Part II

Analysing Spatial Data

Analysing Spatial Data

The analysis of spatial data is usually undertaken to make inferences, that is to try to draw conclusions about a hypothesised data generating process or to use an estimated process to predict values at locations for which observations are unavailable. In some cases, the conclusions are sufficient in themselves, and in others, they are carried through to other hierarchical layers in the model under scrutiny. Haining (2003, pp. 184–185) and Bivand (2002, p. 409) suggest (following Tukey, 1977) that our understanding of the data may be partitioned into

$$\text{data} = \text{smooth} + \text{rough}.$$

If the data are spatial, we can see that there is room for another term, irrespective of whether we are more interested in the fit of the model itself or in calibrating the model in order to predict for new data:

$$\text{data} = \text{smooth} + \text{spatial smooth} + \text{rough}.$$

The added term constitutes the 'added value' of spatial data analysis, bringing better understanding or predictive accuracy at the price of using specialised methods for fitting the spatial smooth term. We will here be concerned with methods for finding out whether fitting this term is worth the effort, and, if so, how we might choose to go about doing so.

Before rushing off to apply such specialised methods, it is worth thinking through the research problem thoroughly. We have already mentioned the importance of the Distributed Statistical Computing conference in Vienna in 2003 for our work. At that meeting, Bill Venables presented a fascinating study of a real research problem in the management of tiger prawn fisheries. The variable of interest was the proportion by weight of two species of tiger prawn in the logbook on a given night at a given location. In a very careful treatment of the context available, the 'location' was not simply taken as a point in space with geographical coordinates:

'Rather than use latitude and longitude directly as predictors, we find it more effective to represent station locations using the following two predictors:

- The shortest distance from the station to the coast (variable R_{land}), and
- The distance from an origin in the west to the nearest point to the station along an arbitrary curve running nearly parallel to the coast (variable R_{dist}).

[...] Rather than use R_{dist} itself as a predictor, we use a natural spline basis that allows the fitted linear predictor to depend on the variable in a flexible curvilinear way.

[...] Similarly, we choose a natural spline term with four internal knots at the quantiles of the corresponding variable for the logbook data for the "distance from dry land" variable, R_{land}.

The major reason to use this system, which is adapted to the coastline, is that interactions between R_{land} and R_{dist} are more likely to be negligible than for latitude and longitude, thus simplifying the model. The fact that they do not form a true co-ordinate system equivalent to latitude and longitude is no real disadvantage for the models we propose.' Venables and Dichmont (2004, pp. 412–413)

The paper deserves to be required reading in its entirety for all spatial data analysts, not least because of its sustained focus on the research problem at hand. It also demonstrates that because applied spatial data analysis builds on and extends applied data analysis, specifically spatial methods should be used when the problem cannot be solved with general methods. Consequently, familiarity with the modelling chapters of textbooks using R for analysis will be of great help in distinguishing between situations calling for spatial solutions, and those that do not, even though the data are spatial. Readers will benefit from having one or more of Fox (2002), Dalgaard (2002), Faraway (2004, 2006), or Venables and Ripley (2002) to refer to in seeking guidance on making often difficult research decisions.

In introducing this part of the book – covering specialised spatial methods but touching in places on non-spatial methods – we use the classification of Cressie (1993) of spatial statistics into three areas, *spatial point patterns*, covered here in Chap. 7, *geostatistical data* in Chap. 8, and *lattice data*, here termed areal data, in Chaps. 9–11. In Chap. 1, we mentioned a number of central books on spatial statistics and spatial data analysis; Table II.1 shows very roughly which of our chapters contain material that illustrates some of the methods presented in more recent spatial statistics books, including treatments of all three areas of spatial statistics discussed earlier (see p. 13).

The coverage here is uneven, because only a limited number of the topics covered in these books could be accommodated; the specialised literature within the three areas will be referenced directly in the relevant chapters. On the other hand, the implementations discussed below may be extended to cover

Table II.1. Thematic cross-tabulation of chapters in this book with chapters and sections of chosen books on spatial statistics and spatial data analysis

Chapter	Cressie (1993)	Schabenberger and Gotway (2005)	Waller and Gotway (2004)	Fortin and Dale (2005)	O'Sullivan and Unwin (2003)
7	8	3	5	2.1–2.2	4–5
8	2–3	4–5	8	3.5	8–9
9–11	6–7	1, 6	6, 7, 9	3.1–3.4, 5	7

alternative methods; for example, the use of WinBUGS with R is introduced in Chap. 11 in general forms capable of extension. The choice of contributed packages is also uneven; we have used the packages that we maintain, but this does not constitute a recommendation of these rather than other approaches (see Fig. 1.1). Note that coloured versions of figures may be found on the book website together with complete code examples, data sets, and other support material.