B.2 Exploratory Spatial Data Analysis

Roger S. Bivand

B.2.1 Introduction

Exploratory spatial data analysis (ESDA) as used in spatial statistics, spatial econometrics and geostatistics, developed from exploratory data analysis (EDA). In particular, two threads that are central to a-spatial EDA have carried over to ESDA – the importance of the data themselves, and the importance of analytical graphics in representing chosen characteristics of the data.

This chapter will present some of the underlying intentions of ESDA, and survey some of the outcomes. This will necessarily involve the use of software, since most EDA and ESDA techniques presuppose the use of computing resources in some form. Here, we will use R-2.8.0 (R Development Core Team 2008), because the integration of its output with the printed page is somewhat less problematic than that of systems with graphical user interfaces. The choice of R also touches nicely on the Bell Labs' inheritance of the S language, with its links to John Tukey and Bill Cleveland, described by Chambers (2008), himself a major contributor to applied statistics.

In his recent book, Chambers (2008, p.1) proposes the principle that: '*our Mission, as users and creators of software for data analysis, is to enable the best and most thorough exploration of the data possible.*' In this tradition, exploration is part of the process of formulating the question and organising the data so as to be able to answer that question. As Cox and Jones (1981) note, the tradition stands some way from the classical division between descriptive and inferential statistics. The substantive research problem is what matters: '*As John Tukey often remarked, better an approximate answer to the right question than an exact answer to the wrong question*' (Chambers 2008, p.3). This may involve exploring distributional assumptions in relation to variables of interest, perhaps including transformations or the removal of trends, but does presuppose that the analyst wants to find the 'right' question, a point to which we will return in conclusion.

Attentive reading of classics in spatial data analysis, such as Cressie (1993), and Bailey and Gatrell (1995), shows that both EDA and ESDA have long played an important part in finding the 'right question'. This heritage is continued in

M.M. Fischer and A. Getis (eds.), *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, DOI 10.1007/978-3-642-03647-7_13,

newer presentations like Waller and Gotway (2004), and Schabenberger and Gotway (2005). While many point to the increasing availability of spatial data as such, it seems typical that the data we need to attack a given research problem is often costly to gather, often collected for other purposes, and often not with the support best suited to the problem. Consequently, we need to try to make the best possible use of the available data, both in connection with the thrust of our research problem, and looking out for signals suggesting potentially richer questions.

Our research problem focuses our attention on components of variation in our response variables of interest, on variables or spatial locations that account for observed variability. In terms proposed by Tukey, the response variables constitute the data, and what we know about the data based on previous knowledge is the smooth, leaving residual variation in the rough. Exploratory data analysis opens up two complementary possibilities: that our prior knowledge – choice of variables in the smooth and their functional form – deserves revision, and that patterning in the rough can be shifted to the smooth. In particular, spatial patterning in the rough can be used as a 'spatial' smooth in some cases, especially when observations on omitted variables shown in the spatial patterning are not available for any reason. Exploratory spatial data analysis plays an important role in the examination of a-spatial residual variation, to try to see whether spatial patterning can be used to account for the variation in the data in a more satisfactory way.

In this chapter, we will work with examples to show some of the available methods that build on the EDA approach to data analysis. The examples use legacy data sets, and will not necessarily start from univariate EDA as perhaps they should, but rather illustrate fresh groups of methods in turn in each section. One example data set that will be used frequently is the French 'Moral Statistics' data set discussed in detail by Friendly (2007) and taken up in connection with geographical visualization by Dykes and Brunsdon (2007).

B.2.2 Plotting and exploratory data analysis

Cox and Jones (1981, p.135) describe one of the basic attitudes of exploratory data analysis as: '*plot both your data and the results of data analysis*' – pointing directly to statistical graphics. Plotting multiple versions of a display by hand is so time-consuming that actually using EDA visualization had to wait until computer graphics resources became available, despite the hopes expressed in Tukey (1977) that paper and pencil would be enough. Naturally, in the 1970's computer graphics were not very sophisticated, and portability across graphics devices other than line printers was very hard to achieve, so early Minitab EDA output was formatted for line printers (as was output from the subroutines provided in Velleman and Hoaglin 1981).

Output to interactive user terminals was hard, with the initial exception of the first Apple Macintosh computers, which provided both a monochrome graphics screen and a pointing device. This was used by DataDesk and other software to provide ways of exploring data visually; other software for PC systems did not have such a standardised graphics library until much later; Systat for example used a pen plotter for graphics output. Workstation systems, largely running Unix, did have mature graphics libraries, but with a plethora of different versions – Silicon Graphics™ machines were well-liked but very costly.

Since those early years, cross-platform software accommodating differences in graphics devices has become more common, in addition to cross-platform graphics libraries – Xgobi transitioned to use the Gnome graphics library as Ggobi[1], and may now be used on many platforms (Cook and Swayne 2007). Other data visualization software has chosen to use Java as a virtualized platform, as we will see in Section B.2.2 in the case of Mondrian[2] (Theus 2002). This is not dissimilar to the use of XLISP to underpin XLispStat on a cross-platform basis, used by Brunsdon (1998) for exploratory spatial data analysis. The use of Tcl/Tk by Dykes (1997, 1998) is a further example of a developer 'borrowing strength' from an underlying programming language, which provided cross-platform support for interactive graphics, for exploratory spatial data analysis.

The concise introduction to exploratory data analysis by Jacoby (1997) provides us with a first data set and details of the computing environment used – S was used for demonstrating many of the techniques presented, and they may be reproduced using R. The univariate EDA methods used are described by Jacoby (1997), and implementation details of the graphics functions used can be found in Murrell (2005). Sarkar (2007) shows how to use lattice graphics in R to display panels accommodating both the variable(s) of interest and conditioning variables – this builds on Trellis graphics introduced in Cleveland (1993) and Becker et al. (1996). The data set contains Medicaid program quality scores for 48 U.S. contiguous states for 1986, here stored externally in a shapefile, and read into a `SpatialPolygonsDataFrame` object.

Figure B.2.1 shows a number of graphical representations of the observed values of the program quality scores (`PQS`), ranging from the simple but informative stem and leaf tally on panel a), through a jittered stripchart on panel b), and a boxplot [see panel c)]; [see Chapter E.2 for a further discussion of the use of boxplots], to a composite of a histogram with default bin widths and starting point, overlaid by density plots for three bandwidths, and furnished with a rug plot along the bottom axis showing the data values in panel d). As in the remainder of this chapter, the code snippets illustrate how the displays may be made, sometimes in abbreviated form to simplify presentation. The `PQS` variable belongs to the `medicaid` object, here a `SpatialPolygonsDataFrame` object, and is accessed using the $ operator.

[1] http://www.ggobi.org/

[2] http://www.theusrus.de/Mondrian/index.html

```
> stem(medicaid$PQS, scale = 2)
> stripchart (medicaid$PQS, method = 'jitter', vertical = TRUE)
> boxplot(medicaid$PQS)
> hist(medicaid$PQS, col = 'grey90', freq = FALSE)
> lines(density(medicaid$PQS, bw = 15), lwd = 2)
> rug(medicaid$PQS)
```

It is helpful to contrast the smoother generalisation of the boxplot, the histogram, and the density plot with the larger bandwidth to the stem and leaf plot, the stripchart, the rug plot, and the density plot with smaller bandwidth. The first group of techniques shows the 'big picture', while the second group gives more detail, and may even suggest some clustering of the observed values.

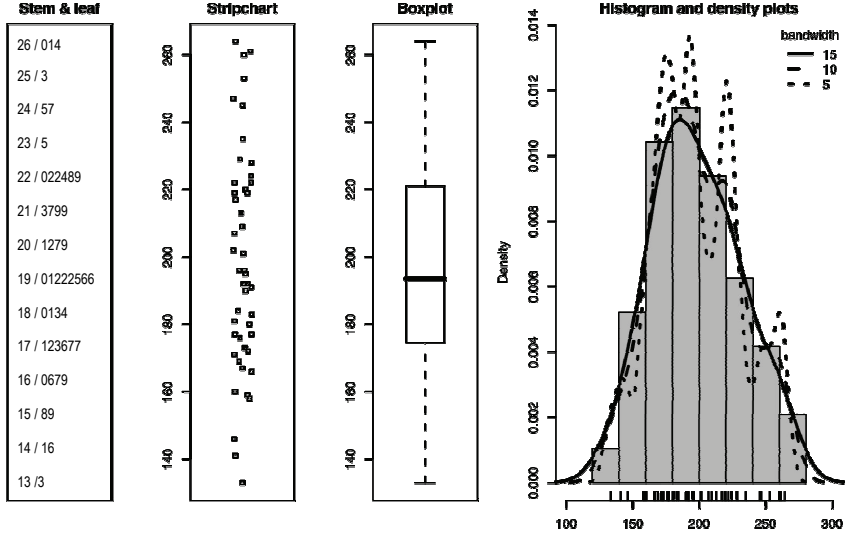

Fig. B.2.1. Displays of the reported Medicaid program quality score values 1986: a) stem and leaf display – here ordered with large values at the top to match the next two panels; b) stripchart with jittered points; c) boxplot with standard whiskers; d) histogram with overplotted density curves for selected bandwidths

All of these techniques use an ordering of the data, as do the two shown in Fig. B.2.2. The plot of the empirical cumulative distribution function of the observed values involves their ordering, and the tallying of ties, to be compared with their rank orders. A uniform distribution gives a more or less straight diagonal line, but the plot is perhaps most useful for exploring unusual breaks between values. The functions can be used in the following way.

```
> plot(ecdf(medicaid$PQS)
> o <- order(medicaid$PQS)
> dotchart(medicaid$PQS[o], labels = as.character (medicaid$STATE_ABBR) [o],
+ groups = medicaid$Division[o])
```

The accompanying dotchart displays all the observed values, with state labels and grouped by statistical division. It introduces the concept of conditioning, here on division, to permit the comparison of ordered values in relation to a structuring variable. With 48 observations, the dotchart is becoming illegible, and would probably benefit from aggregation: curiously, both stem and leaf displays and dotcharts may be viewed 'out of focus' to look at a 'big picture'. Zooming in, it does, however, permit the retrieval of values for identified observations, so that the analyst can see 'which are which'.

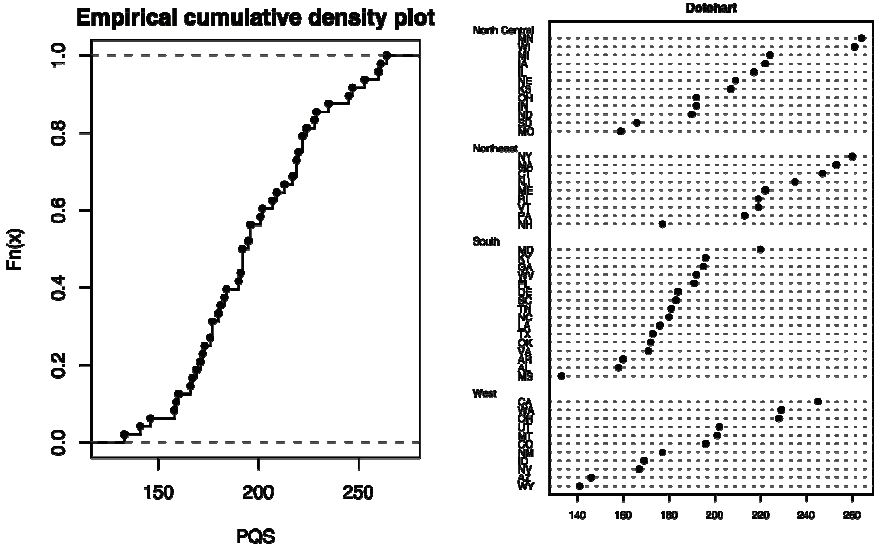

Fig. B.2.2. Medicaid program quality scores 1986: a) empirical cumulative distribution function, and b) dotchart

Dynamically linked graphics. The interactive identification of observations, and groups of observations with apparently shared characteristics, has emerged as an important exploratory tool in data analysis. A pointing device is used to select one or more observation on one graphics display, and the selection is dynamically displayed on all other data displays, both text and graphical. Naturally, this is hard to represent in print, but has generated a rich literature and many software implementations. The background for dynamically linked graphics is discussed in detail by Becker et al. (1987).

One implementation that has served as a research forum for exploring the possibilities offered by multivariate dynamically linked graphics is XGobi (Cook et al. 1996, 1997). From the beginning, XGobi developers were interested in linking to map displays (Symanzik et al. 2000), leaving geographical representation to a desktop GIS. Cook and Swayne (2007) show how dynamically linked graphics have developed and matured, and how dynamic data manipulation, such as 'flying

through' clouds of multivariate data points, can be related to static but reproducible graphic displays. Theus (2002) describes the Mondrian software implementation of many multivariate dynamically linked graphics, including a map view. Naturally, showcasing dynamically linked graphics in print is not possible, but any `SpatialPolygonsDataFrame` object may be exported in the correct format for Mondrian in this way.

```
> library(maptools)
> sp2Mondrian(medicaid, 'medicaid.txt')
```

B.2.3 Geovisualization

While data visualization is perhaps more closely related to data analysis, the work of cartographers brings in scientific and information visualization. This cross-fertilization has led to a range of innovative software tools, many of which are documented in the work of the Commission on GeoVisualization of the International Cartographic Association.[3] Work by cartographers is welcomed in statistical graphics; for example the results of studies into the use and abuse of colour in visualization have diffused widely. Geovisualization is not separate from exploratory spatial data analysis, but rather constitutes the backbone of ESDA, joining up the large range of techniques proposed for examining spatial data in a shared and easily comprehended visualization framework.

Monmonier (1989) introduced the concept of geographical brushing, borrowing from brushing in dynamically linked graphics, selecting observations for linked highlighting from a map representation, most often choosing observations within a map window. Many of these techniques for linked highlighting were implemented in software described by Haslett et al. (1991) and Haslett (1992), and followed up by Dykes (1997, 1998) in the 'cartographic data visualizer' cross-platform implementation. Progress made during the 1990s is summarised by Andrienko and Andrienko (1999) and Gahegan (1999).

Like Mondrian, GeoVISTA studio (Takatsuka and Gahegan 2002) uses Java as an integrating cross-platform framework linking the dynamic display of spatial data with its conceptual underpinnings. The treatment of ontologies as an integral part of geovisualization software is developed by MacEachren et al. (2004a, b). The approach taken by GeoDa (Anselin et al. 2006) is simpler, combining dynamically linked graphics, map views, and numerical exploratory techniques to be discussed in Section B.2.5.

Dykes and Mountain (2003) add the temporal dimension to interactive graphics with spatial data, while the data is smoothed by geographical weighting in the methods described by Dykes and Brunsdon (2007). Many of these proposals seem to address issues of importance for visualization research as such, rather than for

[3] http://geoanalytics.net/ica/

applied data analysis; by contrast, Wood et al. (2007) combine innovative geovisualization with 'mashups', permitting output graphics to be viewed using either browser-based mapping applications, or stand-alone software and geodata distribution systems like Google Earth™.

Thematic cartography. Just as graphical output may be described as lying on a continuum from analytical to presentation in terms of the requirements of its viewers, so may cartographic output (Slocum et al. 2005). Thematic cartography is an important part of exploratory data analysis with spatial data, as well as playing a vital role in presenting model results. It is also crucial in the communication of the intermediate and final results of research, both on screen in applications and documents, and in print. Bailey and Gatrell (1995, pp.48-61) describe the development of computer mapping for analytical purposes. We will not be considering the use of cartograms here, although arguments can be made for their importance in ESDA (Dorling 1993, 1995). There are issues concering the legibility of cartograms, and further difficulties in the algorithmic construction of legible polygons, which led Durham et al. (2006) to complete the construction of acceptable units for the British Census by hand. In this review, we will be using *R* graphics methods largely documented in Bivand et al. (2008, pp.57-80), in particular the `spplot` methods for suitable objects; the first argument here is the object, and the second, a vector of variables to display using the same class intervals, here a single variable.

```
> lbls <- as.character(medicaid$STATE_ABBR)
> spl <- list('sp.text', coordinates(medicaid), lbls, cex = 0.6)
> spplot(medicaid, 'PQS', col.regions = grey.colors(20,
+ 0.95, 0.4), sp.layout = spl, col = 'grey30')
```

The example (see Fig. B.2.3) shows a map view of the program quality score variable; the `sp.layout` argument allows additional graphics components to be added to the output. The `spplot` method can take an argument setting the class intervals, but where none is given, it uses a default of 'pretty' numbers encompassing the range of the data with 19 equally spaced internal intervals, so taking 20 colour values. The `grey.colors` function creates a ramp of grey shades from its second to third argument value for a default gamma of 2.2, which seems to match some computer displays, projectors, and printed output adequately.

The grey shades chosen are not the same as those proposed by Brewer et al. (1997); Brewer and Pickle (2002) in ColorBrewer, mostly in not using the lightest or darkest greys, and by using a larger gamma than the one proposed there.[4] Having good control of class intervals and colours used is an important part of thematic cartography, and is far from easy to achieve in print. Readers willing to try out the code underlying this review are invited to explore alternative palettes to see whether the 'message' of the presented thematic maps is affected.

[4] The gamma correction is a component of the colour space implementation intended to neutralise the effect of the display medium (the default value of 2.2).

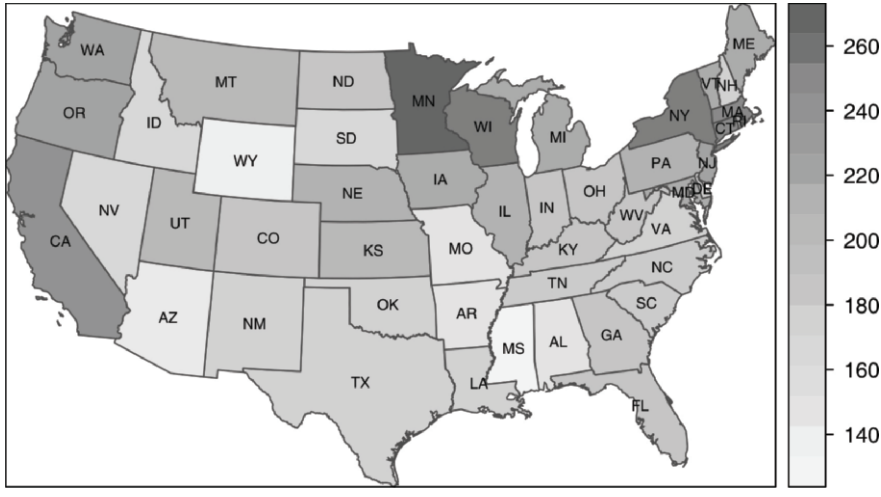

Fig. B.2.3. Medicaid program quality scores, 1986: thematic cartography as a method for statistical display

Conditioned choropleth maps. Trellis graphics displays are intended to permit the researcher to explore multivariate relationships by conditioning on potentially interesting variables (Becker et al. 1996). In an innovative paper building on modern statistical graphics, Carr et al. (2000) propose the use of linked micromaps, matching maps used to provide graphical indices for conditioned panels, and conditioned choropleth maps. They define CC maps in the following way: '*Similar to conditioning on sex and showing separate choropleth maps for males and females, CC maps provide for conditioning on the levels or values of variables and for the display of multiple choropleth maps. The basic difference in the examples here is that the conditioning does not distinguish separate populations within each unit of study but rather partitions the units of study*' (Carr et al. 2000, p.2530). More details and examples can be found in Carr et al. (2005).

In the classic North Carolina sudden infant death syndrome data set, a relationship is found between the Freeman-Tukey transformed SIDS rate for 1974-1978 by county and the Freeman-Tukey transformed nonwhite birth rate (Cressie 1993, pp.548-551).

We can use a lattice of conditioned choropleth maps to explore the spatial footprint of this relationship. We could convert the nonwhite birth rate into a categorical variable (factor) to partition the counties, but follow usual practice when conditioning panels on a numerical variable and use equal count overlapping shingles. The reasons for using overlapping shingles – to avoid the risk of giving the breaks in the conditioning variable too much influence in the display – are discussed by Becker et al. (1996, pp.142-147), and documented for *R* by Sarkar (2007, pp.177-187). With no overlap, `equal.count` would return members corresponding to quantiles for the number of conditioning levels required, but as can

be seen from Fig. B.2.4, the shadings in the panel strips do overlap, reflecting the chosen degree of protection from interval choice artefacts. The `equal.count` function in lattice allows us to construct a shingle, and to use it in `CCmaps`.

```
> library(lattice)
> sh_nw4 <- equal.count(nc.sids$ft.NWBIR74, number = 4
+    overlap = 1/5)
> CCmaps(nc.sids, 'ft.SID74', list(Nonwhite_births = sh_nw4))
```

As we move from lower left to lower right, then upper left to upper right across the panels of Fig. B.2.4, we see that the counties in each level of the shingle seem to be clustered, and that the choropleth map values of the variable of interest increase. This corresponds to the positive relationship reported between the variables, but also suggests that including the conditioning variable may reduce residual autocorrelation in a model of Freeman-Tukey transformed SIDS rates.

```
> gfrance <- readOGR('.', 'gfrance1')
> gfrance$Pop_crime <- gfrance$Crime_prop/100
```

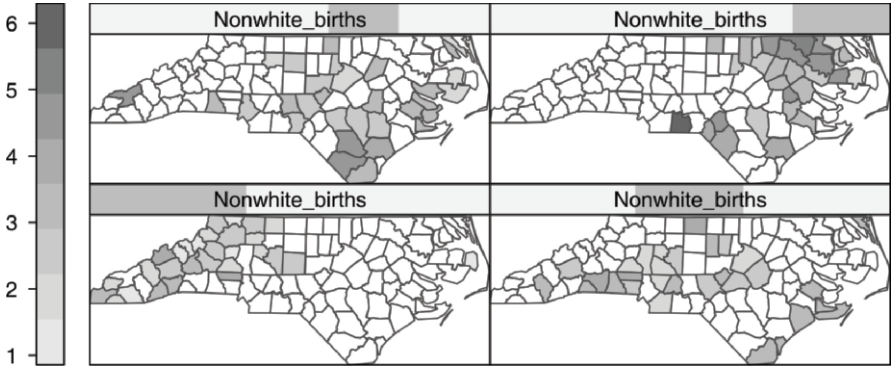

Fig. B.2.4. North Carolina Freeman-Tukey transformed SIDS rates by county for 1974-1978 conditioned on four shingles of the Freeman-Tukey transformed nonwhite live birth rates

Friendly (2007, p.395) includes a conditioned choropleth map of a variable from the Guerry French moral statistics data set: number of population per observation unit per crime against property, conditioned on wealth and literacy. The data set is available from the author's website[5] as a shapefile, which we read in as before. Figure B.2.5 shows the spatial distribution of the three variables being used here.

[5] http://www.math.yorku.ca/SCS/Gallery/guerry/maps.html#spatial

In order to plot a conditioned choropleth map, we need to construct two shingles, here, following Friendly (2007, p.395), with 10 percent overlap and two levels each.

```
> sh_wealth <- equal.count(gfrance$Wealth, number = 2,
+    overlap = 1/10)
> sh_literacy <- equal.count(gfrance$Literacy, number = 2,
+    overlap = 1/10)
> CCmaps(gfrance, 'Pop_crime', list(Wealth = sh_wealth,
+    Literacy = sh_literacy))
```

Figure B.2.6 differs from the original figure in a number of ways. The class intervals used for displaying the crime variable are not the same, and the legend is as provided by the underlying spplot and levelplot methods. The ordering of the panels also differs, but the spatial footprint is the same: wealthy and literate places experience higher rates of crime against property than poor and illiterate places. Note the inverted rate used – population per crime, rather than crime counts per inhabitant.

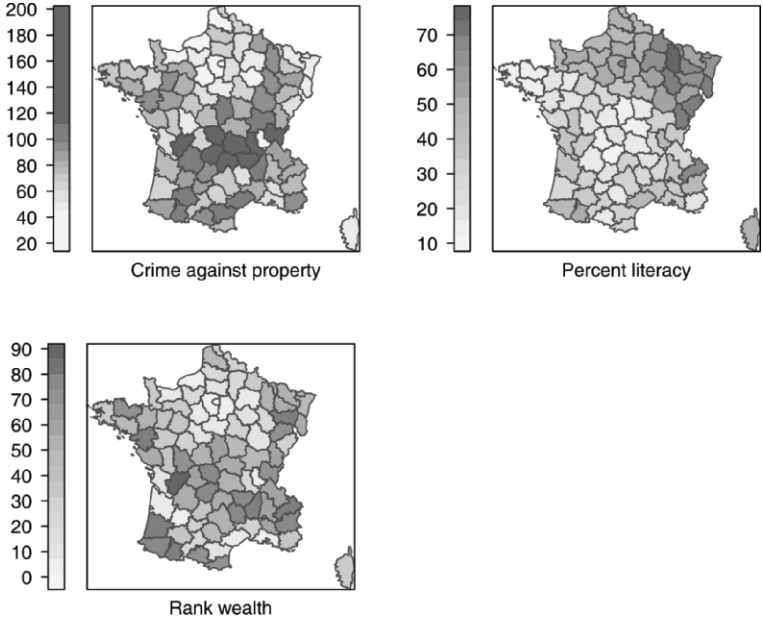

Fig. B.2.5. Choropleth maps of population per crime against property, rank wealth and percentage literacy, France (Friendly 2007)

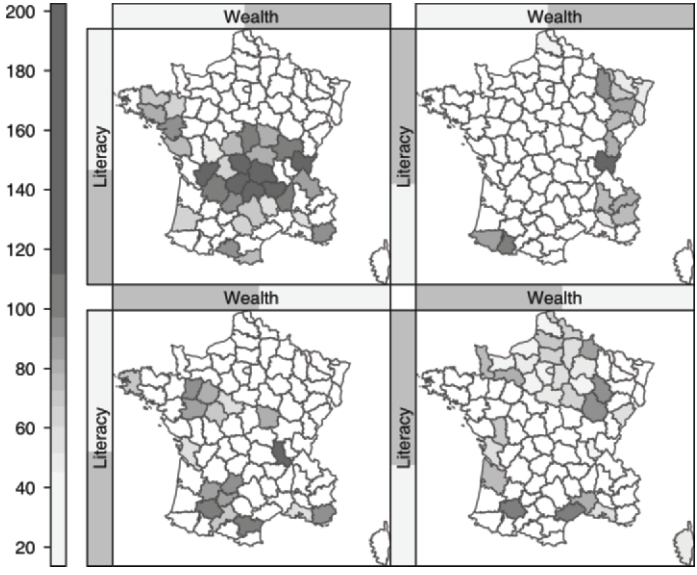

Fig. B.2.6. Choropleth maps of population per crime against property, conditioned on ranked wealth and percentage literacy, France (see Friendly 2007, p.395)

B.2.4 Exploring point patterns and geostatistics

Within the spatial analysis literature, ESDA has often been described as a subset of exploratory data analysis (Anselin 1998; Anselin et al. 2007). In a somewhat broader framework, however, it is perhaps difficult to distinguish ESDA as a subset of EDA, because many other strands feed into it, for example from information visualization and geographical information science, that are not present in EDA itself. It is tempting rather to see EDA as that part of ESDA of relevance to data where observations have no spatial location; such an over-arching view admits geovisualization as a part of ESDA, and places exchanges of knowledge and techniques between cartography and statistical graphics in a more natural context. Note that statisticians often use spatial data sets and objects as vehicles for their presentations (cf. Chambers 2008).

'*Analyzers of spatial data should ... be suspicious of observations when they are unusual with respect to their neighbours*' (Cressie 1993, p.33). This operational definition, buttressed by lively concern about data collection on the one hand and model specification on the other, is reflected in many of the examples presented in Cressie (1993), see also Unwin (1996), Kaluzny et al. (1998), Haining (2003), and Lloyd (2007). Often it is not sufficient to see ESDA as a toolbox of finished tools, because one frequently needs to 'get closer' to the data than the tools allow. This is one of the reasons for placing ESDA within an environment for statistical computing like R (Bivand et al. 2008), where users can engage the

data as far as they might wish. Finally, it should be noted that there are topics not yet adequately covered, such as ESDA for categorical data, surveyed and advanced by Boots (2006).

Exploring point patterns. While ESDA is often seen as being applied to areal data, in fact approaches to data analysis derived from EDA are used throughout spatial data analysis. For example, the $\hat{G}$ nearest neighbour distance measure used in point pattern analysis is simply a binned empirical cumulative density function plot of the nearest neighbour distances. Levine (2006) describes how many exploratory tools are provided in CrimeStat in an accessible fashion, and with the possibility of using simulation to see whether the patterns detected by the user ought to be treated as significant. Diggle (2003) gives many examples of the ways in which care in data analysis – respecting the data – informs even the most technically advanced statistical procedures. Baddeley et al. (2005) show how residuals from modelling a point pattern may be explored diagnostically; the *spatstat* package for `R` provides many ways to explore point patterns (Baddeley and Turner 2005). We will not be considering scan tests in this chapter; their provision in R is reviewed in Gómez-Rubio et al. (2005), and Bivand et al. (2008).

One of the classic data sets provided with R shows the locations of earthquakes near Fiji since 1964; the points in geographical coordinates are accompanied by the depth detected, the magnitude of the event, and the number of stations reporting it. These mean that we can treat it as a marked point pattern, for example using non-overlapping shingles of depth. The `xyplot` function takes a `formula` object as its first argument – this is a symbolic expression of the model to be visualised, here with points to be plotted on longitude and latitude conditioned on a depth shingle.

```
> data(quakes)
> depthgroup <- equal.count(quakes$depth, number = 3, overlap = 0)
> xyplot(lat ~ long | depthgroup, data=quakes, main='Fiji earthquakes',
+     type = c('p', 'g'))
```

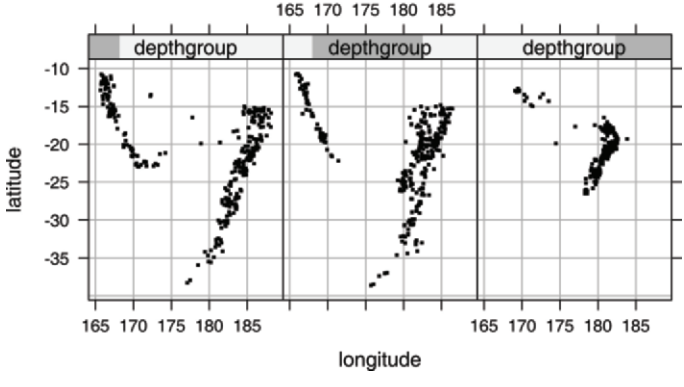

Fig. B.2.7. Seismic events near Fiji since 1964, conditioned on depth

Figure B.2.7 reproduces the conditioning of location on depth for the earthquake events discussed in detail by Murrell (2005, pp.126–141) and Sarkar (2007, pp.67–76). They also show how magnitude may also be visualized on conditioned scatterplots through a further shingle, or shaded symbols. Here we will consider how we might express the relative intensity of the point pattern using kernel smoothing. In order to do this we should project the geographical coordinates to the plane, using an appropriate set of parameters, here a Transverse Mercator projection used on Fiji. We use the default bisquare kernel with three chosen bandwidths, and set kernel values close to zero to `NA`.

```
> coordinates(quakes) <- c('long', 'lat')
> proj4string(quakes) <- CRS('+proj=longlat')
> quakes_tmerc <- spTransform(quakes, CRS('+init=epsg:3460'))
> library(splancs)
> pl <- bboxx(bbox(quakes_tmerc))
> h150k <- spkernel2d(as.points(coordinates(quakes_tmerc)),
+    poly = pl, h0 = 150000)
> is.na(h150k) <- h150k < .Machine$double.eps
```

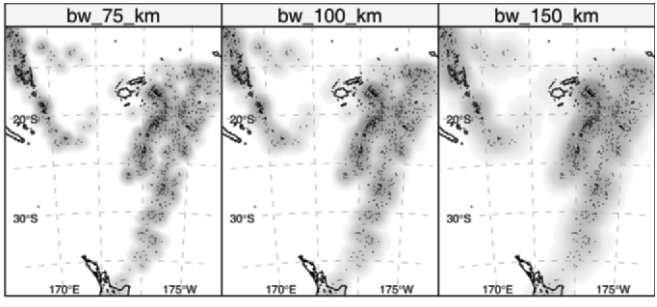

Fig. B.2.8. Kernel density plots of seismic events near Fiji; three increasing bandwidth settings

Figure B.2.8 shows density plots of the earthquake events for three increasing bandwidth values. The panels have also been furnished with shorelines and a graticule to aid in positioning the events. Had we additionally conditioned on depth or magnitude, or added tectonic boundaries, we might have come a little further. However, we can already see clearly that the observed pattern is not likely to be homogeneous. Exploration of point patterns is often helpful in drawing attention to the need to look for covariates that may account for inhomogeneity, or to possible use of a control point pattern to contrast with the observed cases.

Exploratory geostatistics. It is probable that more exploratory spatial data analysis is done in geostatistics than in the remaining domains of spatial data analysis; Cressie (1993) gives many examples. It is easy to grasp why interpolation is crucially dependent on identifying the 'right' model, in terms of the selection of observation locations, the fitting of models of spatial autocorrelation, de-

tecting useful covariates, and checking the appropriateness of assumptions such as isotropy. If a seriously sub-optimal model is chosen, both the predictions themselves and estimates of uncertainty around those values will not be as satisfactory as might have been achieved with the same data. Lloyd (2007) and Müller (2007) provide further discussions of techniques for making good use of the data to hand, and of the design of patterns of sampling locations to improve prediction. Geostatistics is also discussed in Chapter B.6.

Here we will use a data set of precipitation values for Switzerland, discussed in Diggle and Ribeiro (2007, pp.118-121, pp.149-150, pp.169-172), and used in the 'Spatial Interpolation Comparison 97' contest[6]. The examples demonstrate that geostatistics software, here R packages, provides much support for exploratory spatial data analysis, discussed for example by Bivand et al. (2008, pp.192, pp.195-200). Other software adopts the same approach; the Geostatistical Analyst extension to ArcGIS™ is well furnished with ESDA tools.

```
> library(geoR)
> data(SIC)
> plot(sic.100, borders = sic.borders, lowess = TRUE)
```

In the geoR package, the plot method for a `geodata` object is to make an ESDA graphic display. Setting the `lowess=` argument permits a smoothed line to be drawn through scatterplots of the data against the x and y coordinates, so that the four-panel display, shown in Fig. B.2.9, conveys a lot of information. On screen, the map symbols are coloured, to draw more attention to the spatial patterning of the quartiles of the variable of interest. We could of course condition a scatterplot of the location coordinates on a shingle of the variable of interest, as presented above. The histogram overplotted with a density line and rug plot shows that the data deserves more exploration, especially if a trend is mixing distributions of precipitation values together. The trend is here taken as the mean of the data, but the smoothed lines suggest that a spatial trend is present, of course in addition to the effect of station elevation, which has not been included here.

Location diagnostics. Should we attempt to add in a spatial trend, or a covariate, we should pay attention of the warning given by Unwin and Wrigley (1987) to use the same diagnostic tools as in any other modelling exercise. It is, as Fig. B.2.10. shows, quite frequently the case that some observations exert a more than proportional influence on the fitted model. The circles are proportional to Cook's influence statistic, and indicate that the distinguished stations ought to be looked at carefully, to see why they differ so much from their near neighbours. Note that most of the distinguished stations are on the edge of the study area.

[6] http://www.ai-geostats.org/index.php?id=data

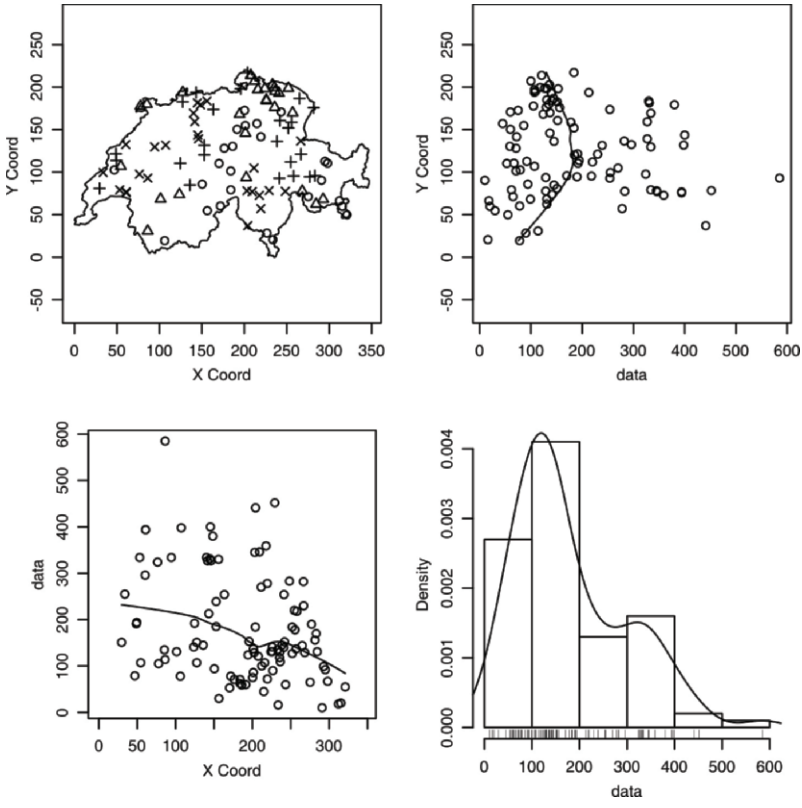

Fig. B.2.9. Exploratory geostatistical display of Swiss precipitation data from the 1997 Spatial Interpolation Comparison contest: a) precipitation quartiles; b) plot of precipitation by northings; c) plot of precipitation by eastings; d) histogram and density of precipitation

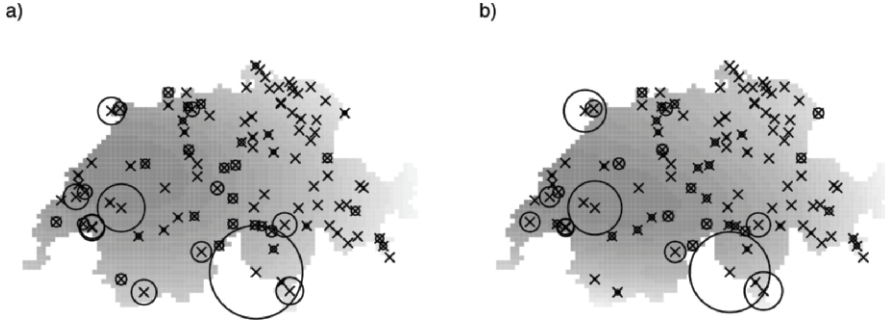

Fig. B.2.10. Influence plots for trend surfaces, Swiss precipitation data, circle radius proportional to Cook's influence statistic (Unwin and Wrigley 1987): a) quadratic trend surface; b) cubic trend surface

Variogram diagnostics. Variogram diagnostics are linked to other steps taken in exploring variables in geostatistics (Pebesma 2004). Using the spatial representations presented in Bivand et al. (2008), we can review some of the tools made available in the *gstat* package. First, we convert the Swiss precipitation data set to a suitable object form, and show a *h*-scatterplot of pairs of observed values conditioned on distance, expressed in the `breaks` argument to `hscat`. The formula interface used here places the variable of interest on the left hand side of the equation, and only the intercept term on the right hand side.

```
> library(gstat)
> sic.100SP <- SpatialPointsDataFrame(SpatialPoints(sic.100$coords),
+     data = data.frame(precip = sic.100$data))
> hscat(precip ~ 1, data = sic.100SP, breaks = seq(0, 120,
+     20))
```

The first diagnostic plot (Fig. B.2.11) is known as an *h*-scatterplot, and conditions a scatterplot of the values at pairs of locations on the binned distance h_{ij} between them; the diagonal lines represent perfect correlation. The sample correlations between the observed values at locations *i* and *j* are perhaps a little hard to read in a monochrome plot, so are repeated in text output, declining from 0.714 in the first 20km bin, to 0.344 between 20 and 40km, and going through zero in the third bin. It appears, then, that nearer observations are more like one another, and that the similarity declines with distance.

By defining a `gstat` object, we can easily create variograms of different kinds by passing this object and additional arguments to `variogram`.

```
> g <- gstat(id = 'precip', formula = precip ~ 1, data = sic.100SP)
> evgm <- variogram(g, cutoff = 100, width = 5)
> revgm <- variogram(g, cutoff = 100, width = 5, cressie = TRUE)
> cevgm <- variogram(g, cutoff = 100, width = 5, cloud = TRUE)
```

Figure B.2.12 shows a variogram cloud plot and a plot of empirical variogram values for twenty 5km wide bins, for classical and robust versions of the variogram. The bin borders are shown to highlight the way in which the empirical variogram is constructed as a measure of central tendency of squared differences in the variable of interest between pairs of points whose inter-point distance falls into the bin. Cressie (1993, pp.74-83) provides the development of a robust estimator, shown with a dashed line in Fig. B.2.12, that reduces the impact of unusually large differences in value between near neighbours. The *fields* package returns number summaries by bin in addition to the classical variogram estimator in output from the `vgram` function.

Figure B.2.13 shows a variogram map, and four empirical variograms for four axes at 0°, 45°, 90° and 135°; the variogram direction lines are coded in the same way on both panels. A variogram map is centred around (0, 0) and has map dimension and cell size similar to cutoff and interval width values; it is constructed by averaging pairs that have distance within a certain bin. In this case, we see that

the structure aligned with the 45° direction corresponds to lower variogram values for nearer bins. Recall that here we taking the trend as the mean only, ignoring the impact of large scale spatial trends and covariates.

Directionality. Finally, we follow Bivand et al. (2008, pp.205-206) in examining possible anisotropy in the data set. Using the same bins as earlier, we add arguments to the `variogram` function to create objects for plotting.

```
> mevgm <- variogram(g, cutoff = 100, width = 5, map = TRUE)
> aevgm <- variogram(g, cutoff = 100, width = 5, alpha = c(0,
+ 45, 90, 135))
```

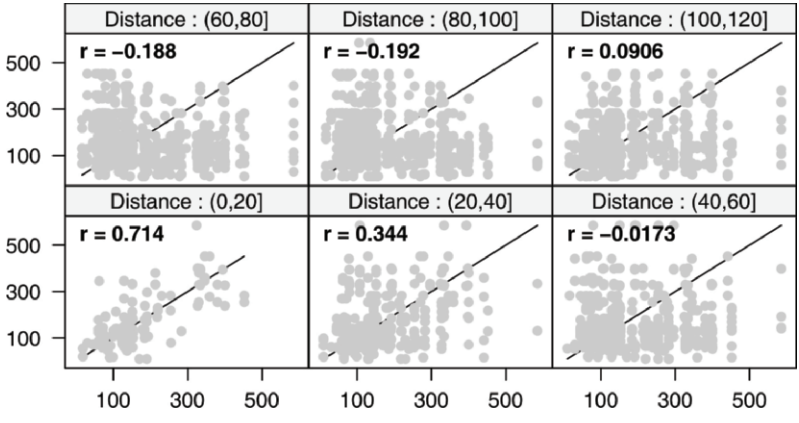

Fig. B.2.11. *h*-scatterplots: scatterplots of pairs of observed values conditioned on distance; sample correlations shown in panels

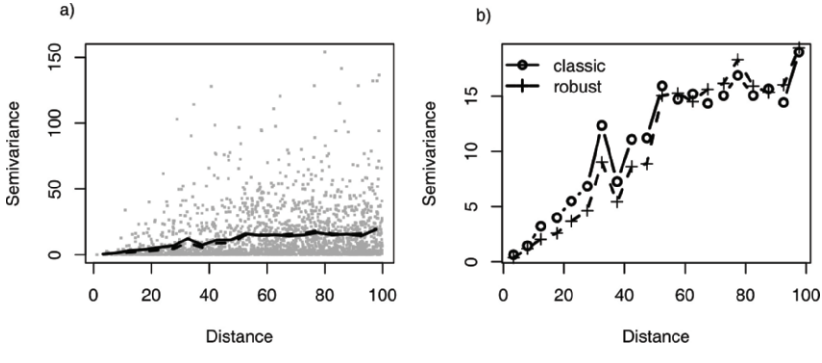

Fig. B.2.12. Swiss precipitation data – binned classic and robust variogram values: a) variogram cloud display; b) variogram values [note that the vertical axis is not in the same scale in a) and b)]

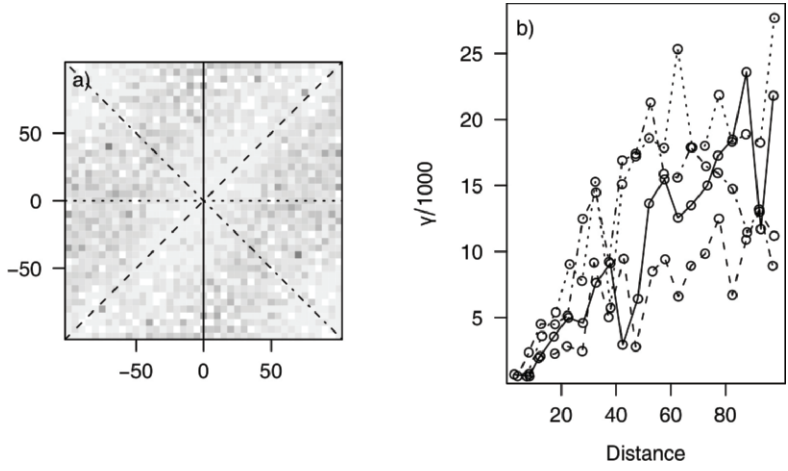

Fig. B.2.13. Detecting directionality in the variogram of Swiss precipitation data: a) variogram map showing binned semivariance values by direction and distance; b) classical variograms for four axes at 0°, 45°, 90° and 135°

B.2.5 Exploring areal data

Much of the literature on exploratory spatial data analysis has focussed on the exploration of areal data with respect to spatial association. In this section, we will look at local indicators of spatial association within this tradition, but will also consider how larger scale regularities may be revealed by using median polish smoothing and Moran eigenvector mapping. A topical area that has not been given enough attention is that of regression diagnostics for fitted spatial regression models (Haining 1994); while users appear to want heteroskedasticity-corrected standard errors, few seem to realise that the mis-specification could arguably be better handled if diagnostic methods had been used (see also Mur and Lauridsen 2007).

Median polish smoothing. Cressie (1993, pp.46-48, pp.393-400) discusses in some detail how smoothing may be used to partition the variation in the data into smooth and rough. Initial use of median polish smoothing is described by Cox and Jones (1981). In order to try it out on the North Carolina SIDS data set, we will use a coarse gridding into four columns and four rows given by Cressie (1993, pp.553-554), where four grid cells are empty; these are given by variables `L_id` and `M_id` in object `nc.sids`. Next we aggregate the number of live births and the number of SIDS cases 1974-1978 for the grid cells.

```
> L_id <- factor(nc.sids$L_id)
> M_id <- factor(nc.sids$M_id)
> both <- interaction(L_id, M_id)
> mBIR74 <- tapply(nc.sids$BIR74, both, sum)
> mSID74 <- tapply(nc.sids$SID74, both, sum)
```

Using the same Freeman-Tukey transformation as is used for the county data, we coerce the data into a correctly configured matrix, some of the cells of which are empty. The `medpolish` function is applied to the matrix, being told to remove empty cells; the function iterates over the rows and columns of the matrix using `median` to extract an overall effect, row and column effects, and residuals.

```
> mFT <- sqrt(1000) * (sqrt(mSID74/mBIR74) + sqrt((mSID74 +
+     1)/mBIR74))
> mFT1 <- t(matrix(mFT, 4, 4, byrow = TRUE))
> med <- medpolish(mFT1, na.rm = TRUE, trace.iter = FALSE)
> med

Median Polish Results (Dataset: 'mFT1')

Overall: 2.909650

Row Effects:
[1]  -0.05686791    -0.37236370     0.05686791      0.79541774

Column Effects:
[1]  -0.005484562   -0.446250551    0.003656375     0.726443256

Residuals:
        [,1]        [,2]            [,3]            [,4]
[1,] NA          -0.45800        0.000000        0.37556
[2,] -0.092554    0.00000        0.101695        0.00000
[3,] 0.092554     0.30464        -0.090726       -0.55364
[4,]  NA          NA              0.000000        NA
```

Returning to the factors linking rows and columns to counties, and generating matrices of dummy variables using `model.matrix`, we can calculate fitted values of the Freeman-Tukey adjusted rate for each county, and residuals by subtracting the fitted value from the observed rate. Naturally, the fitted value will be the same for counties in the same grid cell.

```
> mL_id <- model.matrix(~L_id - 1)
> mM_id <- model.matrix(~M_id - 1)
> nc.sids$pred <- c(med$overall + mL_id %*% med$row + mM_id %*%
+     med$col)
> nc.sids$mp_resid <- nc.sids$ft.SID74 - nc.sids$pred
> nc.sids$ft.SID74_c <- scale(nc.sids$ft.SID74, scale = FALSE)
> nc.sids$pred_c <- scale(nc.sids$pred, scale = FALSE)
```

Figure B.2.14 shows the median polish smoothing results as three maps, the observed Freeman-Tukey transformed SIDS rates, the fitted smoothed values, and the residuals.

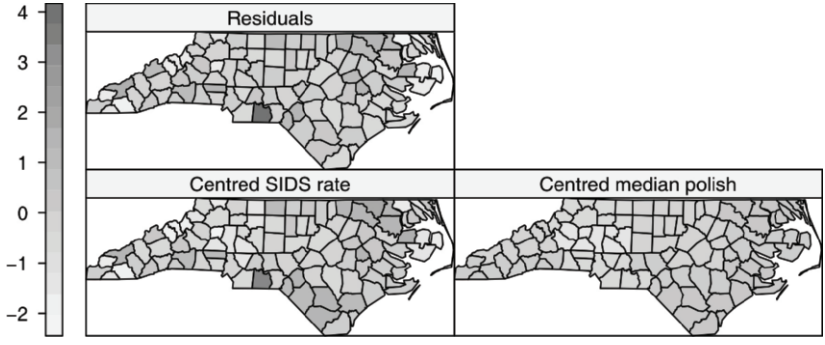

Fig. B.2.14. Median polish for North Carolina SIDS data – the Freeman-Tukey transformed SIDS rates and fitted smoothed values are mean-centred to use the same scale as the residuals

Local indicators of spatial association (*LISA*).While global measures permit us to test for spatial patterning over the whole study area, it may be the case that there is significant autocorrelation in only a smaller section, which is swamped in the context of the whole. Both distance statistics (Getis and Ord 1992, 1996; Ord and Getis 1995), and the local indicators of spatial association derived by Anselin (1995), resemble passing a moving window across the data, and examining dependence within the chosen region for the site on which the window is centred. The specifications for the window can vary, using perhaps contiguity or distance at some spatial lag from the considered zone or point.

There are clear connections here both to the study of point patterns – although methods for boundary correction have not been specifically added to weighting matrix definitions yet – and to geostatistics, since these statistics have application to the exploration of non-homogeneities in relationships between locations across the study area. They are however subject to a correlation problem when cast in a hypothesis testing framework, that estimated values of the local indicator for neighbouring zones or sites will be correlated with each other because they are necessarily calculated from many of the same values, recalling that neighbouring placements of the moving window will most likely overlap. Ord and Getis (1995) provide suitable adjustments to critical values of the G_i and G_i^* statistics. De Castro and Singer (2006) provide further developments for the appropriate handling of the false discovery rate.

The uses to which local statistics have been put are to identify 'hot-spots', to assess stationarity prior to the use of methods assuming that the data do conform to this assumption, and other checks for heterogeneity in the data series (Getis and Ord 1996). A thorny problem is that local indicators do pick up global patterns if they are present for whatever reason (Ord and Getis 2001). Measures of spatial autocorrelation are discussed in more detail in Chapter B.3.

Implementations of LISA techniques can be found in GeoDa (Anselin et al. 2006), in SAM (Rangel et al. 2006), and in the spatial statistics toolbox of Arc

GIS™, as well as the R versions discussed below (Bivand 2006; Bivand et al. 2008). The availability of software implementations has contributed to a wave of applications in many scientific domains. Scanning just the last two years, it appears that one key application area is in sociology and social policy, ranging from social medicine and fertility (Crighton et al. 2007; Schmertmann et al. 2008), through child care (Anselin et al. 2007; Freisthler et al. 2006; Lery 2008; Voss et al. 2006), to language neighbourhoods (Ishizawa and Stevens 2007), deprivation and mortality (Sridharan et al. 2007) and homicide (Ceccato et al. 2007). Another application area with many contributions is concerned with regional economic performance (Patacchini and Rice 2007; Patacchini and Zenou 2007; Yamamoto 2008), and regional and local development (Portnov 2006; Yu and Wei 2008). Penetration in other areas is also occurring, for example in local genetic structures (Sokal and Thomson 2006) and forestry (Räty and Kangas 2007).

Some but not all of the published cases using LISA techniques are exploratory. All of the papers introducing LISA techniques stress the need for caution in drawing conclusions, because apparent hotspots may rather reflect misspecification – for example the omission from the mean model of an important intermediate variable or the choice of an inappropriate functional form, because constructing tests for very small sets of neighbours even in the absence of misspecification is hard (Tiefelsdorf 2000, 2002; Bivand et al. 2009), and because of the multiple and dependent tests problem (de Castro and Singer 2006). Finally, as Waller and Gotway (2004, p.239) show, it may be necessary to create customised tests acknowledging the construction of the dependent variable, in their case using a constant risk hypothesis.

To present LISA techniques, we will return to the Guerry French moral statistics data set. To begin with, a list of contiguous neighbours is constructed, leaving Corsica with no neighbours (for details of the handling of no-neighbour observations, see Bivand and Portnov 2004).

```
> library(spdep)
> gf_cont <- poly2nb(gfrance)
```

Figure B.2.15 shows the G_i and G_i^* statistic values, scaled as standard deviates, for population per crime against property. The contiguity neighbours are converted into spatial weights using row-standardisation, after, in the G_i^* case, adding in the observations as their own neighbours.

```
> lwW <- nb2listw(gf_cont, zero.policy = TRUE)
> gfrance$local_G <- c(localG(gfrance$Pop_crime, lwW, zero.policy= TRUE))
> lwWs <- nb2listw(include.self(gf_cont))
> gfrance$local_G_star <- c(localG(gfrance$Pop_crime, lwWs))
```

Negative values show which observations are surrounded by observations with similar low values. while positive values show which observations are surrounded by observations with similar high values. Recall that high values show many inhabitants per crime, low values few inhabitants per crime. The value for Corsica,

which has no neighbour, is missing for G_i and takes a value proportional to the difference between the global mean and its own inverse crime rate for G_i^*, because then Corsica is its own only neighbour.

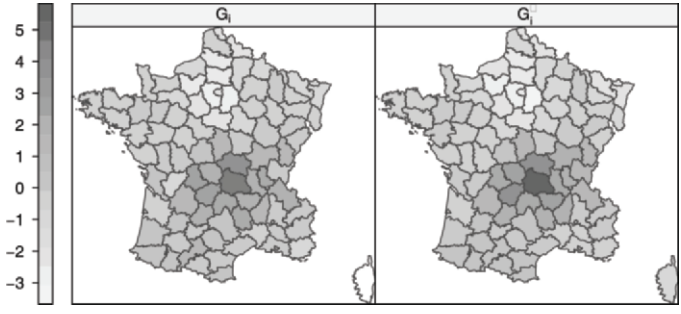

Fig. B.2.15. Local G_i and G_i^* statistics: population per crime against property, France

Since we are using G_i and G_i^* scaled as standard deviates, we will not apply them to residuals of models fitting global coefficients. The local Moran's I_i values are unscaled – they are not standard deviates, so the global Moran's I equals the mean of the local Moran's I_i values.

```
> gfrance$local_I <- localmoran(gfrance$Pop_crime, lwW,
+    zero.policy = TRUE)[, 1]
> mean(gfrance$local_I)

[1] 0.2606168

> moran.test(gfrance$Pop_crime, lwW, zero.policy = TRUE)$estimate[1]

Moran I statistic
        0.2606168
```

Since it may be the case that the local autocorrelation is driven by mis-specification, we will try two variants on the null model of treating the mean of population per crime against property as all we know. In addition to the null model, we will fit a simultaneous autoregressive model with only an intercept; the autoregressive coefficient is significant, and the model fit improves from the null baseline.

```
> C_p_esar <- spautolm(Pop_crime ~ 1, gfrance, lwW, zero.policy = TRUE,
+    method = 'Matrix')
> coef(C_p_esar)

(Intercept)    lambda
76.502332     0.470789

> gfrance$local_I_err <- localmoran(residuals(C_p_esar),
+    lwW, zero.policy = TRUE)[, 1]
```

The second variant is to fit a linear model using the percentage literacy and rank wealth variables as suggested in the conditioned choropleth map example. The coefficient for percentage literacy is negative, which – recalling that the crime rate is inverted – means that higher literacy is associated with more crime. The rank wealth coefficient is positive because lower rank means higher wealth, hence lower rank is linked to more crime.

```
> C_px_lm <- lm(Pop_crime ~ Literacy + Wealth, gfrance)
> coef(C_px_lm)

(Intercept)     Literacy     Wealth
75.7783729      -0.4569233   0.4733127

> lm.morantest(C_px_lm, lwW, zero.policy = TRUE)$estimate[1]

Observed Moran's I
         0.06888486

> gfrance$local_I_xlm <- localmoran(residuals(C_px_lm),
+     lwW, zero.policy = TRUE)[, 1]
```

This model fits the data much better than the simultaneous autoregressive null model, and, as Friendly (2007, p.396) reports, accounts for somewhat over a quarter of the variation in the dependent variable. The residuals of this model show no global autocorrelation, and a simultaneous autoregressive model with these variables included does not improve the fit. As Fig. B.2.16 shows, there is much more 'action' in the left-hand panel, where we only model the data by the mean.

Both of the areas picked out in Fig. B.2.15: the Île-de-France in the north-central part of the country with low values of the statistic, and today's Auvergne region in the south-central part of the country with high values, corresponding to values of the inverted crime rate, have higher values of Moran's I_i. Observations with intermediate values of G_i have low values of I_i, because they represent places with neighbours with inverted crime rates unlike their own. Moving to the right in Fig. B.2.16, we see that the range of shading is compressed, as the effects of misspecifation are removed. The very low value in Rhône (mid-southeast) in the map of I_i for the null model and the residuals of the simultaneous autoregressive null model is removed once the covariates are included (the large and relatively wealthy city of Lyon is atypical of its surroundings). In the map of I_i for the residuals of the linear model with covariates, Puy-de-Dôme in the Auvergne still has a large value of the statistic, suggesting that the inverse crime rate is even higher in the Auvergne than one would expect from the levels of wealth and literacy (or their absence) observed there.

We will make a LISA plot using a conditioned choropleth map, conditioning the observed Moran's I_i for the null model on factors capturing the Moran scatterplot quadrants in which the observations fall (Anselin 1996). The factors take values c ('L', 'H') depending on whether the observations are above or below the mean of the inverse crime rate, and above or below the mean of the spatial lag of the inverse crime rate.

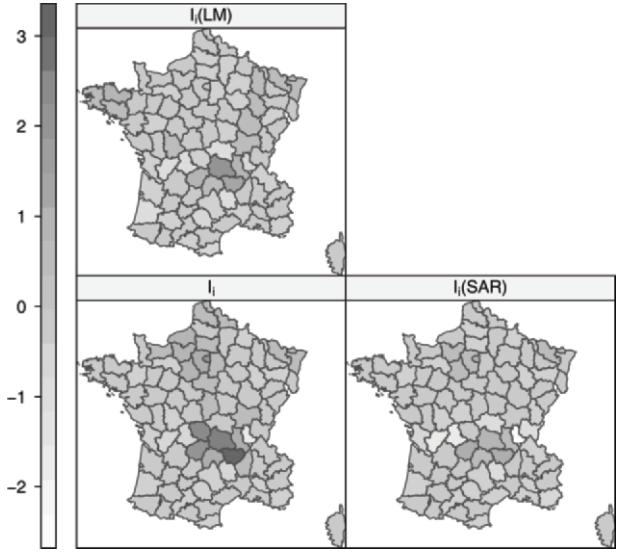

Fig. B.2.16. Local I_i statistics for the null model, the residuals of the simultaneous autoregressive model, and the residuals of the linear model including literacy and wealth: population per crime against property, France

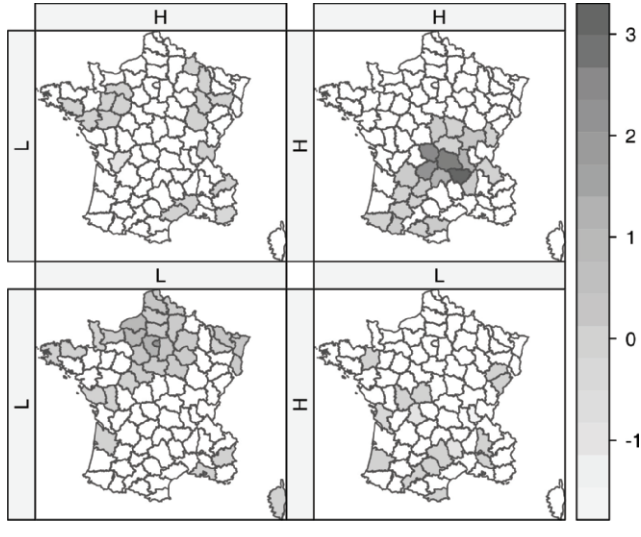

Fig. B.2.17. Conditioned choropleth LISA map: Moran's I_i for the null model conditioned on the LISA quadrant; first letter above, second letter left

Figure B.2.17 shows the split in the null model between the HH'cluster' in the Auvergne, with high values of the inverse crime rate observed for the observations and their neighbours, and the LL 'cluster' in Île-de-France, with low values of the inverse crime rate observed for the observations and their neighbours. The HL and LH panels do not display patterns that are as clear.

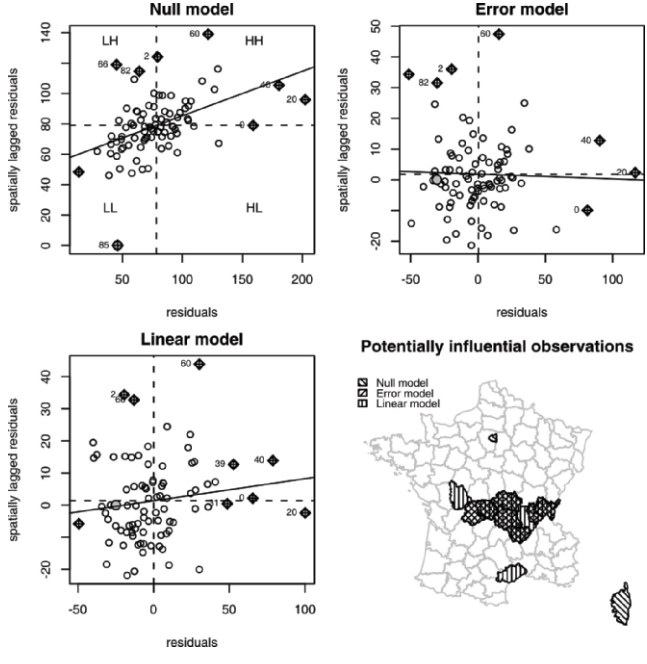

Fig. B.2.18. Moran scatterplots for a) null; b) simultaneous autoregressive; c) linear model with covariates; and d) influence map for the three models; the dashed lines divide the scatterplots into the LISA LL, HL, LH, and HH quadrants

Finally, Fig. B.2.18 shows Moran scatterplots for all three models, the null model, the simultaneous autoregressive null model, and the linear model with covariates (Anselin 1996). Interestingly, the observations found to extert influence on the linear relationship between the residuals from the models of the inverse crime rate and its spatial lag are largely the same ones across models, and form a belt stretching west and east from the Auvergne east to the Swiss border. These observations could be exerting such consistent influence because of measurement issues with the inverted crime rate, or because of remaining model mis-specification and pointing up such unusual observations is among the reasons for engaging in exploratory data analysis. Li et al. (2007) propose a approximate profile-likelihood estimator for spatial autocorrelation, which also has an ESDA extension, including a scatterplot and a local APLE measure.

```
> sPc <- scale(gfrance$Pop_crime, scale = FALSE)
> aple(sPc, lwW)
[1]   0.4810092
> aple_res <- aple.plot(sPc, lwW)
> crossprod(aple_res$Y, aple_res$X)/crossprod(aple_res$X)
         [,1]
[1,] 0.4810092
> gfrance$localAple <- localAple(sPc, lwW)
```

As Fig. B.2.19 shows, the new measure provides a view of the data that is not dissimilar to that of local Moran's I_i. The scatterplot shows that the same observations exert influence, and the map of values shows the same impact of higher positive local autocorrelation in the Auvergne and Île-de-France regions.

Two further avenues will be left unexplored here. First, it is possible that some of the problems in exploring the inverse crime rate come from the greater uncertainty of rate estimates for observations with small populations, and using an Empirical Bayes smoothing procedure may be appropriate. Second, the crime count with a log population offset term could be modelled using Poisson regression, and the deviance or Pearson residuals explored for spatial patterning.

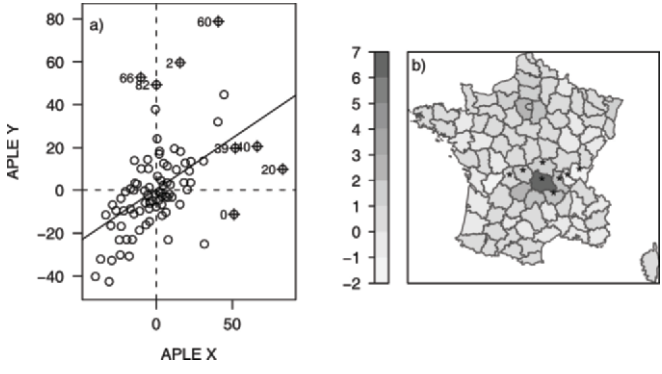

Fig. B.2.19. APLE plot and local APLE values for the population per crime rate:
a) approximate profile-likelihood estimator plot, showing observations with influence;
b) local APLE values, with observations with influence marked by asterisks

Scale. There are close relationships between the graph structure of spatial weights, and the structure exposed by examining the eigenfunctions of a centred weights matrix (Griffith 2003; Tiefelsdorf 2000), relationships underlying the understanding of Moran's *I*. It has been suggested by Griffith (2003) that maps of eigenvectors may be used to explore the effect of scale, because some eigenvectors will show large scale structures, others will capture regional differences, and others again will represent small scale patterns. Naturally, the choice of a different spatial weights matrix may give a different view on patterning at different spatial scales.

```
> SF1 <- SpatialFiltering(Pop_crime ~ 1, data = gfrance,
+ nb = gf_cont, style = 'W', zero.policy = TRUE, tol = 0.5,
+ verbose = FALSE)
> SF2 <- SpatialFiltering(Pop_crime ~ Literacy + Wealth,
+    data = gfrance, nb = gf_cont, style = 'W', zero.policy = TRUE,
+    tol = 0.5, verbose = FALSE)
```

Here we show the eigenvector maps for the eigenvectors chosen by semparametric spatial filtering for the null model and the linear model with covariates (Tiefelsdorf and Griffith 2007). Figure B.2.20 shows the six eigenvectors chosen to remove the residual spatial autocorrelation from the null model. The first eigenvector chosen is shown on the upper left, and displays a smooth, almost linear trend. The next two chosen on the upper row show regional patterns, something like quadratic and cubic trend surfaces. On the lower row, the chosen eigenvectors pick up smaller scale patterns.

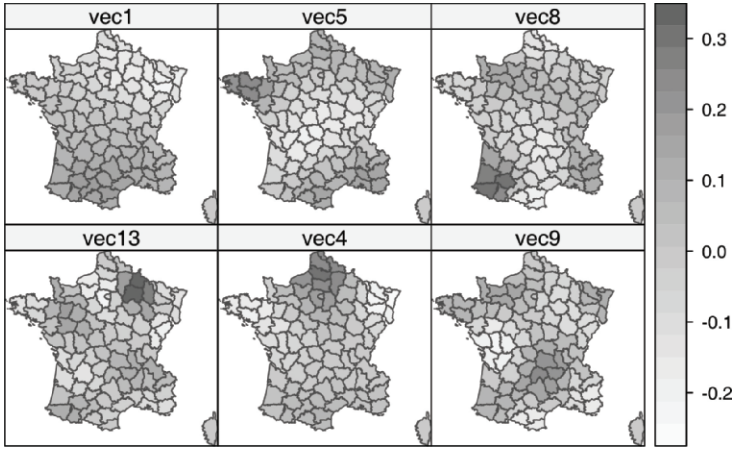

Fig. B.2.20. Six eigenvector maps for eigenvectors: null model

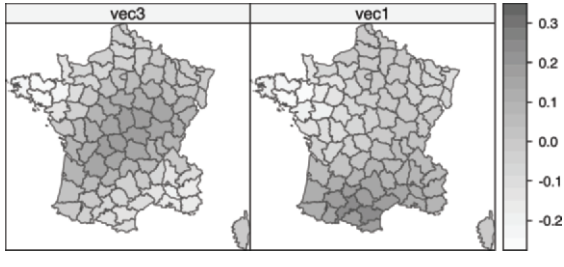

Fig. B.2.21. Two eigenvector maps for eigenvectors: linear model with covariates

Figure B.2.21 shows the two eigenvectors chosen to remove the residual spatial autocorrelation from the linear model with covariates. Because the same palette is used in Figs. B.2.20 and B.2.21, we can see how much of the residual autocorrelation has been removed by the covariates. Note that the eigenvectors differ because they are centred using the model projection matrices, so that their maps are not the same. Perhaps the patterning remaining in the linear model with covariates residuals signals that not all the mis-specification has been removed.

Geographically weighted approaches. Non-stationarity is a further source of misspecification, such as omitted variables or inappropriate functional forms. It may be approached through geographical weighting, passing a kernel with a given bandwidth over the map of data points in order to compute weighted regressions at fit points. The weights are proportional to the distances between the data points and the fit points (Brunsdon et al. 1998; Fotheringham et al. 2002). A change of support is involved, because the observation polygons are replaced by the polygon centroids, here both for the data points and the fit points.

```
> library(spgwr)
> GWfrance_bw100km1 <- gw.cov(gfrance, 'Pop_crime', bw = 1e+05,
+     cor = FALSE)
```

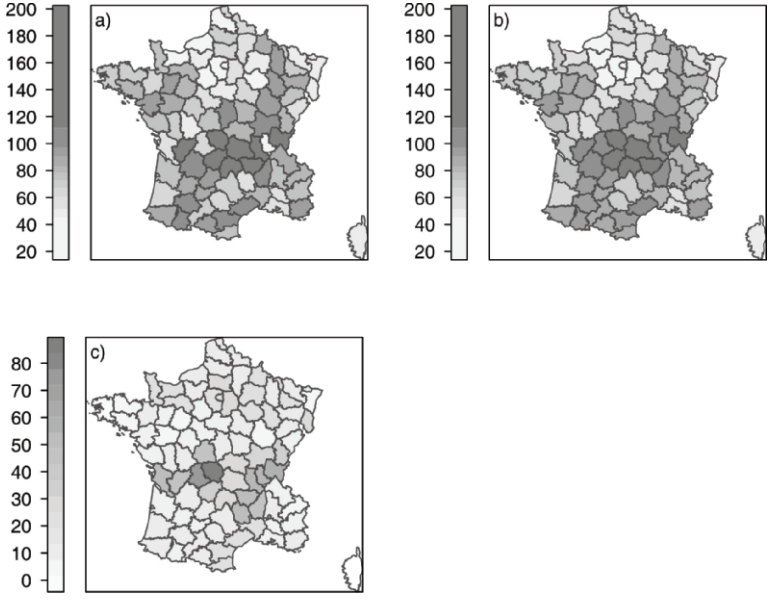

Fig. B.2.22. Population per crime against property: a) population per crime against property; b) geographically weighted means; and c) geographically weighted standard deviations

Taking a bandwidth of 100km and the default Gaussian kernel, we can calculate geographically weighted measures for the inverted crime rate (Dykes and Brunsdon 2007). Figure B.2.22 repeats the map of the inverted crime rate for reference, and shows the input variable and its geographically weighted mean using the same class intervals and palette. A smaller bandwidth would have yielded less smoothing, a larger bandwidth more, as Dykes and Brunsdon (2007) visualize.

Turning to the geographically weighted standard deviations, there seems to be some patterning, with observations apparently very unlike their neighbours being highlighted. However, recall that we are dealing with a rate variable, population per crime against property, where our confidence about the rate estimate should be related to population size. Figure B.2.23 shows a map of geographically weighted standard deviations for the chosen bandwidth conditioned on a shingle of the 1831 population. Although the picture is not very clear, it does seem that some of the observations with smaller populations have larger geographically weighted standard deviations. Obvious exceptions are the observations including the large cities of Lyon and Bordeaux, which were not like their rural neighbours in the first half of the Nineteenth century.

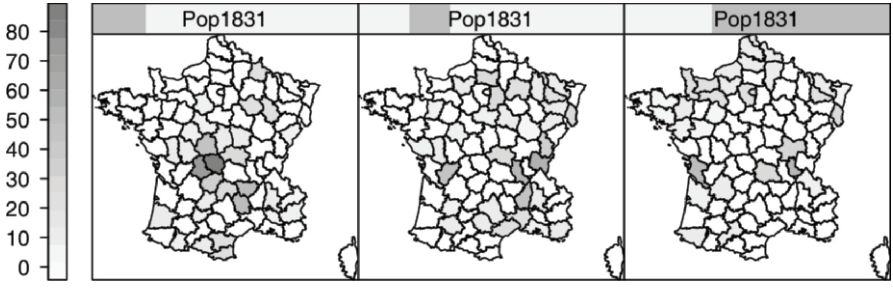

Fig. B.2.23. Conditioned choropleth map of the geographically weighted standard deviation on the inverted crime rate, conditioned on population size

Geographically weighted regression. Extending the geographically weighted approach to geographically weighted regression, we can fit our linear model with covariates using the same bandwidth and support.

```
> GWfrance_bw100km <- gwr(Pop_crime ~ Literacy + Wealth,
+     data = gfrance, bandwidth = 1e+05, hatmatrix = TRUE)
```

Figure B.2.24 shows maps of the geographically weighted regression coefficients and the coefficient of determination. As Wheeler and Tiefelsdorf (2005) point out, the GW coefficients may be highly negatively correlated with each other, as we see is the case between the intercept term and the percent literacy coefficient – the maps are almost mirror images of each other. It may be helpful to refer back to the maps of the variables shown in Fig. B.2.5; there are some similarities in spa-

tial patterning between the covariates and the geographically weighted regression coefficients, given smoothing by the kernel employed. Since collinearity is present, it is hard to conclude unequivocally that the variation in the geographically weighted regression coefficients demonstrates non-stationarity, although it is very possible that the present linear model with covariates remains mis-specified.

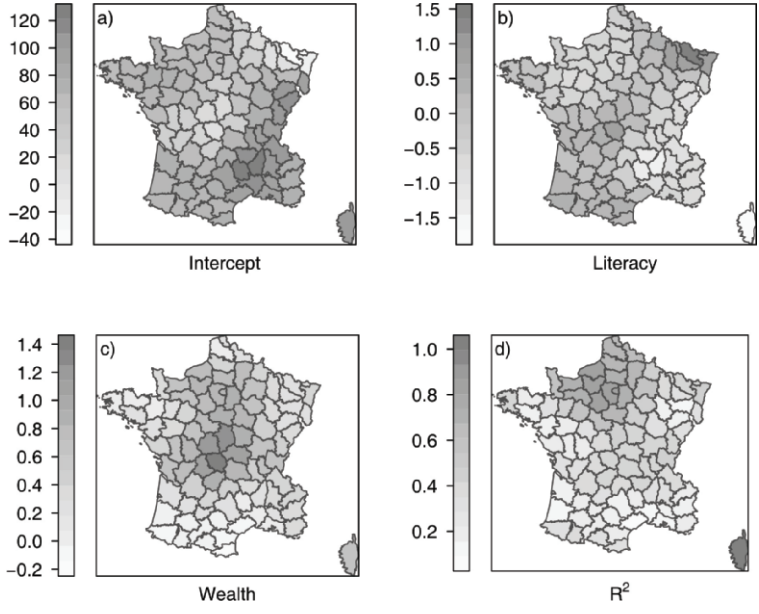

Fig. B.2.24. Maps of geographically weighted regression coefficients; a) intercept; b) percent literacy; c) rank wealth; and d) the coefficient of determination

Finally, as earlier, we also have a problem with Corsica, which had no contiguous spatial neighbours, and which here has almost no weight on any other observation for this bandwidth and kernel (`sum.w`).

```
> Corse <- which(gfrance$Department == 'Corse')
> as(GWfrance_bw100km$SDF, 'data.frame')[, c(1:5)][Corse,
+     ]

     sum.w X.Intercept. Literacy Wealth R2
85   1.032410   101.5770   -1.67253   0.7115072   0.9971373

> sapply(as(GWfrance_bw100km$SDF, 'data.frame')[, c(1:5)],
+     rank)[Corse, ]

     sum.w   X.Intercept.   Literacy   Wealth   R2
        1          77           1        64     86
```

It has extreme local coefficient values (shown by value and rank here) and a coefficient of determination of close to unity, which, although unimportant in themselves, do affect the visualization by stretching the range of values to be displayed. The use of an adaptive kernel perhaps have helped, but may make the interpretation of the output more complex.

B.2.6 Concluding remarks

This chapter should by now have shown that there are many EDA, geovisualization, and ESDA tools and techniques, and that many are implemented and available. There are however still two issues to be addressed: the tendency for *exploratory* analysis – looking for the 'right' question – to slide into inference, be it formalised or not, without considering the implications. In some cases, it can lead to the insertion of a kind of geographical particularism into our understanding of data generation processes. This is unfortunate, because it implies that our understanding of phenomena of interest is dominated by spatially structured (and/or unstructured) random effects, that the undocumented spatial autocorrelation is at the centre of our endeavours.

The second issue was taken up in the introduction: the assumption that the analyst does want to find the 'right' question. Krivoruchko and Bivand (2009, p.17) have discussed the wide range of user motivations encountered: '*In some cases, users are neither able to make nor interested in making an appropriate choice of method … In other cases, users are more like developers, working much more closely with the software in writing scripts and macros, and in trying out new models.*'

This suggests that the problem may be addressed by making the methods easier to use, by documenting them better, and offering training. It may additionally mean drawing attention to the possible benefits of doing the analysis at hand responsibly, something which is far from simple in check-box organisations, or even when academic supervisors or referees impose their views on analyses rather than empower the analyst to move towards a better question. It is not a coincidence that many early publications on EDA appeared in newsletters concerned with the teaching of statistics and data analysis.

Perhaps it is the case that using EDA and ESDA may not get you tenure quickly, getting to right questions takes time, luck, experience, and often participation in a scientific community willing to share insights and advice. On the other hand, when the research questions actually do matter, improving the way that they are framed is not a trivial achievement, and it is this that is the purpose of exploratory data analysis.

Acknowledgements. The author would like to thank the editors, an anonymous referee, and participants at a spatial statistics session at the 55th North American Meetings of the Regional Science Association International, Brooklyn, November 2008, for helpful comments and suggestions for improvements.

References

Andrienko GL, Andrienko NV (1999) Interactive maps for visual data exploration. Int J Geogr Inform Sci 13(4):355-374

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

Anselin L (1996) The moran scatterplot as an esda tool to assess local instability in spatial association. In Fischer MM, Scholten HJ, Unwin D (eds) Spatial analytical perspectives on GIS. CRC Press (Taylor and Francis Group), Boca Raton [FL], London and New York, pp.111-125

Anselin L (1998) Exploratory spatial data analysis in a geocomputational environment. In Longley PA, Brooks SM, McDonnell R, MacMillan W (eds) Geocomputation: a primer.Wiley, New York, Chichester, Toronto and Brisbane, pp.77-94

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

Anselin L, Sridharan S, Gholston S (2007) Using exploratory spatial data analysis to leverage social indicator databases: the discovery of interesting patterns. Soc Ind Res 82(2):287-309

Baddeley A, Turner R (2005) spatstat: An R package for analyzing spatial point patterns. J Stat Software 12(6):1-42

Baddeley A, Turner R, Möller J, Hazelton M (2005) Residual analysis for spatial point processes (with discussion). J Roy Stat Soc B67(5):617-666

Bailey TC, Gatrell AC (1995) Interactive spatial data analysis. Longman, Harlow

Becker RA, Cleveland WS, Shyu MJ (1996) The visual design and control of trellis display. J Comput Graph Stat 5(2):123-155

Becker RA, Cleveland WS, Wilks AR (1987) Dynamic graphics for data analysis. Stat Sci 2(4):355-383

Bivand RS (2006) Implementing spatial data analysis software tools in R. Geogr Anal 38(1):23-40

Bivand RS, Portnov BA (2004) Exploring spatial data analysis techniques using R: the case of observations with no neighbours. In Anselin L, Florax RJGM, Rey SJ (eds) Advances in spatial econometrics: methodology, tools, applications. Springer, Berlin, Heidelberg and New York, pp.121-142

Bivand RS, Müller W, Reder M (2009) Power calculations for global and local Moran's *I*. Comput Stat Data Anal 53(8):2859-2872

Bivand RS, Pebesma EJ, Gómez-Rubio V (2008) Applied spatial data analysis with R. Springer, Berlin, Heidelberg and New York

Boots B (2006) Local configuration measures for categorical spatial data: binary regular lattices. J Geogr Syst 8(1):1-24

Brewer CA, Pickle L (2002) Comparison of methods for classifying epidemiological data on choropleth maps in series. Ann Assoc Am Geogr 92(4):662-681

Brewer CA, MacEachren AM, Pickle LW, Herrmann DJ (1997) Mapping mortality: evaluating color schemes for choropleth maps. Ann Assoc Am Geogr 87(3):411-438

Brunsdon C (1998) Exploratory spatial data analysis and local indicators of spatial association with XLISP-STAT. The Statistician 47(3):471-484

Brunsdon C, Fotheringham AS, Charlton M (1998) Geographically weighted regression – modelling spatial non-stationarity. The Statistician 47(3):431-443

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

Carr DB, White D, MacEachren A (2005) Conditioned choropleth maps and hypothesis generation. Ann Assoc Am Geogr 95(1):32-53

de Castro MC, Singer BH (2006) Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. Geogr Anal 38(2):180-208

Ceccato V, Haining R, Kahn T (2007) The geography of homicide in Sao Paulo, Brazil. Environm Plann A39(7):1632-1653

Chambers JM (2008) Software for data analysis: programming with R. Springer, New York

Cleveland WS (1993) Visualizing data. Hobart Press, Summit [NJ]

Cook D, Swayne DF (2007) Interactive and dynamic graphics for data analysis. Springer, Berlin, Heidelberg and New York

Cook D, Majure J, Symanzik J, Cressie NAC (1996) Dynamic graphics in a GIS: exploring and analyzing multivariate spatial data using linked software. Comput Stat 11(4):467-480

Cook D, Symanzik J, Majure J, Cressie NAC (1997) Dynamic graphics in a GIS: more examples using linked software. Comput Geosci 23(4):371-385

Cox NJ, Jones K (1981) Exploratory data analysis. In Wrigley N, Bennett RJ (eds) Quantitatve geography. Routledge and Kegan Paul, London, pp.135-143

Cressie NAC (1993) Statistics for spatial data (revised edition). Wiley, New York, Chichester, Toronto and Brisbane

Crighton EJ, Elliott SJ, Moineddin R, Kanaroglou P, Upshur REG (2007) An exploratory spatial analysis of pneumonia and influenza hospitalizations in Ontario by age and gender. Epidemi Infect 135(2):253-261

Diggle PJ (2003) Statistical analysis of spatial point patterns (2nd edition). Arnold, London

Diggle PJ, Ribeiro PJR (2007) Model-based geostatistics. Springer, Berlin, Heidelberg and New York

Dorling D (1993) Map design for Census mapping. Cartogr J 30(2):167-183

Dorling D (1995) Visualizing changing social-structure from a Census. Environm Plann A27(3):353-378

Durham H, Dorling D, Rees P (2006) An online Census atlas for everyone. Area 38(3):336-341

Dykes JA (1997) Exploring spatial data representation with dynamic graphics. Comput Geosci 23(4):345-370

Dykes JA (1998) Cartographic visualization: exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv. The Statistician 47(3):485-497

Dykes JA, Brunsdon C (2007) Geographically weighted visualization: interactive graphics for scale-varying exploratory analysis. IEEE Transact Visual Comput Graph 13(6):1161-1168

Dykes JA, Mountain D (2003) Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and applications. Comput Stat Data Anal 43(4):581-603

Fischer MM, Stumpner P (2009) Income distribution dynamics and cross-region convergence in Europe. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.599-627

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

Freisthler B, Lery B, Gruenewald PJ, Chow J (2006) Methods and challenges of analyzing spatial data for social work problems: the case of examining child maltreatment geographically. Soct Work Rest 30(4):198-210

Friendly M (2007) A.-M. Guerry's moral statistics of France: challenges for multivariable spatial analysis. Stat Sci 22(3):368-399

Gahegan M (1999) Four barriers to the development of effective exploratory visualisation tools for the geosciences. Int J Geogr Inform Sci 13(4):289-309

Getis A (2009) Spatial Autocorrelation. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.255-279

Getis A, Ord JK (1992) The analysis of spatial association by the use of distance statistics. Geogr Anal 24(2):189-206

Getis A, Ord JK (1996) Local spatial statistics: an overview. In Longley P, Batty M (eds) Spatial analysis: modelling in a GIS environment. GeoInformation International, Cambridge, pp.261-277

Gómez-Rubio V, Ferrándiz-Ferragud J, López-Quílez A (2005) Detecting clusters of disease with *R*. J Geogr Syst 7(2):189-206

Griffith DA (2003) Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer, Berlin, Heidelberg and New York

Haining R (1994) Diagnostics for regression modeling in spatial econometrics. J Reg Sci 34(3):325-341

Haining RP (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge

Haslett J (1992) Spatial data-analysis challenges. The Statistician 41(3):271-284

Haslett J, Bradley R, Craig P, Unwin A, Wills G (1991) Dynamic graphics for exploring spatial data with application to locating global and local anomalies. Am Stat 45(3):234-242

Ishizawa H, Stevens G (2007) Non-english language neighborhoods in Chicago, Illinois, 2000. Soc Sci Res 36(3):1042-1064

Jacoby WG (1997) Statistical graphics for univariate and bivariate data. Sage, Thousand Oaks [CA]

Kaluzny SP, Vega SC, Cardoso TP, Shelly AA (1998) S+SpatialStats, user manual for Windows and UNIX. Springer, Berlin, Heidelberg and New York

Krivoruchko K, Bivand R (2009) GIS, users, developers, and spatial statistics: on monarchs and their clothing. In Pilz J (ed) Interfacing geostatistics and GIS. Springer, Berlin, Heidelberg and New York, pp.203-222

Lery B (2008) A comparison of foster care entry risk at three spatial scales. Subst UseMisuse 43(2):223-237

Levine N (2006) Crime mapping and the CrimeStat program. Geogr Anal 38(1):41–56

Li H, Calder CA, Cressie NAC (2007) Beyond Moran's *I*: testing for spatial dependence based on the spatial autoregressive model. Geogr Anal 39(4):357-375

Lloyd CD (2007) Local models for spatial analysis. CRC Press (Taylor and Francis Group), Boca Raton [FL], London and New York

MacEachren A, Gahegan M, Pike W (2004a) Visualization for constructing and sharing geo-scientific concepts. Proceedings of the National Academy of Sciences of the United States of America 101 (Suppl. 1), pp.5279-5286

MacEachren A, Gahegan M, Pike W, Brewer I, Cai G, Lengerich E, Hardisty F (2004b) Geovisualization for knowledge construction and decision support. IEEE Comp Graph Appl 24(1):13-17

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

Müller W (2007) Collecting spatial data. Springer, Berlin, Heidelberg and New York

Mur J, Lauridsen J (2007) Outliers and spatial dependence in cross-sectional regressions. Environ Plann A39(7):1752-1769

Murrell P (2005) R Graphics. CRC Press (Taylor and Francis Group), Boca Raton [FL], London and New York

Oliver M (2009) The variogram and kriging. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.319-352

Ord JK, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and an application. Geogr Anal 27(3):286-306

Ord JK, Getis A (2001) Testing for local spatial autocorrelation in the presence of global autocorrelation. J Reg Sci 41(3):411-432

Patacchini E, Rice P (2007) Geography and economic performance: exploratory spatial data analysis for Great Britain. Reg Stud 41(4):489-508

Patacchini E, Zenou Y (2007) Spatial dependence in local unemployment rates. J Econ Geogr 7(2):169-191

Pebesma E (2004) Multivariable geostatistics in S: the gstat package. Comput Geosci 30(7):683-691

Portnov BA (2006) Urban clustering, development similarity, and local growth: a case study of Canada. Europ Plann Stud 14(9):1287-1314

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org, ISBN 3-900051-07-0

Rangel TFLVB, Diniz-Filho JAF, Bini LM (2006) Towards an integrated computational tool for spatial analysis in macroecology and biogeography. Glob Ecol Biogeogr 15(4):321-327

Räty M, Kangas A (2007) Localizing general models based on local indices of spatial association. Europ J Forest Res 126(2):279-289

Sarkar D (2007) Lattice multivariate data visualization with R. Springer, Berlin, Heidelberg and New York

Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis. CRC Press (Taylor and Francis Group), Boca Raton [FL], London and New York

Schmertmann CP, Potter JE, Cavenaghi SM (2008) Exploratory analysis of spatial patterns in Brazil's fertility transition. Popul Res Pol Rev 27(1):1-15

Slocum TA, McMaster RB, Kessler FC, Howard HH (2005) Thematic cartography and geographical visualization. Prentice-Hall, Upper Saddle River [NJ]

Sokal R, Thomson B (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. Am J Phys Anthr 129(1):121-131

Sridharan S, Tunstall H, Lawder R, Mitchell R (2007) An exploratory spatial data analysis approach to understanding the relationship between deprivation and mortality in Scotland. Soc Sci Med 65(9):1942-1952

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

Takatsuka M, Gahegan M (2002) GeoVISTA studio: a codeless visual programming environment for geoscientific data analysis and visualization. Comput Geosci 28(10):1131-1144

Theus M (2002) Interactive data visualization using mondrian. J Stat Software 7(11):1-9

Tiefelsdorf M (2000) Modelling spatial processes: the identification and analysis of spatial relationships in regression residuals by means of Moran's *I*. Springer, Berlin, Heidelberg and New York

Tiefelsdorf M (2002) The saddlepoint approximation of Moran's *I* and local Moran's I_i reference distributions and their numerical evaluation. GeogrAnal 34(3):187-206

Tiefelsdorf M, Griffith DA (2007) Semiparametric filtering of spatial autocorrelation: the eigenvector approach. Environ Plann A39(5):1193-1221

Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading [MA]

Unwin A (1996) Exploratory spatial analysis and local statistics. Comput Stat 11(4):387-400

Unwin DJ, Wrigley N (1987) Towards a general-theory of control point distribution effects in trend surface models. Comput Geosci 13(4):351-355

Velleman P, Hoaglin D (1981) The ABC's of EDA: applications, basics, and computing of exploratory data analysis. Duxbury, Boston

Voss PR, Long DD, Hammer RB, Friedman S (2006) County child poverty rates in the US: a spatial regression approach. Popul Res Pol Rev 25(4):369-391

Waller LA, Gotway CA (2004) Applied spatial statistics for public health data. Wiley, New Jersey

Wheeler DC, Tiefelsdorf M (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. J Geogr Syst 7(2):161-187

Wood J, Dykes J, Slingsby A, Clarke K (2007) Interactive visual exploration of a large spatio-temporal dataset: reflections on a geovisualization mashup. IEEE Transact Visual Compu Graph 13(6):1176-1183

Yamamoto D (2008) Scales of regional income disparities in the USA, 1955- 2003. J Econ Geogr 8(1):79-103

Yu D, Wei YD (2008) Spatial data analysis of regional development in Greater Beijing, China, in a GIS environment. Papers in Reg Sci 87(1):97-117