1
Introduction

1.1 Motivating examples

The term *spatial statistics* is used to describe a wide range of statistical models and methods intended for the analysis of spatially referenced data. Cressie (1993) provides a general overview. Within spatial statistics, the term *geostatistics* refers to models and methods for data with the following characteristics. Firstly, values $Y_i : i = 1, \ldots, n$ are observed at a discrete set of sampling locations x_i within some spatial region A. Secondly, each observed value Y_i is either a direct measurement of, or is statistically related to, the value of an underlying continuous spatial phenomenon, $S(x)$, at the corresponding sampling location x_i. This rather abstract formulation can be translated to a variety of more tangible scientific settings, as the following examples demonstrate.

Example 1.1. *Surface elevations*

The data for this example are taken from Davis (1972). They give the measured surface elevations y_i at each of 52 locations x_i within a square, A, with side-length 6.7 units. The unit of distance is 50 feet (≈15.24 meters), whereas one unit in y represents 10 feet (≈3.05 meters) of elevation.

Figure 1.1 is a *circle plot* of the data. Each datum (x_i, y_i) is represented by a circle with centre at x_i and radius proportional to y_i. The observed elevations range between 690 and 960 units. For the plot, we have subtracted 600 from each observed elevation, to heighten the visual contrast between low and high values. Note in particular the cluster of low values near the top-centre of the plot.

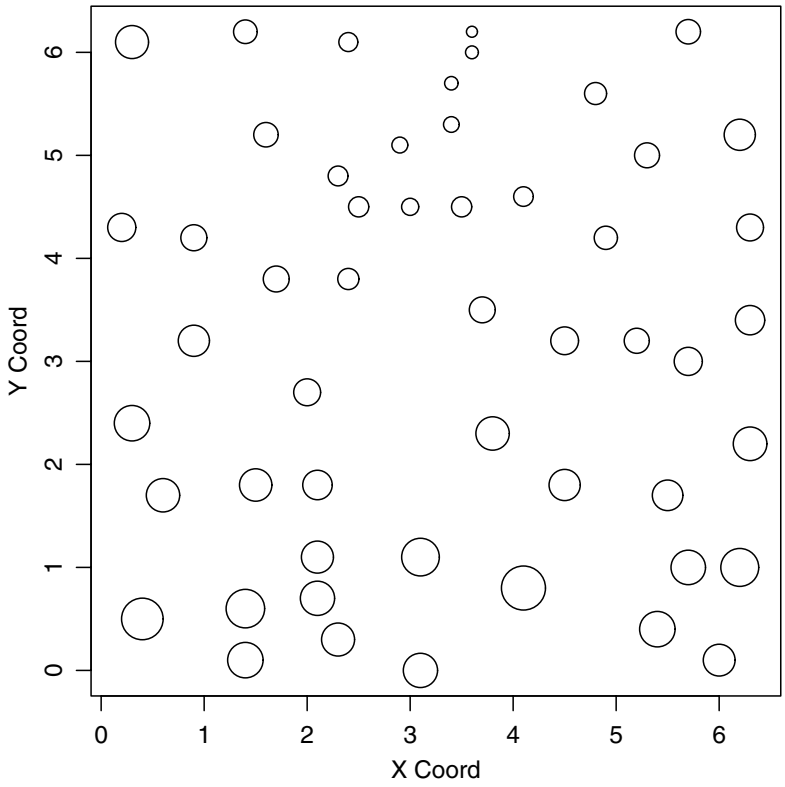

Figure 1.1. Circle plot of the surface elevation data. For the coordinates, the unit of distance is 50 feet. The observed elevations range from 690 to 960 units, where 1 unit represents 10 feet of elevation. Circles are plotted with centres at the sampling locations and radii determined by a linear transformation of the observed elevations (see Section 1.6).

The objective in analysing these data is to construct a continuous elevation map for the whole of the square region A. Let $S(x)$ denote the true elevation at an arbitrary location x. Since surface elevation can be measured with negligible error, in this example each y_i is approximately equal to $S(x_i)$. Hence, a reasonable requirement would be that the map resulting from the analysis should interpolate the data. Our notation, distinguishing between a measurement process Y and an underlying true surface S, is intended to emphasise that this is not always the case.

Example 1.2. *Residual contamination from nuclear weapons testing*

The data for this example were collected from Rongelap Island, the principal island of Rongelap Atoll in the South Pacific, which forms part of the Marshall Islands. The data were previously analysed in Diggle et al. (1998) and have the format $(x_i, y_i, t_i) : i = 1, \ldots, 157$, where x_i identifies a spatial location, y_i is a photon emission count attributable to radioactive caesium, and t_i is the time (in seconds) over which y_i was accumulated.

These data were collected as part of a more wide-ranging, multidisciplinary investigation into the extent of residual contamination from the U.S. nuclear weapons testing programme, which generated heavy fallout over the island in

the 1950s. Rongelap island has been uninhabited since 1985, when the inhabitants left on their own initiative after years of mounting concern about the possible adverse health effects of the residual contamination. Each ratio y_i/t_i gives a crude estimate of the residual contamination at the corresponding location x_i but, in contrast to Example 1.1, these estimates are subject to non-negligible statistical error. For further discussion of the practical background to these data, see Diggle, Harper and Simon (1997).

Figure 1.2 gives a circle plot of the data, using as response variable at each sampling location x_i the observed emission count per unit time, y_i/t_i. Spatial coordinates are in metres, hence the east-west extent of the island is approximately 6.5 kilometres. The sampling design consists of a primary grid covering the island at a spacing of approximately 200 metres together with four secondary 5 by 5 sub-grids at a spacing of 50 metres. The role of the secondary sub-grids is to provide information about short-range spatial effects, which have an important bearing on the detailed specification and performance of spatial prediction methods.

The clustered nature of the sampling design makes it difficult to construct a circle plot of the complete data-set which is easily interpretable on the scale of the printed page. The inset to Figure 1.2 therefore gives an enlarged circle plot for the western extremity of the island. Note that the variability in the emission counts per unit time within each sub-grid is somewhat less than the overall variability across the whole island, which is as we would expect if the underlying variation in the levels of contamination is spatially structured.

In devising a statistical model for the data, we need to distinguish between two sources of variation: spatial variation in the underlying true contamination surface, $T(x)$ say; and statistical variation in the observed photon emission counts, y_i, given the surface $T(x)$. In particular, the physics of photon emissions suggests that a Poisson distribution would provide a reasonable model for the conditional distribution of each y_i given the corresponding value $T(x_i)$. The gamma camera which records the photon emissions integrates information over a circular area whose effective diameter is substantially smaller than the smallest distance (50 metres) between any two locations x_i. It is therefore reasonable to assume that the y_i are conditionally independent given the whole of the underlying surface $T(x)$. In contrast, there is no scientific theory to justify any specific model for $T(x)$, which represents the long-term cumulative effect of variation in the initial deposition, soil properties, human activity and a variety of natural environmental processes. We return to this point in Section 1.2.

One scientific objective in analysing the Rongelap data is to obtain an estimated map of residual contamination. However, in contrast to Example 1.1, we would argue that in this example the map should not interpolate the observed ratios y_i/t_i because each such ratio is a noisy estimate of the corresponding value of $T(x_i)$. Also, because of the health implications of the pattern of contamination across the island, particular properties of the map are of specific interest, for example the location and value of the maximum of $T(x)$, or areas within which $T(x)$ exceeds a prescribed threshold.

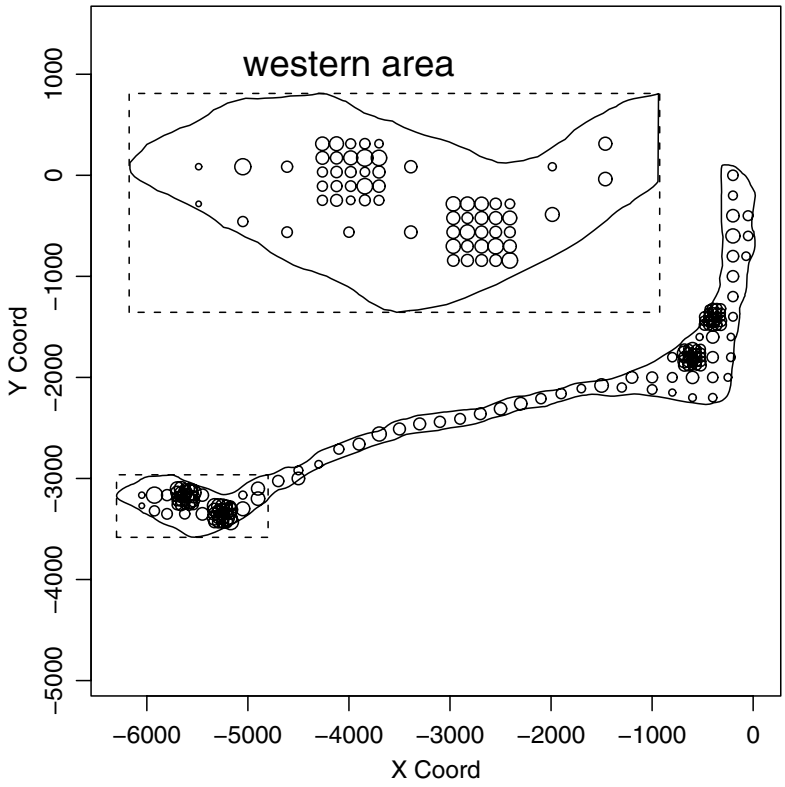

Figure 1.2. Circle plot for data from Rongelap island. Circles are plotted with centres at the sampling locations and radii proportional to observed emission counts per unit time. The unit of distance is 1 metre. The inset shows an enlargement of the western extremity of the island.

Example 1.3. *Childhood malaria in The Gambia*

These data are derived from a field survey into the prevalence of malaria parasites in blood samples taken from children living in village communities in The Gambia, West Africa. For practical reasons, the sampled villages were concentrated into five regions rather than being sampled uniformly across the whole country. Figure 1.3 is a map of The Gambia showing the locations of the sampled villages. The clustered nature of the sampling design is clear.

Within each village, a random sample of children was selected. For each child, a binary response was then obtained, indicating the presence or absence of malaria parasites in a blood sample. Covariate information on each child included their age, sex, an indication of whether they regularly slept under a mosquito net and, if so, whether or not the net was treated with insecticide. Information provided for each village, in addition to its geographical location, included a measure of the greenness of the surrounding vegetation derived from satellite data, and an indication of whether or not the village belonged to the primary health care structure of The Gambia Ministry for Health.

The data format for this example is therefore $(x_i, y_{ij}, d_i, d_{ij})$ where the subscripts i and j identify villages, and individual children within villages, respectively, whilst d_i and d_{ij} similarly represent explanatory variables recorded at the village level, and at the individual level, as described below. Note that if only village-level explanatory variables are used in the analysis, we might choose

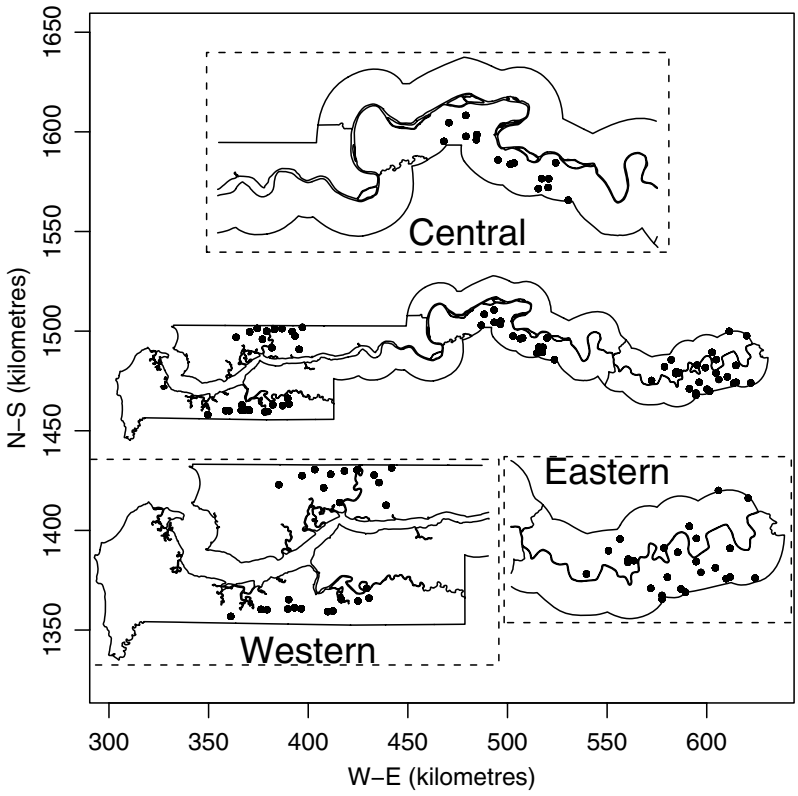

Figure 1.3. Sampling locations for The Gambia childhood malaria survey. The inset plots are enlarged maps of the western, central and eastern regions of The Gambia.

to analyse the data only at the village level, in which case the data format could be reduced to (x_i, n_i, y_i, d_i) where n_i is the number of children sampled in the ith village, and $y_i = \sum_{j=1}^{n_i} y_{ij}$ the number who test positive.

Figure 1.4 is a scatterplot of the observed prevalences, y_i/n_i, against the corresponding greenness values, u_i. This shows a weak positive correlation.

The primary objective in analysing these data is to develop a predictive model for variation in malarial prevalence as a function of the available explanatory variables. A natural starting point is therefore to fit a logistic regression model to the binary responses y_{ij}. However, in so doing we should take account of possible unexplained variation within or between villages. In particular, unexplained spatial variation between villages may give clues about as-yet unmeasured environmental risk factors for malarial infection.

Example 1.4. *Soil data*

These data have the format $(x_i, y_{i1}, y_{i2}, d_{i1}, d_{i2})$, where x_i identifies the location of a soil sample, the two y-variables give the calcium and magnesium content whilst the two d-covariates give the elevation and sub-area code of each sample.

The soil samples were taken from the 0-20 cm depth layer at each of 178 locations. Calcium and magnesium content were measured in $mmol_c/dm^3$ and the elevation in metres. The study region was divided into three sub-regions which have experienced different soil management regimes. The first, in the upper-left corner, is typically flooded during each rainy season and is no longer

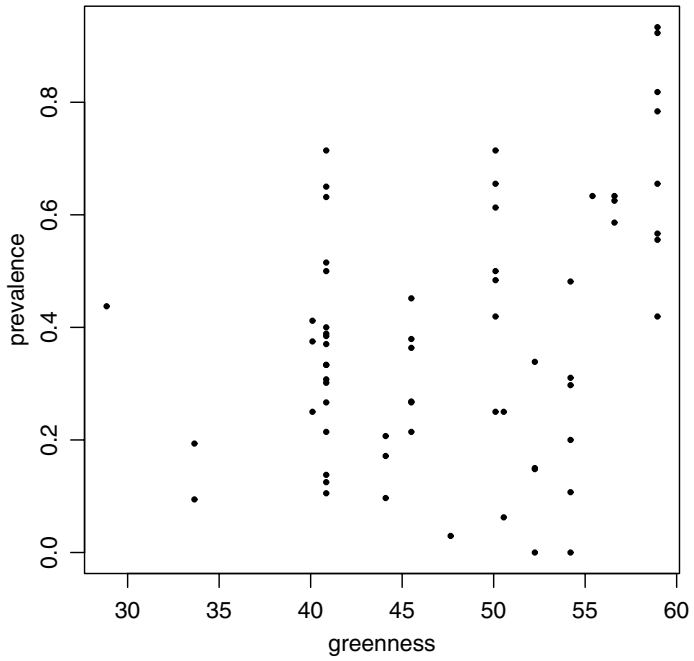

Figure 1.4. Observed prevalences against greenness for villages in The Gambia childhood malaria survey.

used as an experimental area because of its varying elevation. The calcium and magnesium levels in this region therefore represent the pattern of natural spatial variation in background content. The second, corresponding to the lower half of the study region, and the third, in the upper-right corner, have received fertilisers in the past: the second is typically occupied by rice fields, whilst the third is frequently used as an experimental area. Also, the second sub-region was the most recent of the three to which calcium was added to neutralise the effect of aluminium in the soil, which partially explains the generally higher measured calcium values within this sub-region.

The sampling design is an incomplete regular lattice at a spacing of approximately 50 metres. The data were collected by researchers from PESAGRO and EMBRAPA-Solos, Rio de Janeiro, Brasil (Capeche, 1997).

The two panels of Figure 1.5 show circle plots of the calcium (left panel) and magnesium (right panel) data separately, whilst Figure 1.6 shows a scatterplot of calcium against magnesium, ignoring the spatial dimension. This shows a moderate positive correlation between the two variables; the value of the sample correlation between the 178 values of calcium and magnesium content is $r = 0.33$.

Figure 1.7 shows the relationship between the potential covariates and the calcium content. There is a clear trend in the north-south direction, with generally higher values to the south. The relationships between calcium content and either east-west location or elevation are less clear. However, we have included on each of the three scatterplots a lowess smooth curve (Cleveland, 1981) which, in the case of elevation, suggests that there may be a relationship with calcium beyond an elevation threshold. Finally, the boxplots in the bottom right panel of Figure 1.7 suggest that the means of the distributions of calcium content are

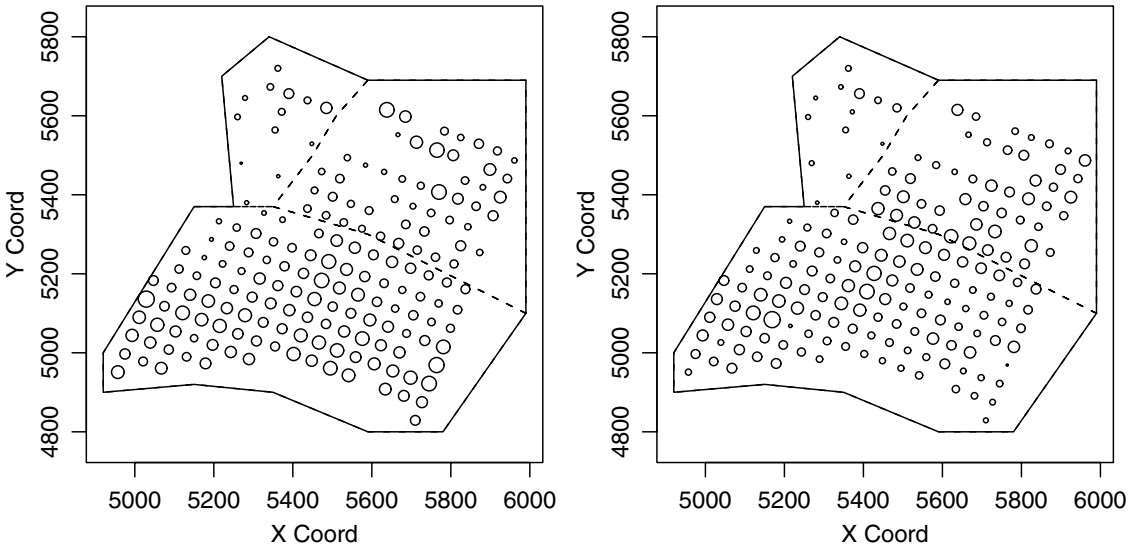

Figure 1.5. Circle plots of calcium (left panel) and magnesium (right panel) content with dashed lines delimiting sub-regions with different soil management practices.

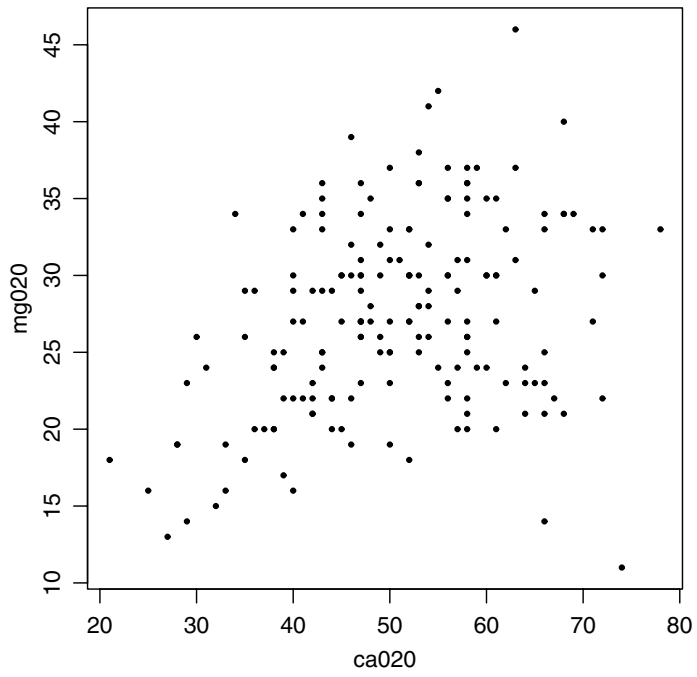

Figure 1.6. Scatterplot of calcium content against magnesium content in the 0-20 cm soil layer.

different in the different sub-regions. In any formal modelling of these data, it would also be sensible to examine covariate effects after allowing for a different mean response in each of the three sub-regions, in view of their different management histories.

One objective for these data is to construct maps of the spatial variation in calcium or magnesium content. Because these characteristics are determined from small soil cores, and repeated sampling at effectively the same location would yield different measurements, the constructed maps should not necessar-

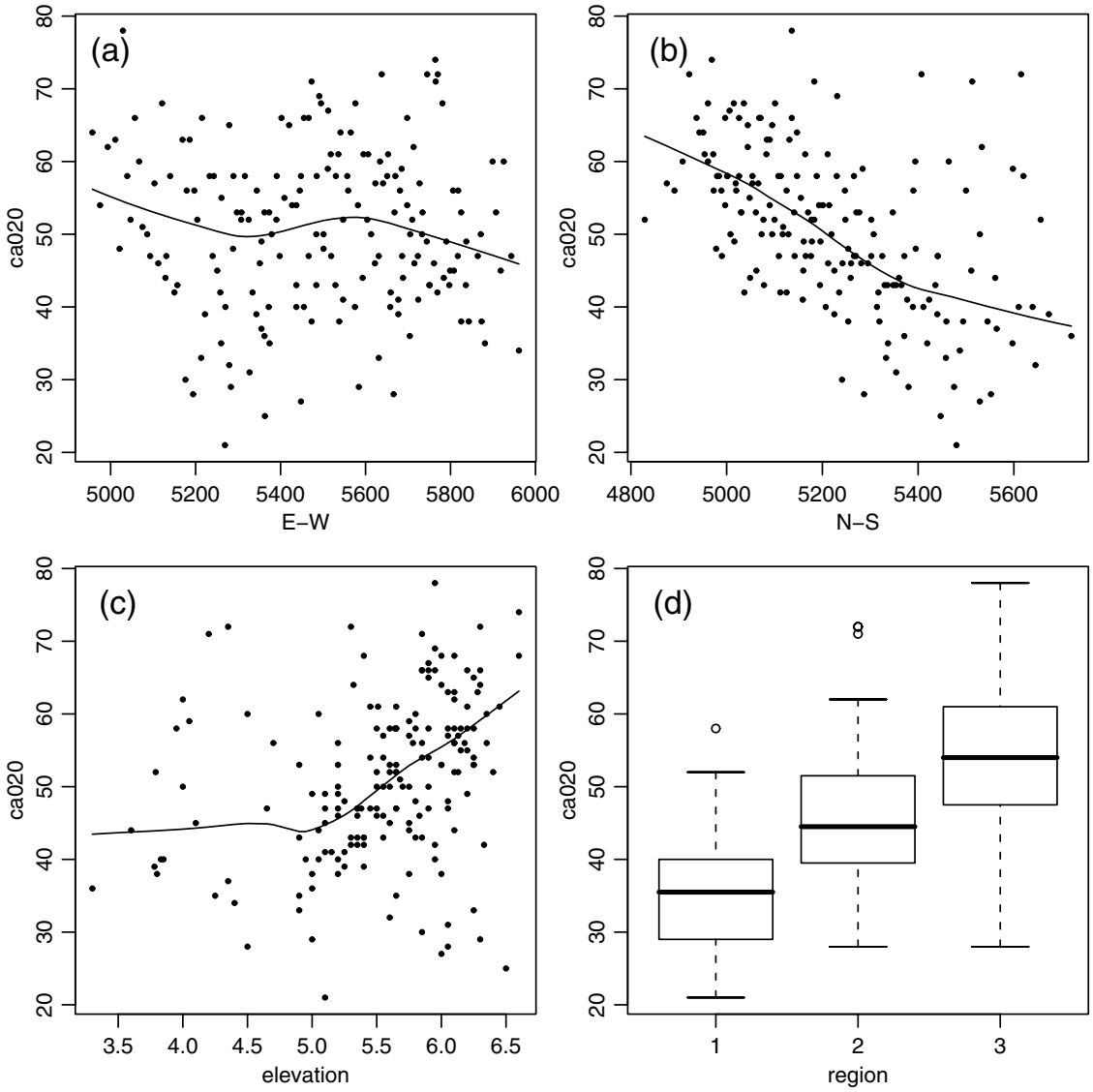

Figure 1.7. Scatterplots of calcium content against: (a) $E-W$ coordinate, (b) $N-S$ coordinate, (c) elevation. Lines are lowess curves. (d) Box-plots of calcium content in each of the three sub-regions.

ily interpolate the data. Another goal is to investigate relationships between calcium or magnesium content and the two covariates. The full data-set also includes the values of the calcium and magnesium content in the 20-40 cm depth layer.

We shall introduce additional examples in due course. However, these four are sufficient to motivate some basic terminology and notation, and to indicate the kinds of problems which geostatistical methods are intended to address.