5

Introductory methods for spatially continuous data

Here and in the following chapter we consider the analysis of data which are 'spatially continuous'. Our objectives are significantly different from those in Part B. There, our observations were the locations of 'events', and we were interested in any pattern in such locations; here, we concentrate on understanding the spatial distribution of values of an attribute over the whole study region, given values at fixed sampling points. Our objective is to model the pattern of variability and to establish any factors to which this might relate. Ultimately, we may wish to use such models to obtain good predictions of values at points where the attribute has not been sampled. Such methods are relevant to many studies in the geosciences, such as soil science, climate study, hydrology, mining geology, and so on.

5.1 Introduction

In Part B we were concerned with the analysis of spatial 'point patterns', where our data consisted of the locations of a series of 'events' occurring in some study region $\mathcal{R}$. We also considered generalisations of this type of situation where additional information was available relating to the 'events', such as a 'labelling' of different types, a time of their occurrence, or background information on variations in the population from which they arose. However, in all these cases our interest remained in possible patterns in the *locations* of the events. Where additional information became involved it was of interest only in so much as it might help us to 'explain' or otherwise improve our analysis of the pattern in the locations.

That objective should be distinguished from the one which we consider in this and the following chapter, and indeed for the remainder of the book. It is now attribute values over $\mathcal{R}$ which we study, based on observed values at a predefined and fixed set of locations. Previously, we were interested in patterns in the locations of observations; now we are interested in patterns in attribute values. The locations are now simply sample sites at which attribute values have been recorded.

Here in Part C the focus of our discussion will be the analysis of an attribute which is conceptually *spatially continuous* over $\mathcal{R}$, and whose value has been sampled at particular fixed point locations s_i. (Again we shall maintain the notation introduced earlier, where $s_i = (s_{i1}, s_{i2})^T$, is a (2×1) vector, representing respectively the 'x' and 'y' coordinates of the ith location.) Typical examples of such spatially continuous data might be geological measures on an ore deposit such as mineral grade, or the concentration of some pollutant, or rainfall measurements, or soil salinity and permeability, and so on. This type of spatially continuous data is often referred to as *geostatistical* data. Clearly, much data arising in the fields of geography, geology and the environmental sciences is of this nature.

We distinguish this from the situation considered subsequently, in Part D of the book, where the attribute values are not considered spatially continuous, but instead relate only to a finite set of areas or zones that partition the study region. Perhaps a useful analogy here is with processes in time. Some processes evolve continuously—for example, temperature during the day can be sampled at any time. Other processes only evolve at discrete fixed intervals—it makes little sense to think of many economic measures as evolving continuously, because by their definition they can only be sampled say monthly, or quarterly. Currently we intend to consider the spatial equivalent of continuous processes in time; later, in Part D, we will be concerned with the spatial equivalent of fixed time period development, the areas being analogous to the time periods. We should acknowledge in passing that it may be perfectly possible to regard certain types of data under either framework—we might treat pollution measurements as spatially continuous for the purpose of predicting pollution levels at sites where we do not have measurements; however, we might also form the average of pollution measurements for a set of administrative health districts, so as to analyse such data in conjunction with health indices available only at that spatial level.

Statistically then, the situation we are considering in the remainder of this chapter and the next, is one where a series of observations y_i, $i = 1, \ldots, n$, on a spatially continuous attribute, have been recorded at corresponding spatial locations, s_i, in the study region $\mathcal{R}$. Occasionally we shall refer to our observations collectively as the $(n \times 1)$ vector $y = (y_1, \ldots, y_n)^T$. The measurements, y_i, are assumed to be observations on a spatial stochastic process $\{Y(s),\ s \in \mathcal{R}\}$, which varies in a spatially continuous way over $\mathcal{R}$ and has been sampled at fixed points. Note that strictly we should probably refer to our observed data values as $y(s_i)$, since they are observations on the random variables $Y(s_i)$. However, we shall mostly use the simpler notation y_i. Where it is convenient we may also refer to the random variable $Y(s_i)$ simply as Y_i, or collectively for all sample sites by the $(n \times 1)$ vector $Y = (Y(s_1), \ldots, Y(s_n))^T$.

The main objective of our analyses will be to infer the nature of spatial variation in the attribute over the whole of $\mathcal{R}$, from the sampled point values.

We may seek description in terms of a smooth surface which captures large scale global trends; or alternatively, we might also wish to study aspects of local variability, particularly if the primary objective is one of accurate interpolation or prediction of the value of the attribute at points other than s_i. Essentially these objectives correspond to the notions discussed in Chapter 1 of modelling first order variation in the mean value of the process, $E(Y(s))$, which we shall refer to as $\mu(s)$; or of modelling second order variation or spatial dependence between $Y(s_i)$ and $Y(s_j)$ for any two locations, s_i and s_j, in $\mathcal{R}$; that is, $COV(Y(s_i), Y(s_j))$.

In general, as with 'point patterns' previously, our approach will involve a mixture of methods, some designed to examine large scale heterogeneity in the mean value of the attribute, $\mu(s)$, over $\mathcal{R}$, others where we shall examine areas within which we assume the process to be stationary or isotropic (see Chapter 1) and look for spatial dependence, in this case through the use of the *variogram* or the *covariogram*, which, as we will see, attempt to capture the covariance structure of the process. When it comes to proposing possible statistical models in order to try to 'explain' any effects we may detect, we shall usually think of these as consisting of two 'components'; a first order component, representing large scale variations in $\mu(s)$; and a stationary second order component representing small scale spatial dependence in the process. The reader should bear in mind that, as in the case of 'point pattern' analysis, there will inevitably be cases where spatial distribution explained in terms of dependence with the assumption of homogeneity, could possibly arise also from heterogeneity in mean value. Ultimately, our division into the two components, although guided by our analyses, will be to some extent arbitrary.

In line with the structure established in Part B, we begin by outlining some case studies that we shall use throughout this part of the book. We then move on to consider methods, under the general headings of visualisation, exploration and modelling.

5.2 Case studies

Throughout this and the following chapter, we shall use a selection of data sets involving spatially continuous variables, to illustrate various forms of analyses. We have provided copies of these data sets on disk. They include:

- Rainfall measurements in California
- Rainfall measurements in central Sudan
- Temperatures for weather stations in England and Wales
- Groundwater levels in Venice
- Radon gas levels in Lancashire
- PCBs in an area of south Wales
- Geochemical data for north Vancouver Island, Canada
- South American climate measures.

We begin by describing all of these data sets in more detail; full references to the sources of the data are given at the end of this chapter.

We have already encountered the first of these data sets in Chapter 1, that concerned with the spatial variability of rainfall in California. We described some analyses there and refer the reader back to that discussion. Recall that the data consist of recordings of average annual precipitation at a set of 30 monitoring stations, distributed across the state of California and shown in Figure 1.1. For these same sites we have measures of altitude, latitude, and distance from the coast, each of which is a possible covariate that might explain the variation in precipitation. With these data our interest is perhaps less in making spatial predictions of rainfall than in trying to explain spatial variation using the available covariates.

The next example seems, on the surface, very similar, in that it too deals with spatial variability of precipitation, but this time in central Sudan. Here, however, we have no covariates. The problem is to try to describe the spatial variation and perhaps to make estimates of precipitation in areas where there are no monitoring stations. Resources for collecting such basic information are obviously limited, particularly in the developing world. The data we include consist of measurements of total annual precipitation in 1942, 1962 and 1982, recorded at 31 sites. The sites are unevenly distributed across the region, with a preponderance of monitoring stations around the capital city of Khartoum, at the confluence of the White Nile and the Blue Nile (Figure 5.1).

Understanding the distribution of rainfall in central Sudan is important, particularly because the area faces some of the most severe population pressures of any part of the Sahel region. The intensity of crop cultivation is considerable, and this, coupled with severely limited rainfall resources, has led to the desertification so characteristic of much of the wider region.

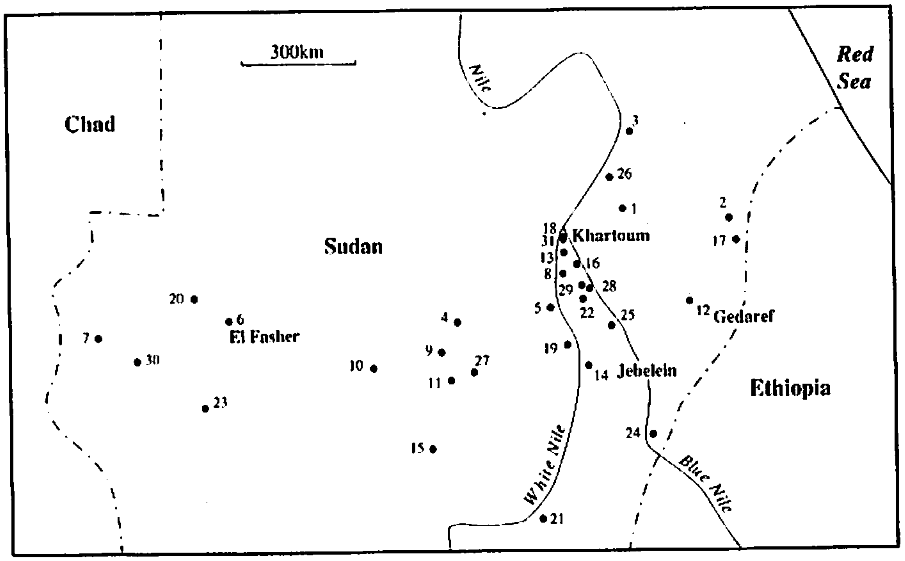

Fig. 5.1 Rainfall measuring sites in Sudan

Although we include here rainfall totals for only three years it is well known that precipitation is highly variable from year to year. Further, there is, of course, very great seasonal variability in rainfall, with the wettest months generally in July and August, though this seasonality itself varies geographically. We restrict attention to seeing how annual rainfall varies across space and how the nature of this spatial variability has changed over two twenty-year periods. Some climatologists have detected no real evidence for long-term reduction in rainfall, the droughts in recent years being thought of as natural fluctuations. Further, we shall be interested to see to what extent rainfall is highly localised in space. We should appreciate that when considering issues of desertification what matters more is not so much annual precipitation as seasonal and diurnal variation. However, the spatial localisation is certainly also of interest.

A further set of climate data which we include relate to England and Wales. Data on the mean daily temperature in August 1981 and August 1991 were extracted from the Monthly Weather Report, for a set of 48 stations distributed across the country. The main criterion for choosing sites was to ensure a reasonable geographical spread, but we also chose only those stations for which values were available in both years. In addition, we have also included the elevation of each site. We shall be interested in seeing whether we can obtain a good description of the geographical variation in temperature. In the simplest case we shall want to see the extent to which temperature variation can be explained solely in terms of geographical location. But we will then wish to see if further explanatory power is achieved by adding in elevation as a further covariate. Given that we have data for two years we might also want to investigate whether the character of spatial variation in one year is different from that in another.

Some parts of the world rely heavily on groundwater for supplies of both drinking water and water for industrial and commercial use and our next example relates to data on levels of such groundwater. In Venice withdrawals from several aquifers (the rock formations that contain water) at different depths have often been heavy, and have led to major problems of land subsidence. The result has been that the local population has been exposed to the risk of flooding from the Adriatic Sea. In order to be able to control the pumping from wells, hydrologists require accurate maps of the subsurface levels of groundwater (the so-called 'piezometric' surface).

The data we include for the Venice region come from a series of sparsely distributed boreholes (Figure 5.2). The groundwater levels have been measured in 1973 and 1977. In !973 data were measured for 40 sites; however, data for only 35 of these sites are available for 1977. Since 1973 withdrawals of groundwater from the major industrial area (Porto Marghera) have declined markedly, leading to a lower risk of subsidence and of saltwater intrusion into the freshwater aquifer. As with other data sets considered in this chapter we

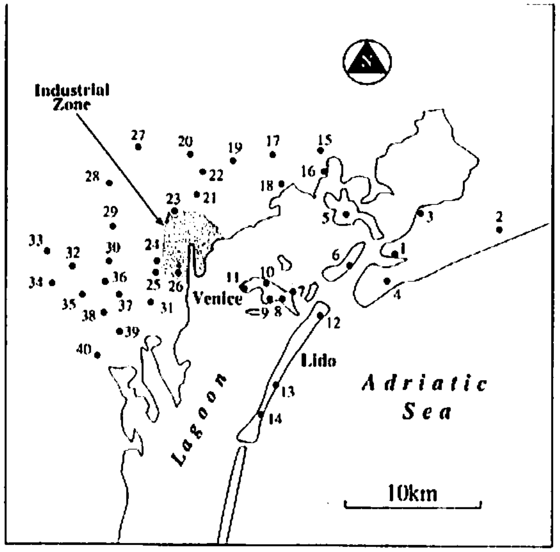

Fig. 5.2 Boreholes at which groundwater levels were measured in Venice 1973

would like to be able to describe the nature of spatial variation as accurately as possible, in order to provide estimates of groundwater levels at locations that have not been sampled. We also wish to know how reliable these estimates are.

Radon-222, commonly called just radon, is a naturally occurring radioactive gas produced by the decay of trace quantities of uranium. Released to the atmosphere it is harmless but when trapped within buildings it can accumulate and is considered to be a serious risk factor for lung and possibly other cancers. As a result, local authorities in many parts of the world are monitoring the gas in homes where the risk is thought to be raised. For example, some areas of South West England comprise uranium-bearing granitic rocks, and radon levels recorded in many properties there are extremely high. However, it is known that levels of radon are very highly variable in space, with levels in one property bearing little relationship to values in adjacent properties; this is thought to be a function of building materials, degree of ventilation and insulation, and so on. It is clearly important to try to characterise this degree of spatial variability. The data we include here relate to Lancashire, an area not thought to be at particular risk. Data are available for 344 homes and were measured in 1989 (Figure 5.3). For reasons of confidentiality the locations are defined with a resolution of 100 metres and have been randomly shifted. As a result, while any results obtained are a faithful description of spatial variation it is not possible to identify individual properties!

In general, the picture is of relatively low radon values to the west and south, higher values to the north and east. This reflects the fact that the area to the

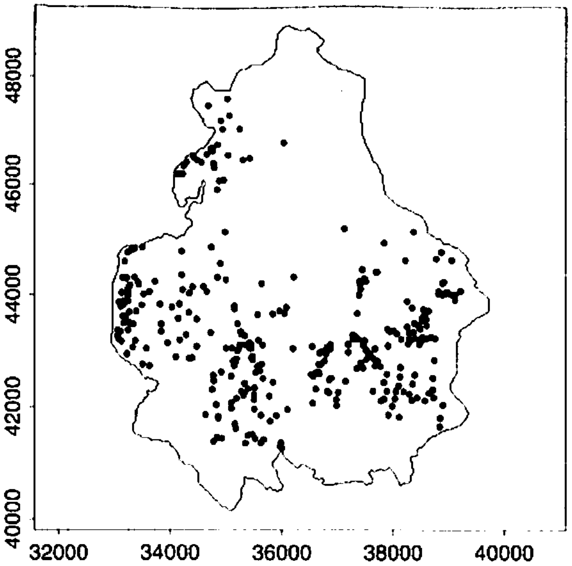

Fig. 5.3 Radon measurement sites in Lancashire

west (known as the 'Fylde') is covered by glacial deposits which trap the radon gas and prevent it rising to the surface. Further north and east there are no such surficial deposits and the gas escapes from the sandstones and limestones to the surface. The distribution of sample sites reflects population distribution. Population is very sparsely distributed to the north-east of the study region (the 'Trough of Bowland') and most of the sites are clustered in the major centres of population: Blackpool, Lancaster, Preston, Blackburn, and Burnley. With these data interest lies in understanding such broad regional trends in more detail, and examining if there is smaller scale variation superimposed on such trends. It might be possible to interpolate the data to provide a regional map of radon levels, or identify areas where sampling is 'thin' and needs to be improved.

The next case study relates to data on environmental pollution in soil samples. The data collection arose as a result of concerns over possible risks associated with the contamination of the environment with polychlorinated biphenyls (PCBs) in a small area of South Wales. Near the town of Pontypool is located a large plant for the incineration of chemical wastes (including PCBs) at very high temperatures. There had been worries that some of these substances had been escaping into the surrounding environment, possibly contaminating soil and vegetation. Data on 70 sites within an area of about 6 square kilometres are included here (Figure 5.15a). The soil samples were taken in late 1991. Measuring minute concentrations of PCBs is notoriously difficult and there were painstaking efforts to do this as accurately as possible, with different laboratories cross-checking the measurements. The research report from which the data are taken lists different sets of data for the sites; we have used the standardised data for the sum of seven different types of PCB (known as 'congeners'). We shall, however, simply refer to the data as PCB measurements. The analytical problem is to try to characterise the pattern of spatial variability; are there locally elevated concentrations around the incineration plant?

A further set of data that we shall find of interest in a spatially continuous setting are taken from the National Geochemical Reconnaissance performed by the Geological Survey of Canada. The particular study area is that part of Vancouver Island north of latitude 50 and west of longitude 126. In the small data set included here we have 916 sites (stream locations) at which five elements (part of a much larger set), have been measured; these are: zinc (Zn), copper (Cu), nickel (Ni), cobalt (Co), and manganese (Mn). The coverage is quite dense over the study area; in general, the reconnaissance programme aims to achieve a sampling density of approximately 1 sample per 13 km^2. As with the radon and PCB data, we shall be interested in characterising the nature and scale of spatial variation in geochemistry.

Broad-scale, multi-element reconnaissance programmes are carried out in many countries, since an understanding of spatial variation is clearly of some economic importance. There is quite a long tradition in geochemistry of mapping and analysing this kind of data, particularly in order to try to identify unusual features, or anomalies, that may suggest potential mineralisation. We comment later on the visualisation of spatially continuous data, but it is useful to point out here that geochemists are not only concerned with mapping single elements. They frequently wish to derive multi-component maps, which show the simultaneous variation of a number of elements. Three elements are usually selected; for instance, we might use the combination: zinc, copper, lead. The fact that interest may be as much in combinations of elements as opposed to separate ones, implies that an understanding of geochemical data is often a multivariate problem. We may want to obtain estimates of the distribution of a single element; but we might wish to look at an ensemble, using the sorts of methods we consider in Chapter 6.

Geochemical data are not of use solely in mineral exploration and resource evaluation. Depending on the particular trace elements being sampled, such data are also of value in an environmental context. High precision geochemical maps can be used to identify areas with raised levels of toxic elements such as lead, cadmium and arsenic. They might also be used to identify areas that are deficient in some elements. Some diseases in animal populations have been correlated with mineral deficiencies. For example, sheep can suffer from a disease called 'swayback', which has been linked to copper deficiency. We might then use a regional geochemical map to identify areas of crop cultivation where the soil could be treated to alleviate the problem. In some cases we might even wish to look for associations with human health problems.

One feature of geochemical data is that they often tend to be log-normally distributed; at least, the distribution will often be highly skewed, with a preponderance of generally low values, together with a much smaller set of extreme values. This is a common feature of many data sets which involve measurement of concentrations of some description (such as that on radon and PCBs). It will therefore often be advisable to transform the data before analysis, a logarithmic transformation being the obvious choice.

During this century several different schemes for climate classification have been devised, many because of the view that such schemes are relevant to human activity. A major constraint, however, in devising a reasonably objective classification concerns the type of data recorded at climate stations. Schemes that rely solely on temperature and precipitation, both of which can be measured with simple equipment, are going to be most useful in characterising climates across the global land surface. Because we are wanting to use a variety of both temperature and precipitation measures we need multivariate methods to help devise our classification. So with our next data set we are interested less in the interpolation of a single variable, as with the Sudan rainfall data, but more in using the available sample information to devise a multivariate classification.

The data relate to 76 climate stations in South America (Figure 6.1). Stations were selected that were all under 200 metres above sea level, in order to remove the effects of elevation. There are 16 climate variables, all of which are concerned with particular aspects of temperature and precipitation. The full list of variables is given in Table 5.1 and includes information on seasonality of climate and not simply annual totals or means. Clearly, we can expect broad regional differences, separating the tropical low latitude regimes in much of Amazonia from the middle latitude steppe conditions in parts of Argentina, for example. But to what extent can we paint a richer picture of climatic variability? We shall be using these data not so much in the present chapter but in Chapter 6, where we explore the role of multivariate methods.

Table 5.1 Description of climate variables for South America

1.	Average annual temperature
2.	Average daily January temperature
3.	Maximum January temperature
4.	Minimum January temperature
5.	Average daily July temperature
6.	Maximum July temperature
7.	Minimum July temperature
8.	Average annual precipitation
9.	Average January precipitation
10.	Average July precipitation
11.	Average annual number of days precipitation > 1 mm
12.	Average number of days in January precipitation > 1 mm
13.	Average number of days in July precipitation > 1 mm
14.	Temperature range (January–July)
15.	Precipitation ratio (July/January)
16.	Rain days ratio (July/January)

Having outlined the sort of data that we might be interested in analysing, let us now consider what kinds of visual representations we can obtain of such data before we go on to consider more sophisticated ways of exploring spatial variation. The simplest type of map that we can produce from data of this type is one in which the data value is written alongside the sampled location. However, this will not look very elegant or informative! A better solution is to use a symbol at each site, the nature of which carries useful information about the data value. We might use a variety of symbol types to represent different classes of variation; however, a preferred solution is to use *proportional symbol* maps.

Different geometric symbols can be used in this way, but commonly, and most simply, circles are used, where the area of the circle is proportional to the data value at that location. Alternatively, we might use rectangles, the height of which represents value. Although more difficult to produce, spheres are sometimes used to represent a variable such that the volume of the sphere is proportional to the data value (see Figure 2.1 for an example). Regardless of what symbol we use there are some important cartographic issues that arise in such mapping. The main problems concern how the map user perceives data values, since it has long been known that people tend to underestimate areas on maps. For example, it seems sensible to draw circles whose radii are proportional to the square root of the data values. In practice, however, cartographers introduce a correction factor to compensate for this perceptual underestimation, so that instead of taking square roots (or an exponent of 0.5) they use an (empirically derived) exponent of 0.57. Other problems concern the overlap of symbols, since with many locations the map will appear cluttered. Clever cartographic software allows for 'halos' around symbols to alleviate this problem.

As an aside, proportional symbol maps are widely used in contexts other than that of spatially continuous data. For example, we might map retail sales for stores in this way, even though this variable is in no way spatially continuous. Or, we could take data such as total population, representing an aggregate value within an areal unit, and map it using proportional symbols; we return to such uses later, when dealing with area data in Chapter 7.

If we use proportional circles to represent one data item we may then fill the circles with colours (or shading) to indicate the classified value of another variable. However, this too makes demands on the map reader. Another strategy is to shade the symbols with classified values of the same variable used to derive the symbol. In cases where many data values are present and proportional symbols overlap to a large extent, it might even be useful to do away with proportional symbols altogether and use small symbols of fixed size, shaded or coloured according to such a classification of data values. We need to say something about such classification since both the number of classes we use and the type of class intervals we select will determine the 'message' that our map communicates. Precisely the same issues concerned with class interval selection arise in analysing data collected for areal units, and we will refer back to this discussion in Chapter 7, where we consider that type of data.

There are no hard and fast rules about numbers of classes; clearly, this is largely a function of how many data values we have. For example, if we have only a small sample of 20 or 30 sites (as for the Sudan rainfall data) it hardly makes sense to use seven or eight classes; however, with perhaps two or three hundred measurements (as in the example of the radon data) a set of seven or eight classes is likely to prove informative. As a general rule of thumb some statisticians recommend a number of classes of $(1 + 3.3 \log n)$, where n is the number of observations.

As for class interval selection, there is a wide variety of schemes that may be considered. We do not review these in detail here, but simply comment on some possibilities. 'Equal intervals' are self-explanatory; they are valuable where data are reasonably uniformly distributed over their range, but if the data are markedly skewed they will give large numbers of values in just a few classes. This is not necessarily a problem, since unusually high (or low) values are easily picked out on the map. An extension of this scheme is to use 'trimmed equal' intervals where the top and bottom of the frequency distribution (for example, the top and bottom 10 per cent) are each treated as separate classes and the remainder of the values are divided into equal intervals. A way of ensuring that an equal number of observations fall into each class is to base class intervals on the percentiles of the distribution, that is, to use 'quantiles': for example, selecting five classes will produce a quintile map. Another idea might be to use 'standard deviates' where the classes are based on intervals distributed around the mean in units of standard deviation. Finally, of course, we could let cartographers define their own class intervals. For instance, if we are mapping income data we may want to fix certain intervals (such as a 'poverty line'). In general, there is no substitute for a careful examination of the distribution of data values before selecting class intervals. As should be clear, we can obtain a huge variety of maps simply by varying the class intervals. This turns out to be of particular significance if mapping data for areal units and we will return to this point again in Chapter 7.

We can try out some of these ideas and get some initial sense of spatial variation in the data that comprise some of our case studies by visualising them using proportional symbols. Regardless of which year is taken the Sudan rainfall data give the same general impression when mapped, of higher levels to the south-east of the study area (Figure 5.4). This reflects the fact that the southern part of the study area is rather more humid. Further west and north we are moving into arid and semi-arid areas.

Groundwater levels in the Venice region may be mapped for both 1973 and 1977. In 1973 (Figure 5.5), note the cluster of negative values at sites 7–10 (in the city of Venice: see Figure 5.2 for the specific locations) and in the industrial region to the north west of Venice (sites 23–26). As noted earlier, considerable volumes of groundwater were being extracted in these areas. By 1977, while there are still some negative values these are lower in magnitude. If we examine

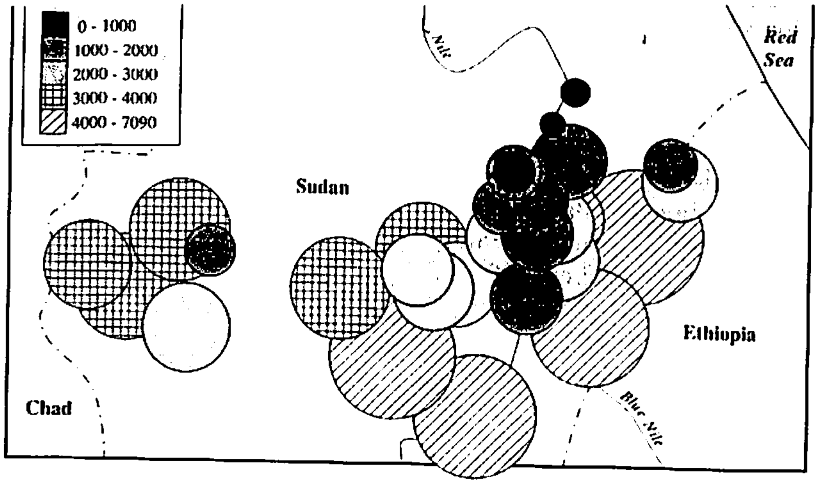

Fig. 5.4 Proportional symbol map of Sudan rainfall in 1982

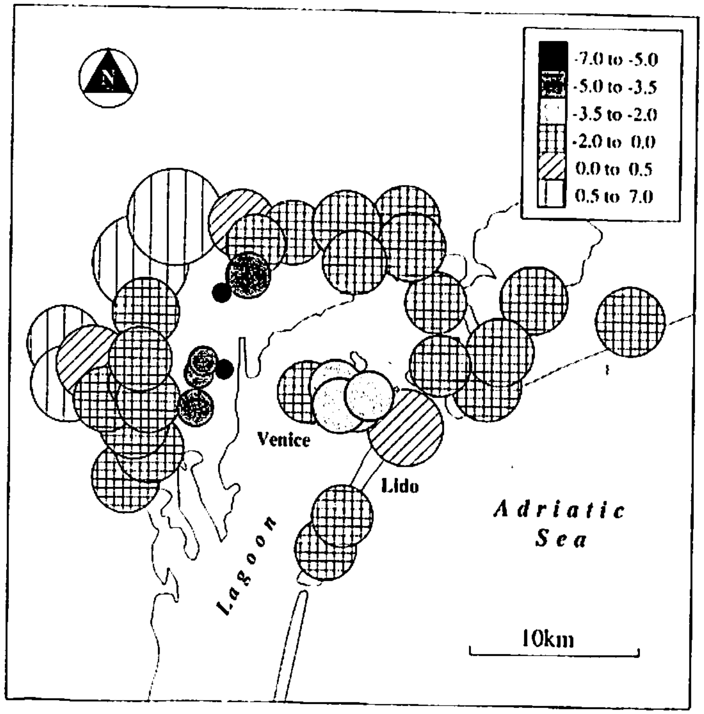

Fig. 5.5 Proportional symbol map of Venice groundwater, 1973

the relationship between 1977 and 1973 values there is a close correlation, but some sites with negative values in 1973 had become positive in 1977.

As mentioned earlier, some environmental data have highly skewed distributions and transformation is necessary before such data can sensibly be visualised in the sort of ways we have discussed. This would certainly be true, for example, if we wished to produce proportional symbol maps for either the PCB or radon data, where a logarithmic transformation is advisable.

These sorts of mapping tools are useful preliminary visualisations of data, but, typically, when we wish to look at variation in spatially continuous data we really need to use maps that *show* such continuity. For instance, the temperature maps that we see on television or in newspapers show not the values at weather stations but an interpolated or 'contoured' surface of temperature variation. How we might obtain such maps is the subject of the next section.

5.4 Exploring spatially continuous data

In this section we consider methods involving various summary statistics or plots which may be derived from the observed data and used informally to investigate hypotheses of interest or suggest possible models. Some of these are more concerned with investigating first order effects in the process, while others address the possibility of spatial dependence or second order effects.

We start by discussing various simple approaches to estimating how $\mu(s)$, the mean value of the attribute of interest, varies across the study region. The 'surfaces' that result from such techniques may be viewed as contour maps for purposes of interpretation. In a sense, the methods discussed here are equivalent to those we considered in relation to the estimation of the intensity of a point process, in Chapter 3. We prefer to think of these methods as exploratory techniques useful mainly in examining global trends in variation. It is tempting to use some of the methods for prediction and indeed this is often done. However, it should be appreciated that the justification for doing so is rather weak. In particular, they do not involve an explicit statistical model for the data under consideration and make no attempt to incorporate explicitly the possibility of spatial dependence. As a result, none of the methods provides any estimates of the errors that can be expected in the results. Such provisos make such methods very dubious for the purposes of prediction.

We then move on to discuss techniques for exploring spatial dependence in the data, through the *covariogram* and *variogram*. These are equivalent to the techniques used in Chapter 3 for examining second order properties of a spatial point process, such as the distribution of inter-event distances and the *K* function.

5.4.1 Spatial moving averages

A very simple way to estimate $\mu(s)$ is by the average of the values at neighbouring sampled data points. For example, $\mu(s)$ may be estimated as the unweighted average of the sample values at the three sampling points nearest to s—a *three-point spatial moving average*. If this averaging is also applied at the sample points s_i, then the resulting map will be smoother than the original observations and will indicate global trends in the data. The more points included in the moving average, the greater the smoothing will be. Patterns that initially show a lot of 'noise' can be 'cleaned up' in this way. The effect, of course, is to remove local place-to-place variation in the data.

The obvious problem with using this approach is that it does not allow for spatial variations in the distribution of sample sites. For instance, there is no discrimination between a site that is a considerable distance from its 'neighbours' (as defined by the averaging scheme used) and one which is very close to them. To counter this problem, we might want to use a weighted average of neighbouring points:

$$\hat{\mu}(s) = \sum_{i-1}^{n} w_i(s)y_i$$

where $\sum w_i(s) = 1$ and $w_i(s) \propto h_i^{-\alpha}$ or $w_i(s) \propto e^{-\alpha h_i}$ where h_i is the distance from s to s_i and α is a parameter with a value chosen to provide a suitable degree of smoothing. Usually $w_i(s)$ is chosen to be zero beyond some appropriate maximum distance, since, again, we might only want to examine a few neighbours. The advantage over the unweighted scheme is that we 'downgrade' the influence of neighbours that are some distance from the site under consideration. As we shall see, there are more sophisticated ways of performing this sort of weighting.

5.4.2 Methods based on tessellation

There are a number of estimation techniques for $\mu(s)$, which have been developed on the basis of *tessellation* (or tiling) of the observed sample locations s_i. The most commonly used employ the *Delaunay triangulation*, also known as a *triangulated irregular network* (TIN).

Given n distinct locations in a planar region $\mathcal{R}$ we can assign to each location, s_i, a 'territory' consisting of that part of $\mathcal{R}$ which is closer to s_i than to any other of the locations. This construction is referred to as the *Dirichlet tessellation* of the locations in $\mathcal{R}$ (Figure 5.6). The 'tiles' so constructed are sometimes referred to as *Voronoi* or *Thiessen polygons*.

Except possibly along the boundary of $\mathcal{R}$ each Thiessen polygon is a convex region. Locations s_i and s_j whose Thiessen polygons share a common boundary can be thought of as being 'contiguous'. The lines joining all such pairs of 'contiguous' locations define a triangulation of the locations called the *Delaunay triangulation*. The Delaunay triangulation and the Dirichlet tessellation can be thought of as two sides of the same coin (Figure 5.6). One way to form the Dirichlet tessellation is from polygon sides which are the perpendicular bisectors of the edges of the Delaunay triangulation and polygon vertices which are the corresponding circumcentres.

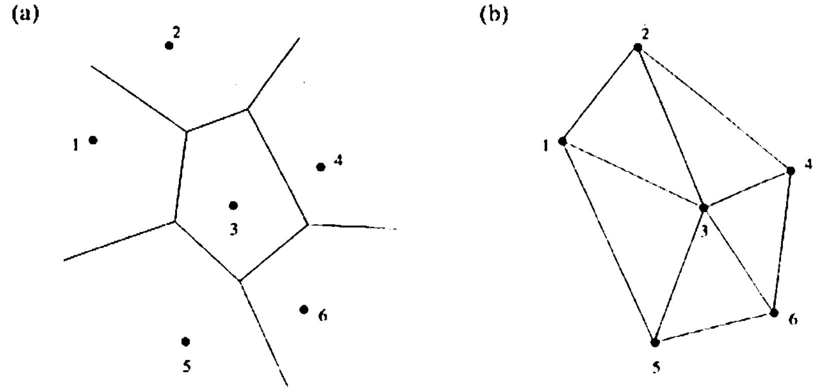

Fig. 5.6 (a) Dirichlet tessellation and (b) Delaunay triangulation

The argument for using such tessellations of the sample sites s_i, when estimating $\mu(s)$ at a unsampled site, s, arises by analogy with curve fitting in one dimension. It is common in fitting a curve to a set of data points in one dimension to use a series of simple functions, such as polynomials, each of which is fitted to successive groups of data points and constrained to give some degree of continuity at their joins. This is the basis of what is known as *spline smoothing*, a common example being the use of *cubic splines*, successive cubic functions. The problem with applying this idea to the spatial or two-dimensional case is that the sample points s_i do not have a natural ordering—they do not divide $\mathcal{R}$ into obvious sub-regions over which to fit two-dimensional splines. One obvious way to produce a natural partitioning of $\mathcal{R}$ is to base it on a 'good' tessellation of the observed points. The Delaunay triangulation provides such a tessellation in that the triangles produced are as close to equilateral as possible. Other triangulations can be defined with 'good' properties of various other kinds, but we do not go into details here.

Once the set of non-overlapping triangles is defined, each vertex of which represents a site or sampled location, we may imagine a vertical line being constructed at each site, the height of which is proportional to the value at that site. The triangles on the base map may then be projected to give a set of tilted planes (Figure 5.7a). A simple way to proceed is then to assume that the surface we wish to estimate can be approximated by these tilted planes. Given that the heights are known at the three vertices of each of the planes, we can use simple geometry to estimate the value at any point on the map.

Once we have, in principle, estimated a value at any location on the map, we can then go on to draw *isolines* (also known as *isarithms*), lines joining points of equal value. We need to select the number of isolines we want and to define the

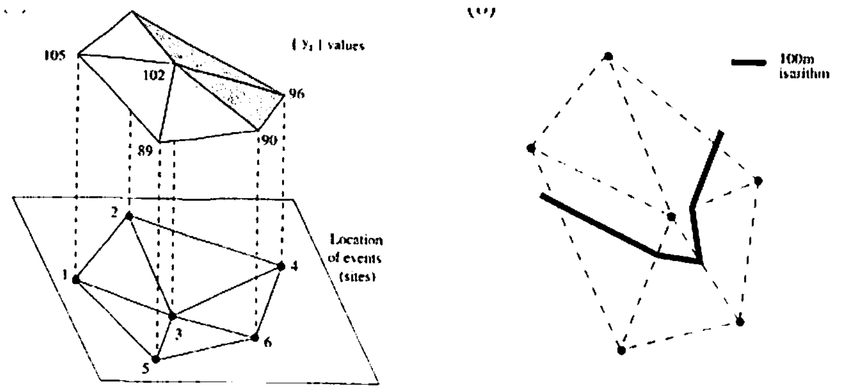

Fig. 5.7 Interpolation using a TIN

contour intervals (where we use the word 'contour' by analogy with the mapping of height above sea level). What the number of contours and contour intervals should be, raises the same sort of issues as in class interval selection for proportional symbol maps discussed earlier. Once these decisions are taken we can 'thread' an isarithm through the data points, initially as a set of straight line segments (Figure 5.7b).

We emphasise that our aim here is not cartographic excellence—we are not interested in 'elegant' contouring! We are using the contouring as an exploratory device to examine variations in $\mu(s)$. In particular, note that the contours obtained in Figure 5.7 are angular rather than smooth; this reflects the nature of the TIN 'model' that is used in the interpolation. Too much generalisation and 'cleaning up' of the contours to give smoothly varying lines can prove somewhat misleading, since the contours may be extrapolated well beyond the spatial range of sample locations. Note in particular that the Delaunay triangulation is only defined over the 'convex hull' formed by the data points and will not necessarily cover the edges of $\mathcal{R}$.

We can illustrate these ideas by 'contouring', in this way, one of the data sets we have already examined, that of groundwater in Venice. A TIN contour map of the 1973 levels is given in Figure 5.8 and clearly highlights the low values in the industrial area on the mainland.

There are many variations on the basic idea of TIN interpolation that we have described. The simple approach of using a series of planes each of which corresponds to a Delaunay triangle being fitted to the observed y_i at the corners, can be extended by fitting quadratic surfaces to the tiles and constraining them to have certain amounts of continuity at the joining edges. Another related approach with several variations falls under the heading of *natural neighbourhood interpolation*. This is based on repeated Dirichlet tessellations. Essentially, a weighted average of neighbouring sample points is used to interpolate the value at s:

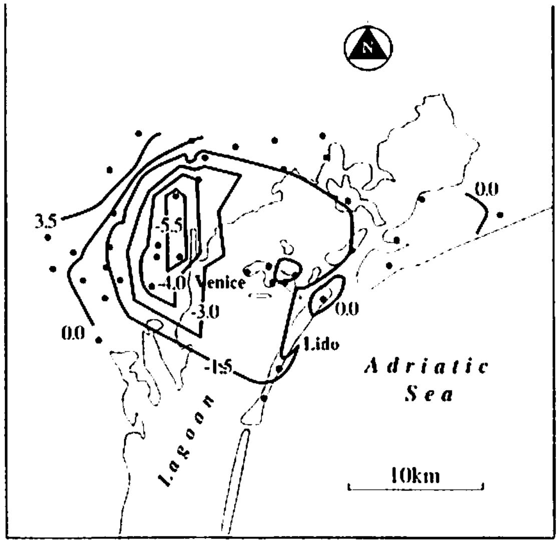

Fig. 5.8 TIN contours of Venice groundwater in 1973

$$\hat{\mu}(s) = \sum_{i=1}^{n} w_i(s) y_i$$

However, unlike the case in our earlier discussion of spatial moving averages, the weights $w_i(s)$ are now the proportion of the area of the Dirichlet tile around s_i which is 'stolen' by the tile around s, when a Dirichlet tessellation is performed on $(s, s_1, \ldots, s_n)$ as opposed to just $(s_1, \ldots, s_n)$. A new tessellation needs to be performed for each interpolated point; however since the new tessellation involves just one extra point, efficient methods may be developed to modify the base tessellation on $(s_1, \ldots, s_n)$ in each case, rather than compute an entirely fresh tessellation.

5.4.3 Kernel estimation

In the case of the analysis of 'point patterns' in Part B a flexible exploratory approach to the estimation of intensity, $\lambda(s)$, or 'events per unit area', over $\mathcal{R}$ was provided by kernel estimation. Ignoring edge corrections our basic kernel estimate was:

$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right)$$

where $k()$ was a bivariate probability density function (*the kernel*) which was symmetric about the origin, and $\tau > 0$ (*the bandwidth*) determined the amount of smoothing—essentially the radius of a disc centred on s within which points s_i will contribute 'significantly' to the estimate $\hat{\lambda}_\tau(s)$. We refer the reader back to the discussion in Chapter 3 for more details.

We are now interested not in 'events per unit area', $\lambda(s)$, but in the mean value $\mu(s)$ of an attribute whose values y_i, have been sampled at locations s_i. The question arises as to whether we can adjust the kernel technique to estimate $\mu(s)$. A formal and rigorous derivation of an appropriate kernel estimate is possible, but an informal argument will suffice to justify it here.

An intuitive way to introduce the attribute values into our previous kernel estimate would be to consider:

$$\sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right) y_i$$

If the original estimate represents 'the number of observations per unit area', then this extension in some sense represents 'the total amount of the attribute per unit area'. It follows that to turn it into an estimate of the mean value of the attribute we need to divide it by the 'number of observations per unit area'. This suggests that an appropriate kernel estimate for $\mu(s)$ would be:

$$\hat{\mu}_\tau(s) = \frac{\sum_{i=1}^{n} k\left(\frac{(s - s_i)}{\tau}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{(s - s_i)}{\tau}\right)}$$

where τ^{-2} in the denominator and the numerator have been cancelled. At points s where the denominator is 0 the numerator must also be 0 and by convention $\hat{\mu}_\tau(s)$ is set to 0 at these points.

Corrections for edge effects have been ignored so far, but since the form of $\hat{\mu}_\tau(s)$ is a ratio of two kernel estimates, involving the same set of s_i, we do not need to incorporate any edge correction since it would appear identically in both the numerator and the denominator and cancel.

As discussed in Chapter 3, the choice of the specific form of the kernel is of secondary importance relative to the choice of an appropriate bandwidth in terms of significantly affecting the resulting estimate. A typical choice for $k()$ might be the quartic kernel which we have already discussed in some detail in Chapter 3 (see Figure 3.3), that is:

$$k(u) = \begin{cases} \frac{3}{\pi}(1 - u^T u)^2 & \textit{for} \;\; u^T u \leq 1 \\ 0 & \textit{otherwise} \end{cases}$$

The effect of increasing the bandwidth τ is to expand the region around s within which observed values y_i influence the estimate at s. For very large τ, $\hat{\mu}_\tau(s)$ will appear flat and local features will be obscured; if τ is small then $\hat{\mu}_\tau(s)$ tends to a collection of spikes centred on the s_i.

Note that the kernel estimate $\hat{\mu}_\tau(s)$ is really just a more sophisticated version of the weighted moving average scheme discussed earlier. That is, for a given τ, it is essentially just a weighted average of the sample data points $\sum w_i(s) y_i$, where the weights

$$w_i(s) = \frac{k\left(\frac{(s - s_i)}{\tau}\right)}{\sum_{j=1}^{n} k\left(\frac{(s - s_j)}{\tau}\right)}$$

depend upon τ, upon s through the distance between s and s_i *and* upon the local intensity of sample points in the neighbourhood of s through the denominator. Sample observations at a given distance from s obtain more weight in regions of $\mathcal{R}$ where sample points are sparse than where they are dense.

For any chosen kernel and bandwidth, values of $\hat{\mu}_\tau(s)$ can be examined at locations on a suitably chosen fine grid over $\mathcal{R}$ to provide a useful visual indication of the variation in the mean value of the attribute over $\mathcal{R}$. The bandwidth τ can be used to vary the level of 'smoothness' of this estimate, as described previously. As also discussed there, methods exist which attempt to determine the optimal bandwidth for any particular data set, and for 'adapting' the value of τ over $\mathcal{R}$ to reflect the local density of sample points. We refer the reader back to our section on kernel estimation in Chapter 3 for a fuller discussion.

5.4.4 Covariogram and variogram

The exploratory methods discussed so far in this chapter have been concerned with first order variations in attribute values, in other words with estimating the way in which the mean or expected value of the process varies in the study region. We now turn our attention to methods designed more explicitly to explore the spatial dependence of deviations in attribute values from their mean, the second order properties.

In Chapter 3 we used the K function or *reduced second moment measure* as a tool to summarise and analyse second order properties of a point pattern. In the current situation the equivalent tool is the *covariance function* or *covariogram*. Before looking at this formally, what do such functions represent in an intuitive sense?

Recall that, in elementary statistics, we speak of covariance as measuring the extent to which two variables vary together. For instance, if, as one increases, so too does another, we say there is positive covariance and this is estimated as the sum of cross-products of deviations of observations from the respective means of the two variables. If we divide this covariance by the product of the two standard deviations we obtain an estimated correlation coefficient. In a spatial context, as we have discussed in Chapter 1, the same ideas apply, except that we are interested not in the covariance or correlation between two variables but rather the way in which the deviations of observations from their mean value at different locations on the map co-vary or are correlated. Because we know that many variables show 'spatial persistence' (for example, the deviation from mean rainfall measured at one location is likely to be similar to that recorded a kilometre away, but less likely to co-vary with that twenty kilometres away) we can anticipate typically observing positive covariance or correlation at short distances and lower covariance or correlation at greater distances for many spatially continuous phenomena.

More formally, if we have a spatial stochastic process $\{Y(s),\ s \in \mathcal{R}\}$ where we denote $E(Y(s))$ as $\mu(s)$ and $VAR(Y(s))$ as $\sigma^2(s)$ then the covariance of this process at any two particular points s_i and s_j is defined as:

$$C(s_i, s_j) = E((Y(s_i) - \mu(s_i))(Y(s_j) - \mu(s_j)))$$

with the corresponding correlation defined as:

$$\rho(s_i, s_j) = \frac{C(s_i, s_j)}{\sigma(s_i)\sigma(s_j)}$$

Notice that $C(s, s) = \sigma^2(s)$.

Such a process is said to be *stationary* if $\mu(s) = \mu$ and $\sigma^2(s) = \sigma^2$ (that is, the mean and variance are independent of location and constant throughout $\mathcal{R}$) and, in addition,

$$C(s_i, s_j) = C(s_i - s_j) = C(\boldsymbol{h})$$

This means that $C(s_i, s_j)$ depends only on the vector difference, $\boldsymbol{h}$, between s_i and s_j (that is, direction and distance of separation) and not on their absolute locations. $C(\boldsymbol{h})$ is often referred to as the *covariance function* or the *covariogram* of the process and the corresponding correlation $\rho(\boldsymbol{h})$ as the *correlogram*. Note that $C(0) = \sigma^2$.

The process is said to be *isotropic* if the dependence is purely a function of the distance between s_i and s_j and not the direction, that is, dependent only on the length of the vector $\boldsymbol{h}$, which we shall denote h. Then, $C(s_i, s_j) = C(h)$. If this holds, then clearly $\rho(s_i, s_j) = \rho(h)$ as well.

A weaker assumption than that of stationarity is *intrinsic stationarity* which is defined through a constant mean and a constant variance in the differences between values at locations separated by a given distance and direction. That is:

$$E(Y(s + \boldsymbol{h}) - Y(s)) = 0$$

$$VAR(Y(s + \boldsymbol{h}) - Y(s)) = 2\gamma(\boldsymbol{h})$$

The function $\gamma(\boldsymbol{h})$ is known as the *variogram*. Strictly it is the semi-variogram but the prefix 'semi' is often omitted and we shall follow this convention in our subsequent references to it (the factor 2 is simply added for convenience so that for large separations $\gamma(\boldsymbol{h})$ is equal to σ^2, rather than twice this).

For a stationary process the covariogram, correlogram and variogram are directly related by:

$$\rho(\boldsymbol{h}) = \frac{C(\boldsymbol{h})}{\sigma^2}$$

$$\gamma(\boldsymbol{h}) = \sigma^2 - C(\boldsymbol{h})$$

So, for a stationary spatial process the covariogram, correlogram and variogram provide similar information in a slightly different form (Figure 5.9). The covariogram and correlogram have the same shape, with the correlogram being scaled so that its maximum value is 1. The variogram also has the same shape as the covariogram, except that it is 'inverted'. While the covariogram starts from a maximum of σ^2 at $\boldsymbol{h} = 0$ and decreases to 0, the variogram starts at 0 and increases, at a distance referred to as the *range*, to a maximum of σ^2, often referred to as the *sill*.

In this section we are interested in using the covariogram, or variogram, simply as exploratory devices to examine spatial dependence in the observed data. Later we will see how these functions play a major role in the modelling of such data. In order to estimate $C(\boldsymbol{h})$ or $\gamma(\boldsymbol{h})$ for an observed spatial process it is in practice necessary to make some sort of stationarity assumption—we need to assume that the behaviour of the process in some parts of the space is the same as that in others, otherwise we do not have the repetition required to form an estimate of the second order properties.

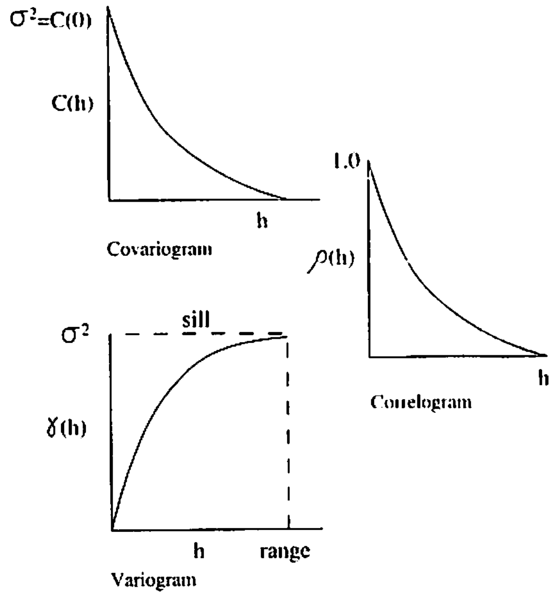

Fig. 5.9 Covariogram, variogram and correlogram

It is a matter of judgement whether to assume stationarity or intrinsic stationarity. Often, we will need to assume isotropy as well, in order to obtain a description of the covariance structure which is sufficiently simple to work with. In general, statisticians and geographers have tended to prefer to work with the covariogram or correlogram, whereas geologists and environmental scientists have favoured the variogram. It should be noted that estimates of the variogram are in general more robust to minor departures from stationarity in the form of a first order trend in the process.

The natural sample estimator of the variogram is:

$$2\hat{\gamma}(\boldsymbol{h}) = \frac{1}{n(\boldsymbol{h})} \sum_{s_i - s_j = \boldsymbol{h}} (y_i - y_j)^2$$

where the summation is over all pairs of observed data points with a vector separation of $\boldsymbol{h}$ and $n(\boldsymbol{h})$ is the number of these pairs. As $\boldsymbol{h}$ is varied, a set of values is obtained, so giving a sample variogram (Figure 5.10). Of course in practice, for irregularly spaced sample points, there will rarely be enough observations with an exact vector separation of $\boldsymbol{h}$ and so a series of intervals are used and the estimator calculated on the basis of separations which lie in these intervals, that is, which are 'close' to $\boldsymbol{h}$. In general this will involve specifying a tolerance on both distance and direction. Usually, disjoint intervals are chosen.

Under the assumption of isotropy the variogram is estimated over all directions for a given distance separation h:

$$2\hat{\gamma}(h) = \frac{1}{n(h)} \sum_{|s_i - s_j| = h} (y_i - y_j)^2$$

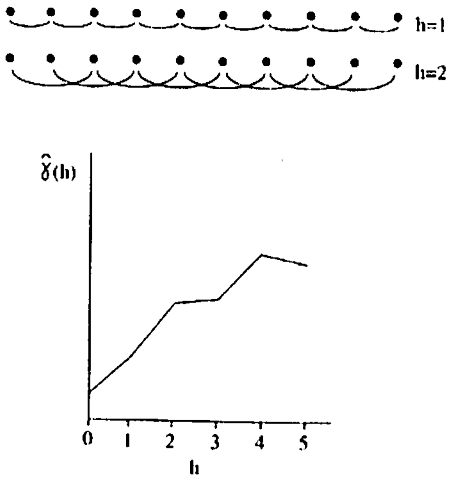

Fig. 5.10 Typical sample variogram

The equivalent estimate for the sample covariogram of a stationary process is:

$$\hat{C}(\boldsymbol{h}) = \frac{1}{n(\boldsymbol{h})} \sum_{s_i - s_j = \boldsymbol{h}} (y_i - \bar{y})(y_j - \bar{y})$$

where $\bar{y}$ is the mean of all the observed sample values and similar remarks apply concerning tolerances around $\boldsymbol{h}$ and simplification under the assumption of isotropy, as in the variogram case.

Note that in general the theoretical relationship between variogram and covariogram for a stationary process will not necessarily hold for the estimates. That is, in general, $\hat{\gamma}(\boldsymbol{h}) \neq \hat{C}(0) - \hat{C}(\boldsymbol{h})$.

Although theoretically $\gamma(0) = 0$, sampling errors and small scale variability may often cause sample values with small separations to be quite dissimilar. This causes a discontinuity at the origin of the sample variogram. This is often referred to as the *nugget effect*. Clearly a variogram consisting of pure nugget effect, that is, horizontal except at the origin, corresponds to a process with no spatial dependence.

Notice that in general $n(\boldsymbol{h})$ will increase as h, the length of the vector separation $\boldsymbol{h}$, increases, so that the reliability of the sample estimate of variogram or covariogram decreases as h decreases. Unfortunately, it is the more local behaviour which is likely to be of practical interest. In other words, in the area in which we are most interested, we have the least reliability in our estimate. One pragmatic suggestion is to attempt to smooth the variability caused by small numbers of data pairs at short distances by using a modified estimate of $\gamma(\boldsymbol{h})$ as:

$$\tilde{\gamma}(\boldsymbol{h}) = \frac{\hat{\gamma}(\boldsymbol{h})}{\bar{y}^2(\boldsymbol{h})}$$

where $\bar{y}^2(\boldsymbol{h})$ is the mean of all the data values used to calculate $\hat{\gamma}(\boldsymbol{h})$ at different values of $\boldsymbol{h}$. $\tilde{\gamma}(\boldsymbol{h})$ is referred to as the *relative variogram*.

How do all these ideas concerning covariograms, variograms and their sample estimates relate to practical spatial data analysis? One would typically begin an exploratory analysis of second order properties or covariance structure by first estimating an isotropic variogram or covariogram. This estimate is based on more sample pairs than any directional variograms and so should be less erratic and have a more interpretable structure. It should also serve to establish the most appropriate h lags at which to compute estimated values. Subsequent analysis can then proceed to calculate directional variograms in two or three broad directions in order to explore for possible directional effects or *anisotropy*. Examination of the resulting plots allows a purely informal assessment of the degree of spatial dependence in the data and any particularly strong directional effects that may be relevant in relation to this.

An additional general exploratory technique which may be useful for gaining insight into covariance structure and to back up any informal inferences drawn from the variogram, is to plot $(y_i - y_j)^2$ (or $\sqrt{(y_i - y_j)^2}$) for all possible (i,j) pairs against the distance between them h. This *variogram cloud* may reveal extreme outlying points that are dominating the estimate of the sample variogram. It may also reveal skewness in the distribution of the differences at any lag which implies that the variogram will be a poor estimate of the true covariance structure at that lag. It may also be useful to examine simple scatter plots of y_i values against 'neighbouring' y_j values at various different spatial lags.

For a stationary process the sample variogram should rise to an upper bound, the *sill* referred to earlier, corresponding to σ^2. The distance at which this occurs is referred to as the *range* (see Figure 5.9). Failure to exhibit an upper bound will indicate some degree of non-stationarity in the process. For example, if the variogram rises as an unbounded, concave-upwards curve away from the origin, this may indicate a first order 'drift' or trend in the process. In this case the trend may be removed from the data using a trend surface model, and further analysis of second order effects carried out using the residuals from this model. In order to explain what this means we need to consider some models for global trends in spatially continuous data, which we will come on to in the next section. But before doing this let us apply some of these second order exploratory tools to real data.

The sample variogram obtained for the logarithms of the radon data is shown in Figure 5.11. Notice that even at very short distances there is still a high degree of variability. In particular, there is clear evidence of a substantial 'nugget effect', the sample variogram being virtually horizontal. We might expect this, given what we know about radon gas levels as discussed earlier. We noted then that they display considerable local spatial variability, with variations even between neighbouring properties depending upon their building materials, age, insulation, and so on. The variogram simply confirms this absence of spatial continuity. In particular, this implies that there is little point in attempting to exploit the sampled values in the locality of a site at which the value is unknown, in order to form estimates of levels at that site. In fact, noting the general global trends, pointed out when we introduced this data set, is probably about as far as we can go with such data in understanding the distribution of the highly variable radon gas levels.

In contrast, the sample variogram for one of the Canadian geochemical elements (nickel) shows much more spatial continuity (Figure 5.12). Again, we have used the logarithms because of the marked skewness in the raw data. Although there is some visual evidence of a nugget effect this is much less pronounced than for the radon data. The variogram increases gradually with distance, beginning to flatten out at a distance of about 1.5–2.0 kilometres.

5.5 Modelling spatially continuous data

So far in this chapter we have been concerned with exploring spatially continuous data in a fairly informal way. In this section we consider the

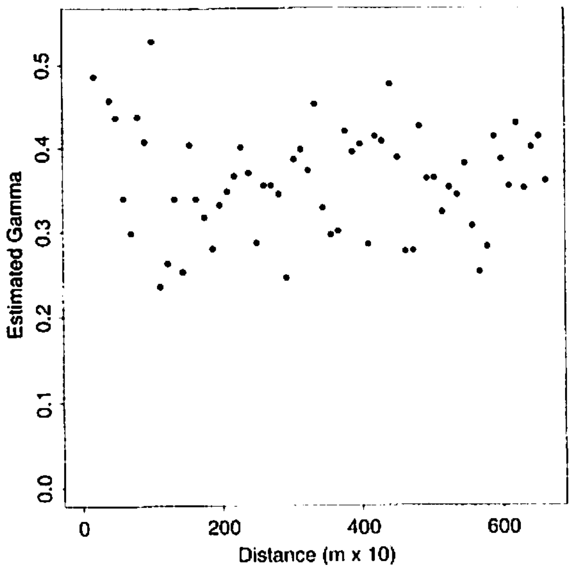

Fig. 5.11 Sample variogram for logarithms of radon levels

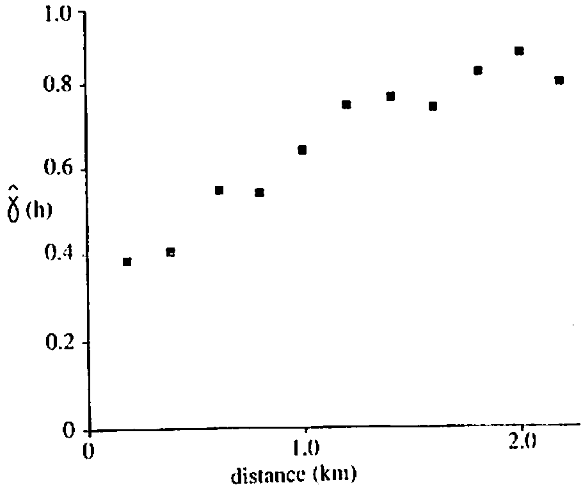

Fig. 5.12 Sample variogram for logarithms of nickel concentrations on north Vancouver Island