1

Hello World: Introducing Spatial Data

1.1 Applied Spatial Data Analysis

Spatial data are everywhere. Besides those we collect ourselves ('is it raining?'), they confront us on television, in newspapers, on route planners, on computer screens, and on plain paper maps. Making a map that is suited to its purpose and does not distort the underlying data unnecessarily is not easy. Beyond creating and viewing maps, spatial data *analysis* is concerned with questions not directly answered by looking at the data themselves. These questions refer to hypothetical processes that generate the observed data. Statistical inference for such spatial processes is often challenging, but is necessary when we try to draw conclusions about questions that interest us.

Possible questions that may arise include the following:

- Does the spatial patterning of disease incidences give rise to the conclusion that they are clustered, and if so, are the clusters found related to factors such as age, relative poverty, or pollution sources?
- Given a number of observed soil samples, which part of a study area is polluted?
- Given scattered air quality measurements, how many people are exposed to high levels of black smoke or particulate matter (e.g. PM_{10}),[1] and where do they live?
- Do governments tend to compare their policies with those of their neighbours, or do they behave independently?

In this book we will be concerned with *applied* spatial data analysis, meaning that we will deal with data sets, explain the problems they confront us with, and show how we can attempt to reach a conclusion. This book will refer to the theoretical background of methods and models for data analysis, but emphasise hands-on, do-it-yourself examples using R; readers needing this background should consult the references. All data sets used in this book and all examples given are available, and interested readers will be able to reproduce them.

[1] Particulate matter smaller than about 10 µm.

In this chapter we discuss the following:

(i) Why we use R for analysing spatial data
(ii) The relation between R and geographical information systems (GIS)
(iii) What spatial data are, and the types of spatial data we distinguish
(iv) The challenges posed by their storage and display
(v) The analysis of observed spatial data in relation to processes thought to have generated them
(vi) Sources of information about the use of R for spatial data analysis and the structure of the book.

1.2 Why Do We Use R

1.2.1 ... In General?

The R system[2] (R Development Core Team, 2008) is a free software environment for statistical computing and graphics. It is an implementation of the S language for statistical computing and graphics (Becker et al., 1988). For data analysis, it can be highly efficient to use a special-purpose language like S, compared to using a general-purpose language.

For new R users without earlier scripting or programming experience, meeting a programming language may be unsettling, but the investment[3] will quickly pay off. The user soon discovers how analysis components – written or copied from examples — can easily be stored, replayed, modified for another data set, or extended. R can be extended easily with new dedicated components, and can be used to develop and exchange data sets and data analysis approaches. It is often much harder to achieve this with programs that require long series of mouse clicks to operate.

R provides many standard and innovative statistical analysis methods. New users may find access to both well-tried and trusted methods, and speculative and novel approaches, worrying. This can, however, be a major strength, because if required, innovations can be tested in a robust environment against legacy techniques. Many methods for analysing spatial data are less frequently used than the most common statistical techniques, and thus benefit proportionally more from the nearness to both the data and the methods that R permits. R uses well-known libraries for numerical analysis, and can easily be extended by or linked to code written in S, C, C++, Fortran, or Java. Links to various relational data base systems and geographical information systems exist, many well-known data formats can be read and/or written.

The level of voluntary support and the development speed of R are high, and experience has shown R to be environment suitable for developing professional, mission-critical software applications, both for the public and the

[2] `http://www.r-project.org`.
[3] A steep learning curve – the user learns a lot per unit time.

private sector. The S language can not only be used for low-level computation on numbers, vectors, or matrices but can also be easily extended with classes for new data types and analysis methods for these classes, such as methods for summarising, plotting, printing, performing tests, or model fitting (Chambers, 1998).

In addition to the core R software system, R is also a social movement, with many participants on a continuum from useRs just beginning to analyse data with R to developeRs contributing packages to the Comprehensive R Archive Network[4] (CRAN) for others to download and employ.

Just as R itself benefits from the open source development model, contributed package authors benefit from a world-class infrastructure, allowing their work to be published and revised with improbable speed and reliability, including the publication of source packages and binary packages for many popular platforms. Contributed add-on packages are very much part of the R community, and most core developers also write and maintain contributed packages. A contributed package contains R functions, optional sample data sets, and documentation including examples of how to use the functions.

1.2.2 ... for Spatial Data Analysis?

For over 10 years, R has had an increasing number of contributed packages for handling and analysing spatial data. All these packages used to make different assumptions about how spatial data were organised, and R itself had no capabilities for distinguishing coordinates from other numbers. In addition, methods for plotting spatial data and other tasks were scattered, made different assumptions on the organisation of the data, and were rudimentary. This was not unlike the situation for time series data at the time.

After some joint effort and wider discussion, a group[5] of R developers have written the R package **sp** to extend R with classes and methods for spatial data (Pebesma and Bivand, 2005). Classes specify a structure and define how spatial data are organised and stored. Methods are instances of functions specialised for a particular data class. For example, the summary method for all spatial data classes may tell the range spanned by the spatial coordinates, and show which coordinate reference system is used (such as degrees longitude/latitude, or the UTM zone). It may in addition show some more details for objects of a specific spatial class. A plot method may, for example create a map of the spatial data.

The **sp** package provides classes and methods for points, lines, polygons, and grids (Sect. 1.4, Chap. 2). Adopting a single set of classes for spatial data offers a number of important advantages:

[4] CRAN mirrors are linked from `http://www.r-project.org/`.

[5] Mostly the authors of this book with help from Barry Rowlingson and Paulo J. Ribeiro Jr.

(i) It is much easier to move data across spatial statistics packages. The classes are either supported directly by the packages, reading and writing data in the new spatial classes, or indirectly, for example by supplying data conversion between the **sp** classes and the package's classes in an interface package. This last option requires one-to-many links between the packages, which are easier to provide and maintain than many-to-many links.
(ii) The new classes come with a well-tested set of methods (functions) for plotting, printing, subsetting, and summarising spatial objects, or combining (overlaying) spatial data types.
(iii) Packages with interfaces to geographical information systems (GIS), for reading and writing GIS file formats, and for coordinate (re)projection code support the new classes.
(iv) The new methods include Lattice plots, conditioning plots, plot methods that combine points, lines, polygons, and grids with map elements (reference grids, scale bars, north arrows), degree symbols (as in 52°N) in axis labels, etc.

Chapter 2 introduces the classes and methods provided by **sp**, and discusses some of the implementation details. Further chapters will show the degree of integration of **sp** classes and methods and the packages used for statistical analysis of spatial data.

Figure 1.1 shows how the reception of **sp** classes has already influenced the landscape of contributed packages; interfacing other packages for handling and analysing spatial data is usually simple as we see in Part II. The shaded nodes of the dependency graph are packages (co)-written and/or maintained by the authors of this book, and will be used extensively in the following chapters.

1.3 R and GIS

1.3.1 What is GIS?

Storage and analysis of spatial data is traditionally done in Geographical Information Systems (GIS). According to the toolbox-based definition of Burrough and McDonnell (1998, p. 11), a GIS is '...a powerful set of tools for collecting, storing, retrieving at will, transforming, and displaying spatial data from the real world for a particular set of purposes'. Another definition mentioned in the same source refers to '...checking, manipulating, and analysing data, which are spatially referenced to the Earth'.

Its capacity to analyse and visualise data makes R a good choice for spatial data analysis. For some spatial analysis projects, using only R may be sufficient for the job. In many cases, however, R will be used in conjunction with GIS software and possibly a GIS data base as well. Chapter 4 will show how spatial data are imported from and exported to GIS file formats. As is often the case in applied data analysis, the real issue is not whether a given problem *can* be

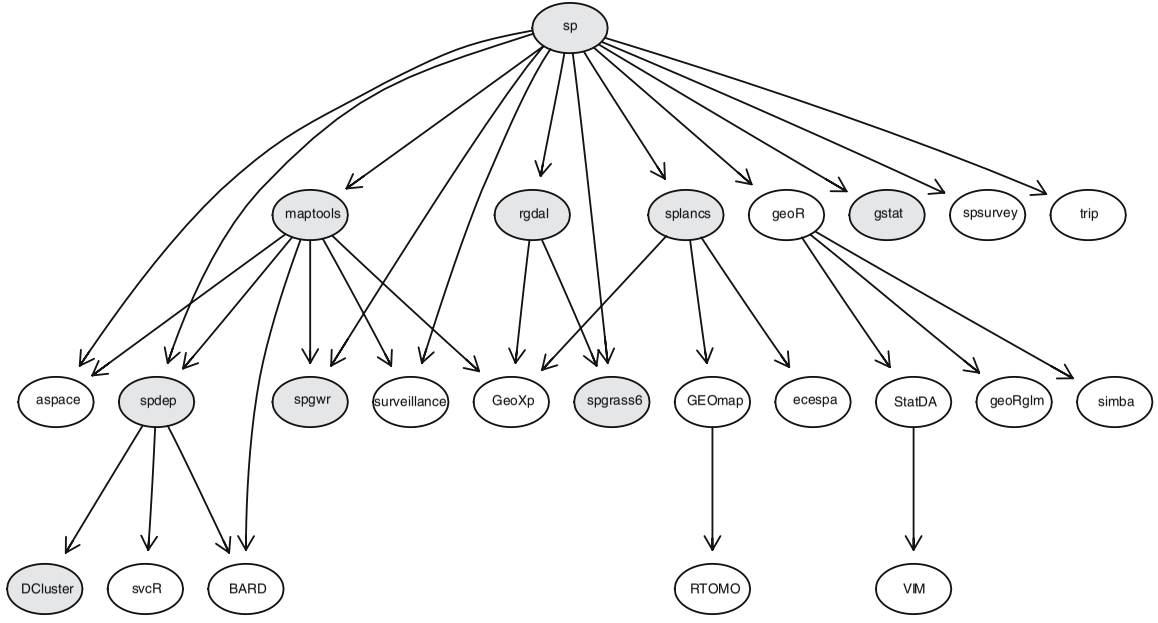

Fig. 1.1. Tree of R contributed packages on CRAN depending on or importing **sp** directly or indirectly; others suggest **sp** or use it without declaration in their package descriptions (status as of 2008-04-06)

solved using an environment such as R, but whether it can be solved *efficiently* with R. In some cases, combining different software components in a workflow may be the most robust solution, for example scripting in languages such as Python.

1.3.2 Service-Oriented Architectures

Today, much of the practice and research in geographical information systems is moving from toolbox-centred architectures (think of the 'classic' Arc/Info™ or ArcGIS™ applications) towards *service-centred* architectures (such as Google Earth™). In toolbox-centred architectures, the GIS application and data are situated on the user's computer or local area network. In service-centred architectures, the tools and data are situated on remote computers, typically accessed through Internet connections.

Reasons for this change are the increasing availability and bandwidth of the Internet, and also ownership and maintenance of data and/or analysis methods. For instance, data themselves may not be freely distributable, but certain derived products (such as visualisations or generalisations) may be. A service can be kept and maintained by the provider without end users having to bother about updating their installed software or data bases. The R system operates well under both toolbox-centred and service-centred architectures.

1.3.3 Further Reading on GIS

It seems appropriate to give some recommendations for further reading concerning GIS, not least because a more systematic treatment would not be appropriate here. Chrisman (2002) gives a concise and conceptually elegant introduction to GIS, with weight on using the data stored in the system; the domain focus is on land planning. A slightly older text by Burrough and McDonnell (1998) remains thorough, comprehensive, and perhaps a shade closer to the earth sciences in domain terms than Chrisman.

Two newer comprehensive introductions to GIS cover much of the same ground, but are published in colour. Heywood et al. (2006) contains less extra material than Longley et al. (2005), but both provide very adequate coverage of GIS as it is seen from within the GIS community today. To supplement these, Wise (2002) provides a lot of very distilled experience on the technicalities of handling geometries in computers in a compact form, often without dwelling on the computer science foundations; these foundations are given by Worboys and Duckham (2004). Neteler and Mitasova (2008) provide an excellent analytical introduction to GIS in their book, which also shows how to use the open source GRASS GIS, and how it can be interfaced with R.

It is harder to provide guidance with regard to service-centred architectures for GIS. The book by Shekar and Xiong (2008) work is a monumental, forward-looking collection with strong roots in computer and information science, and reflects the ongoing embedding of GIS technologies into database

systems far more than the standard texts. Two hands-on alternatives show how service-centred architectures can be implemented at low cost by non-specialists, working, for example in environmental advocacy groups, or volunteer search and rescue teams (Mitchell, 2005; Erle et al., 2005); their approach is certainly not academic, but gets the job done quickly and effectively.

In books describing the handling of spatial data for data analysts (looking at GIS from the outside), Waller and Gotway (2004, pp. 38–67) cover most of the key topics, with a useful set of references to more detailed treatments; Banerjee et al. (2004, pp. 10–18) give a brief overview of cartography sufficient to get readers started in the right direction.

1.4 Types of Spatial Data

Spatial data have spatial reference: they have coordinate values and a system of reference for these coordinates. As a fairly simple example, consider the locations of volcano peaks on the Earth. We could list the coordinates for all known volcanoes as pairs of longitude/latitude decimal degree values with respect to the prime meridian at Greenwich and zero latitude at the equator. The World Geodetic System (WGS84) is a frequently used representation of the Earth.

Suppose we are interested in the volcanoes that have shown activity between 1980 and 2000, according to some agreed seismic registration system. This data set consists of points only. When we want to draw these points on a (flat) map, we are faced with the problem of projection: we have to translate from the spherical longitude/latitude system to a new, non-spherical coordinate system, which inevitably changes their relative positions. In Fig. 1.2, these data are projected using a Mollweide projection, and, for reference purposes, coast lines have been added. Chapter 4 deals with coordinate reference systems, and with transformations between them.

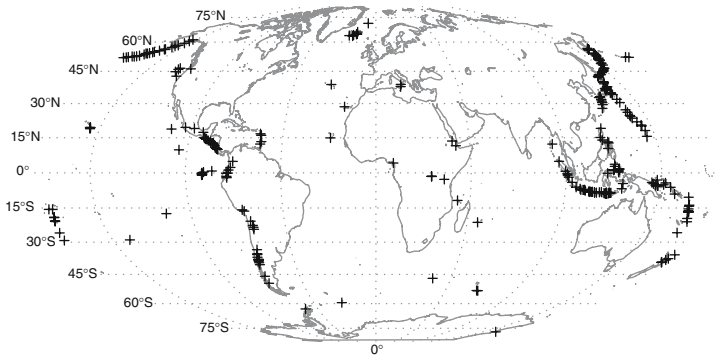

Fig. 1.2. Volcanoes of the world, with last known eruption 1964 or later (+); source: National Geophysical Data Center

If we also have the date and time of the last observed eruption at the volcano, this information is called an *attribute*: it is non-spatial in itself, but this attribute information is believed to exist for each spatial entity (volcano).

Without explicit attributes, points usually carry implicit attributes, for example all points in this map have the constant implicit attribute – they mark a 'volcano peak', in contrast to other points that do not. We represent the *purely spatial* information of entities by data models.The different types of data models that we distinguish here include the following:

Point, a single point location, such as a GPS reading or a geocoded address
Line, a set of ordered points, connected by straight line segments
Polygon, an area, marked by one or more enclosing lines, possibly containing holes
Grid, a collection of points or rectangular cells, organised in a regular lattice

The first three are vector data models and represent entities as exactly as possible, while the final data model is a raster data model, representing continuous surfaces by using a regular tessellation. All spatial data consist of positional information, answering the question 'where is it?'. In many applications these will be extended by attributes, answering the question 'what is where?'; Chrisman (2002, pp. 37–69) distinguishes a range of spatial and spatio-temporal queries of this kind. Examples for these four basic data models and of types with attributes will now follow.

The location (x, y coordinates) of a volcano may be sufficient to establish its position relative to other volcanoes on the Earth, but for describing a single volcano we can use more information. Let us, for example try to describe the topography of a volcano. Figure 1.3 shows a number of different ways to represent a continuous surface (such as topography) in a computer.

First, we can use a large number of points on a dense regular *grid* and store the attribute *altitude* for each point to approximate the surface. Grey tones are used to specify classes of these points on Fig. 1.3a.

Second, we can form contour *lines* connecting ordered points with equal altitude; these are overlayed on the same figure, and separately shown on Fig. 1.3b. Note that in this case, the contour lines were derived from the point values on the regular grid.

A *polygon* is formed when a set of line segments forms a closed object with no lines intersecting. On Fig. 1.3a, the contour lines for higher altitudes are closed and form polygons.

Lines and polygons may have *attributes*, for example the 140 contour line of Fig. 1.3a may have the label '140 m above sea level', or simply 140. Two closed contour lines have the attribute 160 m, but within the domain of this study area several non-closed contour lines have the attribute 110 m. The complete area inside the 140 m polygon (Fig. 1.3c) has the attribute 'more than 140 m above sea level', or >140. The area above the 160 m contour is represented by a polygon with a *hole* (Fig. 1.3d): its centre is part of the crater, which is below 160 m.

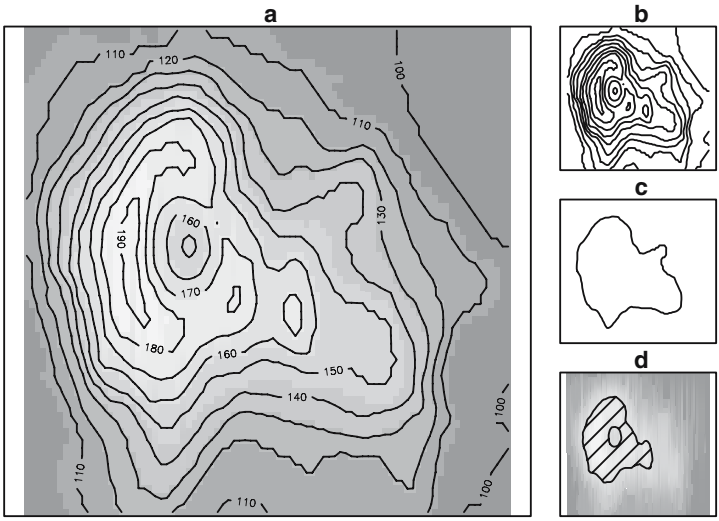

Fig. 1.3. Maunga Whau (Mt Eden) is one of about 50 volcanoes in the Auckland volcanic field. (**a**) Topographic information (altitude, m) for Maunga Whau on a $10 \times 10\,\mathrm{m}^2$ grid, (**b**) contour lines, (**c**) 140 m contour line: a closed polygon, (**d**) area above 160 m (*hashed*): a polygon with a hole

Polygons formed by contour lines of volcanoes usually have a more or less circular shape. In general, polygons can have arbitrary form, and may for certain cases even overlap. A special, but common case is when they represent the boundary of a single categorical variable, such as an administrative region. In that case, they cannot overlap and should divide up the entire study area: each point in the study area can and must be attributed to a single polygon, or lies on a boundary of one or more polygons.

A special form to represent spatial data is that of a grid: the values in each grid cell may represent an average over the area of the cell, or the value at the midpoint of the cell, or something more vague – think of image sensors. In the first case, we can see a grid as a special case of ordered points; in the second case, they are a collection of rectangular polygons. In any case, we can *derive* the position of each cell from the grid location, grid cell size, and the organisation of the grid cells. Grids are a common way to tessellate a plane. They are important because

- Devices such as digital cameras and remote sensing instruments register data on a regular grid
- Computer screens and projectors show data on a grid
- Many spatial or spatio-temporal models, such as climate models, discretise space by using a regular grid.

1.5 Storage and Display

As R is open source, we can find out the meaning of every single bit and byte manipulated by the software if we need to do so. Most users will, however, be happy to find that this is unlikely to be required, and is left to a small group of developers and experts. They will rely on the fact that many users have seen, tested, or used the code before.

When running an R session, data are usually read or imported using explicit commands, after which all data are *kept in memory*; users may choose to load a saved workspace or data objects. During an R session, the workspace can be saved to disk or chosen objects can be saved in a portable binary form for loading into the next session. When leaving an interactive R session, the question *Save workspace image?* may be answered positively to save results to disk; saving the session history is a very useful way of documenting what has been done, and is recommended as normal practice – consider choosing an informative file name.

Despite the fact that computers have greater memory capacity than they used to, R may not be suitable for the analysis of massive data sets, because data being analysed is held in memory. Massive data sets may, for example come from satellite imagery, or detailed global coast line information. It is in such cases necessary to have some idea about data size and memory management and requirements. Under such circumstances it is often still possible to use R as an analysis engine on part of the data sets. Smaller useful data sets can be obtained by selecting a certain region or by sub-sampling, aggregating or generalising the original data. Chapters 4 and 6 will give hints on how to do this.

Spatial data are usually displayed on maps, where the x- and y-axes show the coordinate values, with the aspect ratio chosen such that a unit in x equals a unit in y. Another property of maps is that elements are added for reference purposes, such as coast lines, rivers, administrative boundaries, or even satellite images.

Display of spatial data in R is a challenge on its own, and is dealt with in Chap. 3. For many users, the graphical display of statistical data is among the most compelling reasons to use R, as maps are traditionally amongst the strongest graphics we know.

The core R engine was not designed specifically for the display and analysis of maps, and the limited interactive facilities it offers have drawbacks in this area. Still, a large number of visualisations come naturally to R graphics, while they would take a substantial effort to accomplish in legacy GIS. For one thing, most GIS do not provide *conditioning plots*, where series of plots are organised in a regular lattice, share axes, and legends, and allow for systematic comparison across a large number of settings, scenarios, time, or other variables (e.g. Fig. 3.10). R provides on-screen graphics and has many graphics drivers, for example for vector graphics output to PostScript, Windows metafiles, PDF, and many bitmapped graphics formats. And, as mentioned, it works equally well as a front end or as a service providing back end for statistical analysis.

1.6 Applied Spatial Data Analysis

Statistical inference is concerned with drawing conclusions based on data and prior assumptions. The presence of a model of the data generating process may be more or less acknowledged in the analysis, but its reality will make itself felt sooner or later. The model may be manifest in the design of data collection, in the distributional assumptions employed, and in many other ways. A key insight is that observations in space cannot in general be assumed to be mutually independent, and that observations that are close to each other are likely to be similar (*ceteris paribus*). This spatial patterning – spatial autocorrelation – may be treated as useful information about unobserved influences, but it does challenge the application of methods of statistical inference that assume the mutual independence of observations.

Not infrequently, the prior assumptions are not made explicit, but are rather taken for granted as part of the research tradition of a particular scientific subdiscipline. Too little attention typically is paid to the assumptions, and too much to superficial differences; for example Venables and Ripley (2002, p. 428) comment on the difference between the covariance function and the semi-variogram in geostatistics, that '[m]uch heat and little light emerges from discussions of their comparison'.

To illustrate the kinds of debates that rage in disparate scientific communities analysing spatial data, we sketch two current issues: red herrings in geographical ecology and the interpretation of spatial autocorrelation in urban economics.

The red herring debate in geographical ecology was ignited by Lennon (2000), who claimed that substantive conclusions about the impact of environmental factors on, for example species richness had been undermined by not taking spatial autocorrelation into account. Diniz-Filho et al. (2003) replied challenging not only the interpretation of the problem in statistical terms, but pointing out that geographical ecology also involves the scale problem, that the influence of environmental factors is moderated by spatial scale.

They followed this up in a study in which the data were sub-sampled to attempt to isolate the scale problem. But they begin: 'It is important to note that we do not present a formal evaluation of this issue using statistical theory..., our goal is to illustrate heuristically that the often presumed bias due to spatial autocorrelation in OLS regression does not apply to real data sets' (Hawkins et al., 2007, p. 376).

The debate continues with verve in Beale et al. (2007) and Diniz-Filho et al. (2007). This is quite natural, as doubts about the impacts of environmental drivers on species richness raise questions about, for example, the effects of climate change. How to analyse spatial data is obviously of importance within geographical ecology. However, Diniz-Filho et al. (2007, p. 850) conclude that '[w]hen multiple assumptions are not being met, as in the case of virtually all geographical analyses, can a result from any single method (whether spatial or non-spatial) be claimed to be better? ... If different spatial

methods themselves are unstable and generate conflicting results in real data, it makes no sense to claim that any particular method is always superior to any other'.

The urban economics debate is not as vigorous, but is of some practical interest, as it concerns the efficiency of services provided by local government. Revelli (2003) asks whether the spatial patterns observed in model residuals are a reaction to model misspecification, or do they signal the presence of substantive interaction between observations in space? In doing so, he reaches back to evocations of the same problem in the legacy literature of spatial statistics. As Cliff and Ord (1981, pp. 141–142) put it, 'two adjacent supermarkets will compete for trade, and yet their turnover will be a function of general factors such as the distribution of population and accessibility'. They stress that 'the presence of spatial autocorrelation may be attributable either to trends in the data or to interactions; ... [t]he choice of model must involve the scientific judgement of the investigator *and* careful testing of the assumptions'. When the fitted model is misspecified, it will be hard to draw meaningful conclusions, and the care advised by Cliff and Ord will be required.

One way of testing the assumptions is through changes in the policy context over time, where a behavioural model predicts changes in spatial autocorrelation – if the policy changes, the level of spatial interaction should change (Bivand and Szymanski, 1997; Revelli, 2003). Alternatives include using multiple levels in local government (Revelli, 2003), or different electoral settings, such as lame-duck administrations as controls (Bordignon et al., 2003). A recent careful study has used answers to a questionnaire survey to check whether interaction has occurred or not. It yields a clear finding that the observed spatial patterning in local government efficiency scores is related to the degree to which they compare their performance with that of other local government entities (Revelli and Tovmo, 2007).

This book will not provide explicit guidance on the choice of models, because the judgement of researchers in different scientific domains will vary. One aspect shared by both examples is that the participants stress the importance of familiarity with the core literature of spatial statistics. It turns out that many of the insights found there remain fundamental, despite the passage of time. Applied spatial data analysis seems to be an undertaking that, from time to time, requires the analyst to make use of this core literature.

Without attempting to be exhaustive in reviewing key books covering all the three acknowledged areas of spatial statistics – point processes, geostatistics, and areal data – we can make some choices. Bivand (2008, pp. 16–17) documents the enduring position of Ripley (1981)[6] and Cliff and Ord (1981) in terms of paper citations. Ripley (1988) supplements and extends the earlier work, and is worth careful attention. The comprehensive text by Cressie (1993) is referred to very widely; careful reading of the often very short passages of relevance to a research problem can be highly rewarding. Schabenberger and

[6] Reprinted in 2004.

Gotway (2005) cover much of the same material, incorporating advances made over the intervening period. Banerjee et al. (2004) show how the Bayesian approach to statistics can be used in applied spatial data analysis.

Beyond the core statistical literature, many disciplines have their own traditions, often collated in widely used textbooks. Public health and disease mapping are well provided for by Waller and Gotway (2004), as is ecology by Fortin and Dale (2005). O'Sullivan and Unwin (2003) cover similar topics from the point of view of geography and GIS. Like Banerjee et al. (2004), the disciplinary texts differ from the core literature not only in the way theoretical material is presented, but also in the availability of the data sets used in the books for downloading and analysis. Haining (2003) is another book providing some data sets, and an interesting bridge to the use of Bayesian approaches in the geographies of health and crime. Despite its age, Bailey and Gatrell (1995) remains a good text, with support for its data sets in R packages.

In an *R News* summary, Ripley (2001) said that one of the reasons for the relatively limited availability of spatial statistics functions in R at that time was the success of the S-PLUS™ spatial statistics module (Kaluzny et al., 1998). Many of the methods for data handling and analysis are now available in R complement and extend those in the S-PLUS™ module. We also feel that the new packaging system in S-PLUS™ constitutes an invitation, for instance to release packages like **sp** for S-PLUS™– during the development of the package, it was tested regularly under both compute engines. Although the names of functions and arguments for spatial data analysis differ between S-PLUS™ and R, users of the S-PLUS™ spatial statistics module should have no difficulty in 'finding their way around' our presentation.

To summarise the approach to applied spatial data analysis adopted here, we can say that – as with the definition of geography as 'what geographers do' – applied spatial data analysis can best be understood by observing what practitioners do and how they do it. Since practitioners may choose to conduct analyses in different ways, it becomes vital to keep attention on 'how they do it', which R facilitates, with its unrivalled closeness to both data and the implementation of methods. It is equally important to create and maintain bridges between communities of practitioners, be they innovative statisticians or dedicated field scientists, or (rarely) both in the same person. The R Spatial community attempts to offer such opportunities, without necessarily prescribing or proscribing particular methods, and this approach will be reflected in this book.

1.7 R Spatial Resources

There are a range of resources for analysing spatial data with R, one being this book. In using the book, it is worth bearing in mind the close relationships between the increase in the availability of software for spatial data analysis on CRAN and the activities of the informal community of users interested in

spatial data analysis. Indeed, without contributions, advice, bug reports, and fruitful questions from users, very little would have been achieved. So before going on to present the structure of the book, we mention some of the more helpful online resources.

1.7.1 Online Resources

Since CRAN has grown to over 1,200 packages, finding resources is not simple. One opportunity is to use the collection of 'Task Views' available on CRAN itself. One of these covers spatial data analysis, and is kept more-or-less up to date. Other task views may also be relevant. These web pages are intended to be very concise, but because they are linked to the resources listed, including packages on CRAN, they can be considered as a kind of 'shop window'. By installing the **ctv** package and executing the command `install.views("Spatial")`, you will install almost all the contributed packages needed to reproduce the examples in this book (which may be downloaded from the book website).

The spatial task view is available on all CRAN mirrors, but may be accessed directly;[7] it provides a very concise summary of available contributed packages. It also specifically links two other resources, a mailing list dedicated to spatial data analysis with R and an R-Geo website. The R-sig-geo mailing list was started in 2003 after sessions on spatial statistics at the Distributed Statistical Computing conference organised in Vienna earlier the same year. By late 2007, the mailing list was being used by over 800 members, off-loading some of the spatial topic traffic from the main R-help mailing list. While R-help can see over 100 messages a day, R-sig-geo has moderate volume.

The archives of the mailing list are hosted in Zurich with the other R mailing list archives, and copies are held on Gmane and Nabble. This means that list traffic on an interesting thread can be accessed by general Internet search engines as well as the `RSiteSearch()` internal R search engine; a Google™ search on `R gstat kriging` picks up list traffic easily.

The second linked resource is the R-Geo website, generously hosted since its inception by Luc Anselin, and is currently hosted at the Spatial Analysis Laboratory (SAL) in the Department of Geography at the University of Illinois, Urbana-Champaign. Because the site uses a content management system, it may be updated at will, but does not duplicate the CRAN task view. When users report news or issues, including installation issues, with packages, this is the site where postings will be made.

1.7.2 Layout of the Book

This book is divided into two basic parts, the first presenting the shared R packages, functions, classes, and methods for handling spatial data. This part

[7] `http://CRAN.R-project.org/view=Spatial`.

is of interest to users who need to access and visualise spatial data, but who are not initially concerned with drawing conclusions from analysing spatial data per se. The second part showcases more specialised kinds of spatial data analysis, in which the relative position of observations in space may contribute to understanding the data generation process. This part is not an introduction to spatial statistics in itself, and should be read with relevant textbooks and papers referred to in the chapters.

Chapters 2 through 6 introduce spatial data handling in R. Readers needing to get to work quickly may choose to read Chap. 4 first, and return to other chapters later to see how things work. Those who prefer to see the naked structure first before using it will read the chapters in sequence, probably omitting technical subsections. The functions, classes, and methods are indexed, and so navigation from one section to another should be feasible.

Chapter 2 discusses in detail the classes for spatial data in R, as implemented in the **sp** package, and Chap. 3 discusses a number of ways of visualising for spatial data. Chapter 4 explains how coordinate reference systems work in the **sp** representation of spatial data in R, how they can be defined and how data can be transformed from one system to another, how spatial data can be imported into R or exported from R to GIS formats, and how R and the open source GRASS GIS are integrated. Chapter 5 covers methods for handling the classes defined in Chap. 2, especially for combining and integrating spatial data. Finally, Chap. 6 explains how the methods and classes introduced in Chap. 2 can be extended to suit one's own needs.

If we use the classification of Cressie (1993), we can introduce the applied spatial data analysis part of the book as follows: Chap. 7 covers the analysis of spatial point patterns, in which the relative position of points is compared with clustered, random, or regular generating processes. Chapter 8 presents the analysis of geostatistical data, with interpolation from values at observation points to prediction points. Chapters 9 and 10 deal with the statistical analysis of areal data, where the observed entities form a tessellation of the study area, and are often containers for data arising at other scales; Chap. 11 covers the special topic of disease mapping in R, and together they cover the analysis of lattice data, here termed areal data.

Data sets and code for reproducing the examples in this book are available from `http://www.asdar-book.org`; the website also includes coloured versions of the figures and other support material.