, Luxembourg, pp 45–54

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

Anselin L, Dodson RF, Hudak S (1993) Linking GIS and spatial data analysis in practice. Geogr Sys 1(1):3–23

Anselin L, Syabri I, Kho Y (2006) GeoDa: an introduction to spatial data analysis. Geogr Anal 38(1):5–22

Bao S, Anselin L (1997) Linking spatial statistics with GIS: operational issues in the SpaceStat–ArcView link and the S + Grassland link. In: 1997 proceedings of the section on statistical graphics. American Statistical Association, Alexandria, pp 61–66

Bivand RS (2010) Exploratory spatial data analysis. In: Fischer MM, Getis A (eds) Handbook of applied spatial analysis: software tools, methods and applications. Springer, Berlin/Heidelberg, pp 219–254

Carr DB, Pickle LW (2010) Visualizing data patterns with micromaps. Chapman & Hall/CRC, Boca Raton

Carr DB, Wallin JF, Carr DA (2000) Two new templates for epidemiology applications: linked micromap plots and conditioned choropleth maps. Stat Med 19(17–18): 2521–2538

Carr DB, Chen J, Bell BS, Pickle LW, Zhang Y (2002) Interactive linked micromap plots and dynamically conditioned choropleth maps. In: Proceedings of the second national conference on digital government research, Digital Government Research Center (DGRC), pp 61–67, http://www.dgrc.org/conferences/2002_proceedings.jsp

Cook D, Majure JJ, Symanzik J, Cressie N (1996) Dynamic graphics in a GIS: exploring and analyzing multivariate spatial data using linked software. Comput Stat 11(4):467–480. Special issue on computeraided analysis of spatial data

Cressie NAC (1993) Statistics for spatial data, revised edn. Wiley, New York

Fischer MM, Wang J (2011) Spatial data analysis:models, methods and techniques. Springer, Berlin/Heidelberg/New York

Fotheringham AS (1992) Exploratory spatial data analysis and GIS. Environ Plann A 24(2):1675–1678

Harrower MA, Brewer CA (2003) ColorBrewer.org: an online tool for selecting color schemes for maps. Cartogr J 40(1):27–37

Haslett J, Wills G, Unwin A (1990) SPIDER – an interactive statistical tool for the analysis of spatially distributed data. Int J Geogr Inform Syst 4(3):285–296

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5(3):299–314

Liu L, Marble D (1997) Brushing spatial flow data sets. In: 1997 proceedings of the section on statistical graphics, American Statistical Association, Alexandria, pp 67–72

Monmonier M (1989) Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. Geogr Anal 21(1):81–84

Monmonier M (1996) How to lie with maps, 2nd edn. University of Chicago Press, Chicago

Openshaw S (1984) The modifiable areal unit problem. In: Concepts and techniques in modern geography No. 38. Geo Books, Regency House, Norwich

R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3–900051–07–0. http://www.R-project.org/

Ripley BD (1981) Spatial statistics. Wiley, New York

Rosling H, Johansson C (2009) Gapminder: liberating the x–axis from the burden of time. Stat Comput Stat Graph Newslett 20(1):4–7

Snow J (1936) Snow on cholera: being a reprint of two papers by John Snow, M.D. together with a biographical memoir by B. W. Richardson, M.D. and an introduction by Wade Hampton Frost, M.D. The Commonwealth Fund/Oxford University Press, New York/London

Symanzik J (2004) Interactive and dynamic graphics. In: Gentle JE, Härdle W, Mori Y (eds) Handbook of computational statistics – concepts and methods. Springer, Berlin/Heidelberg, pp 293–336

Symanzik J, Carr DB (2008) Interactive linked micromap plots for the display of geographically referenced statistical data. In: Chen C, Härdle W, Unwin A (eds) Handbook of data visualization. Springer, Berlin/Heidelberg, pp 267–294 & 2 color plates

Symanzik J, Cook D, Lewin-Koh N, Majure JJ, Megretskaia I (2000) Linking ArcView and XGobi: insight behind the front end. J Comput Graph Stat 9(3):470–490

Tukey JW (1977) Exploratory data analysis. Addison Wesley, Reading

Unwin A (1994) REGARDing geographic data. In: Dirschedl P, Ostermann R (eds) Computational statistics. Physica–Verlag, Heidelberg, pp 315–326

Unwin A, Wills G, Haslett J (1990) REGARD – graphical analysis of regional data. In: 1990 proceedings of the section on statistical graphics, American Statistical Association, Alexandria, pp 36–41

Wang X, Chen JX, Carr DB, Bell BS, Pickle LW (2002) Geographic statistics visualization: web–based linked micromap plots. Comput Sci Eng 4(3):90–94