PART B

Spatial data: obtaining data and quality issues

3

Obtaining spatial data through sampling

Spatial data are acquired in many different ways. In section 3.1 primary, secondary and other types of data sources are briefly reviewed. Section 3.2 considers the problem of obtaining data through spatial sampling. Section 3.3 briefly reviews simulation as a means of obtaining spatial 'data'.

Most of the chapter is concerned with obtaining data through sampling. Three types of attribute properties, for which spatial sampling is often needed, are considered. The first is sampling to estimate global properties of an attribute in an area such as its mean value. This arises in estimating crop yield or the proportion of an area under a particular type of land use. The second is sampling for the purpose of constructing maps and making predictions of attribute values at specific locations. This arises when designing a sampling system for rainfall (e.g. rainfall by area) or predicting yields from mining. The third is sampling to detect areas with critically high (or low) values of an attribute ('hot' and 'cold' spots). This arises when seeking to detect land parcels that have been seriously contaminated.

The chapter deals only with attribute sampling. There are situations where a sample design is needed to represent some object such as the location of a line or the boundary of a lake. In geographic information system databases the representation and resolution of boundary lines between polygons is the product of a sampling process. The detail with which any spatial object is represented is a function of the density of sample points (Griffith, 1989; Longley et al., 2001). This aspect of spatial sampling is not considered here.

3.1 Sources of spatial data

Primary data are collected by a researcher to meet the specific objectives of a project. In observational science primary data originate from fieldwork and sample surveys. If hypotheses are to be tested that have a spatial

or geographical dimension then surveys should ensure accurate and careful geo-referencing of each observation – as precise as confidentiality will allow. This will help with later stages that may require linkage with data from other surveys. If local-area and contextual influences are to be examined, then focused sampling in contrasting areas is needed. The results of national surveys when applied to local areas may not produce estimates with sufficient precision because the sample size in the local area may turn out to be small. Stratification needs to be built into the sampling strategy if local-area estimates are needed.

Secondary spatial data sources include maps, national and regional social, economic and demographic census data, data generated by public bodies such as health, police and local authorities as well as commercial data sets generated by private institutions in for example the retail and financial sectors. Even when such data (e.g. national censuses) ostensibly represent complete enumerations of the population it is sometimes safer to view them as samples. One of the benefits of this is that the analyst can invoke sampling theory and avoid the criticism of having placed too much emphasis on any particular data set when it is known that had the data been collected for even a slightly different time period, counts would almost certainly have been different (Craglia et al., 2002). The huge growth in certain types of secondary data, generated by public agencies, has frequently been remarked upon.

Satellites are an important source of environmental data and are useful in conjunction with socio-economic, topographic and other ancillary data in constructing descriptions of, for example, urban areas (Harris and Longley, 2000). These developments owe much to modern devlopments in hardware and the creation of geographic information systems that allow the handling, including linkage, of large geographically referenced data sets. Decker (2001) provides a review of data sources for GIS and Lawson and Williams (2001) for spatial epidemiology. With data integration, the process of assigning different spatial data sets to a common spatial framework, comes a range of technical issues about how such integration should be implemented and the reliability of such integrated data sets (Brusegard and Menger, 1989). Data sets may be of varying qualities, have different lineages, different frequencies of collection as well as be on different spatial frameworks that change over time.

Some data are generated by inputing sample data into a model to generate a spatial surface of data values. This may be done if the variable of interest is difficult or expensive to collect. Air pollution monitoring is expensive. Air pollution maps for an area are constructed by combining data on known point, line and area sources of air pollution with climatological data and assumptions about how pollutants disperse. After calibrating and validating model output against such sample data as are available, model output is then used to provide

maps of air pollution (Collins, 1998). If a probability model is specified for a variable of interest then not only will the average surface be of interest but also variability about the average. Simulation methods are used to display this variability (see section 3.3).

3.2 Spatial sampling

3.2.1 The purpose and conduct of spatial sampling

The purpose behind spatial sampling is to make inferences about a population where each member has a geographical reference or geo-coding, on the basis of a subset of individuals drawn from that population. Sampling is used rather than say undertaking a complete census for various reasons. The population may be so large a complete census would be physically impossible or impractical (e.g. estimating the average length of pebbles on a beach). There may be an (uncountable) infinity of locations where measurements could be taken as in the case of ground-level air quality, or soil depth in an area with a continuous covering of soil. The cost of acquiring information on each individual may rule out a complete census. The 1991 UK Census only provides data on household employment on the basis of a 10% sample of all the returns partly for confidentiality reasons but also because of the costs of manual coding. Remotely sensed data provides a complete census (at a given resolution) but for cost reasons the data are interpreted by ground truthing based on a sample of sites.

In other situations it is not the size of the population or the cost of acquiring the data that calls for sampling but the level of precision, on the quantity of interest, required by the application. Sampling introduces error in the sense that the property of interest is estimated to within some level of precision – the inverse of the error variance or sampling error associated with the estimator. This error variance can be held to pre-determined limits by the choice of sample size. Taking a complete enumeration (or even a very large sample) may be wasteful of effort if such accuracy is not necessary. Further, the accuracy of a census may be illusory if measurement error is present, and, in a reversal of what might be considered the normal relationship between census and sample, sampling may be needed to improve the quality of 'census' information. In the case of crime data, counts based on police records are known to produce undercounts so household sampling is undertaken in order to improve estimates. Population censuses miss certain groups (such as the homeless) and sampling may be undertaken to improve data on them.

Drawing inferences about a geographical population through sampling calls for a series of decisions to be taken on the sample design. These decisons

are taken in relation to the following questions: (i) What is to be estimated? (ii) What sample size (n) is required in order to achieve the desired level of precision? (iii) Since the sampling is spatial, what n locations should be selected for the sample? (iv) What estimator should be used to compute the quantity of interest? (v) What measure of distance (between the estimate and the attribute of interest in the population) should the sample design seek to minimize? The answers to these questions are not necessarily independent of each other. The choice of sampling plan, (iii), is dependent on the selected estimator, (iv), and whether it is a spatial or non-spatial property of the population that is of primary interest, (i). An example of a non-spatial property of a population is the mean level of an attribute in an area or the proportion of the population that exceeds a certain threshold value. They are termed non-spatial properties because the analyst is only interested in 'how much' not 'where' (Brus and de Gruijter, 1997). However spatial properties of a population involve 'where' questions: identifying where in the population threshold values of an attribute are exceeded or where the extreme values are located, and being able to make optimal predictions of attribute values at unsampled locations. 'Where' questions extend to constructing maps of population variability or providing quantitative summaries of that variability in terms of autocorrelation or semi-variogram functions (see chapter 2).

There will be other issues that influence the design of a spatial sample. Increasing estimator precision, by increasing the sample size, always raises sampling costs but the extra costs of collecting larger volumes of data may be particularly important in spatial sampling. There may be problems of accessibility to certain sites and transport costs associated with sampling a geographically dispersed population. It is usually necessary to make a trade-off between economic and statistical criteria in designing a sample. Cressie (1991, pp. 321–2), briefly reviews attempts to formalize such decision making.

Sampling designed to give acceptable levels of sampling error for a regional survey will not achieve the same level of sampling error at subregional levels (because of fewer observations) and unless stratification is incorporated into the sample design some areas are likely to have very few, in some cases no, samples. A city-wide survey can be designed to deliver acceptable levels of precision for the city as a whole, but unless there is stratification say at the ward scale then intra-city comparisons across wards may be undermined by a lack of sample size at the ward level or highly varying numbers of samples between the wards so that estimator precision varies greatly from ward to ward. It is at the sample design stage when attention must be paid to deciding whether intra-area comparisons are important and what aspects of spatial variation are important. If inter-area comparisons are important, the sample must be

stratified according to the geographic units to be compared and sample sizes selected for each area to achieve the desired level of precision for the estimator. This will increase the costs of the survey and may lead to levels of precision for the area-wide estimates that are higher than really necessary. Research, particularly at small spatial scales, may want to consider the effects of other nearby ecological systems. If inter-area comparisons in terms of preventative health behaviours are between deprived and affluent neighbourhoods, the researcher may want to differentitate between deprived neighbourhoods that are spatially embedded within other deprived neighbourhoods and deprived neighbourhoods that are spatially embedded within more affluent neighbourhoods. This calls for further levels of spatial stratification.

Even if an estimator of a city-wide attribute, based on a large sample, has high precision, if it is used as the estimator for the value of the same attribute at the ward level it is likely to be a biased estimator. The estimator that uses just the subset of data taken from a particular ward and is an unbiased estimator of the ward-level attribute value is likely to have low precision. Low precision in ward-level estimators is likely to be problematic for the analyst who wants to examine differences at the ward scale. Part of the solution may lie in stratifying the sampling plan and setting sample sizes by strata to ensure appropriate levels of precision. Further improvements can be made by combining the evidence of two scales of sampling (local area and regional) using estimators that combine information or 'borrow strength' for the purpose of local- or small-area estimation (Ghosh and Rao, 1994; Longford, 1999). The city-wide estimator and the ward-level estimator can be combined in a new estimator for the ward-level quantity of interest that weights the two estimators in a way that reflects their relative precision. In this sense the ward-level estimator which has lower precision (larger error variance), 'borrows strength' from the higher precision city-wide estimator. An influential paper by James and Stein (1960) provides an early development of this approach. Gelman et al., 1995, pp. 42–4 provide a Bayesian treatment of this problem. Either the ward-level estimator can be viewed as having been 'shrunk' towards the city-wide estimator (the prior mean) or it can be viewed as the city-wide estimator 'adjusted' towards the observed value of the ward-level estimator. There is further discussion of this in sections 9.1.4 and 10.3.

Designing a sample, particularly a sample for monitoring purposes keeping records through time as well as across a region, may not always start from a 'blank sheet of paper'. There may be a set of sample sites already in position. The problem may be to cut back an existing sampling plan because of costs or because the network is unnecessarily dense. The converse problem is how to add to an existing network of sample points (Arbia and Lafratta, 1997). If

the objective is to provide good coverage of the area then kriging theory can be used to identify areas on the map where the most serious gaps exist or where there is overprovision. The starting point is to analyse spatial variation in prediction error given the existing network and then from this to identify which areas need more sample sites (because prediction errors are unacceptably high) or which areas can shed sites because the data for some sites can be predicted from the data generated at other sites (see below, section 3.2.4(c)). There are good examples of these types of problem in the area of rainfall monitoring, particularly for water quality assessment and flood control (Rodriguez-Iturbe and Mejia, 1974; O'Connell et al., 1979; Hughes and Lettenmaier, 1981). The picture is complicated however if the monitoring stations are required to monitor a range of attributes that have different spatial variablity or if the events generating the attributes occur with differing spatial properties. Monitoring rainfall events is complicated by the fact that different types of rainfall events have different spatial properties. Frontal rainfall creates large-scale patterns that could presumably be monitored by a relatively sparse network. Convection rainfall is often associated with highly localized but very intense patterns of rainfall (Bras and Rodriguez-Iturbe, 1976).

Pilot surveys play an important role in survey research, particularly for purposes of evaluating questionnaires. In the case of spatial sampling, the identification of an optimal sampling plan often depends on the pattern of spatial variability in the population. This is usually unknown. In the case of spatial sampling a pilot survey may be needed in order to estimate the pattern of spatial variability in the population in order to make a judgement as to the most appropriate sampling plan. This is likely to be of particular importance where the intention is to put in place a medium- to long-term system for monitoring.

3.2.2 Design- and model-based approaches to spatial sampling

(a) Design-based approach to sampling

The *design-based approach* or classical sampling theory approach to spatial sampling views the population of values in the region as a set of unknown values which are, apart from any measurement error, fixed in value. Randomness enters through the process for selecting the locations to sample. In the case of a discrete population, the target of inference is some global property such as:

$$(1/N)\Sigma_{k=1,\ldots,N} z(k) \tag{3.1}$$

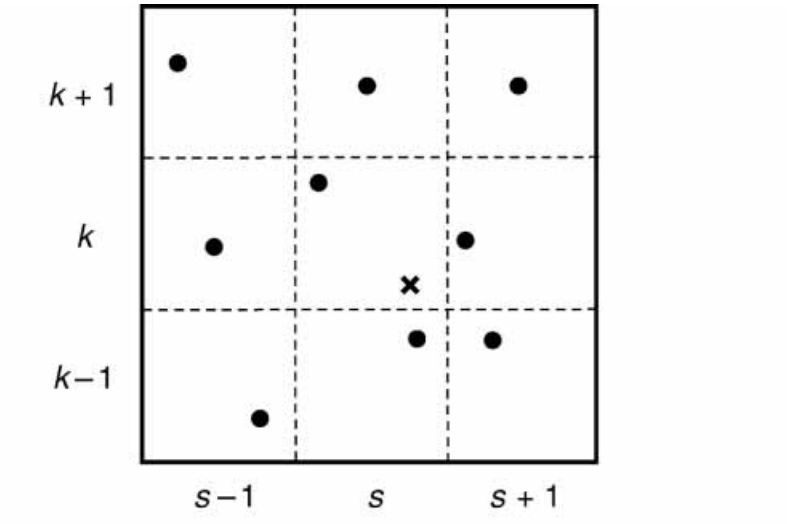
where N is the number of members of the population, so (3.1) is the population mean. If $z(k)$ is binary depending on whether the kth member of the population

is of a certain category or not, then (3.1) is the population proportion of some specified attribute. In the case of a continuous population in region A of area $|A|$ then (3.1) would be replaced by the integral:

$$(1/|A|)\int_A z(x)\,dx \tag{3.2}$$

Design-based *estimators* of quantities (3.1) or (3.2) weight the sample observations by their probabilities of being included in the sample. The merits of different sampling plans in a design-based sampling strategy depend on the structure of spatial variation in the population as will be discussed in section 3.2.4.

The design-based approach is principally used for tackling 'how much' questions such as estimating (3.1) or (3.2). In principal, individual $z(k)$ could be targets of inference but, because design-based estimators disregard most of the information that is available on where the samples are located in the study area, in practice this is either not possible or gives rise to estimators with poor properties (see figure 3.1).



Design-based sample with one observation per strata. In the absence of spatial information the point ✖ in strata (k,s) would have to be estimated using the other point in the strata (k,s) even though in fact the samples in two other strata are closer and may well provide better estimates.

Figure 3.1 Using spatial information for estimation from a sample

(b) Model-based approach to sampling

The model-based approach or superpopulation approach to spatial sampling views the population of values in the study region as but one realization of some stochastic model. The source of randomness that is present in a sample derives from a stochastic model. Again, the target of inference could be (3.1) or (3.2). Under the superpopulation approach, (3.1) for example now represents the mean of just one realization. Were other realizations to be generated, (3.1) would differ across realizations. Under this strategy, since (3.1) is a sum of random variables, (3.1) is itself a random variable and it is usual to speak of *predicting* its value.

A model-based sampling strategy provides predictors that depend on model properties and are optimal with respect to the selected model. Results may be dismissed if the model is subsequently rejected or disputed. Matern (1986, p. 69) remarks that model-based predictors of (3.1) should only be used when detailed knowledge is available about the structure of the underlying population – information that is rarely available. Hansen et al. (1983, p. 792) further suggest that design-based methods lose relatively little efficiency as compared with model-based methods even when the models are perfect descriptors of the data.

In the model-based approach it is the mean (μ) of the stochastic model assumed to have generated the realized population that is the usual target of inference rather than a quantity such as (3.1). This model mean can be considered the underlying signal of which (3.1) is a 'noisy' reflection. Since μ is a (fixed) parameter of the underlying stochastic model, if it is the target of inference, it is usual to speak of *estimating* its value. In spatial epidemiology for example it is the true underlying relative risk for an area rather than the observed or realized relative risk revealed by the specific data set that is of interest. Another important target of inference within the model-based strategy is often $z(i)$ – the value of Z at the location i. Since Z is a random variable it is usual to speak of *predicting* the value $z(i)$.

There is now no need for randomness in the sample plan. Whereas in the design-based approach the surface never changes and the evaluation of a sampling strategy must consider taking repeated probability samples, in the model-based approach each realization produces a new surface of values and the same sampling plan could be adopted in each case (Brus and de Gruijter, 1997). In fact even in cases where a complete 'census' has been taken – for example, a count by area of all new cases of a disease in a given period of time – this may still be viewed as a sample from the underlying model and the data used to estimate the parameters of that model. As implied above, the model-based approach is also of importance for tackling 'where' questions. Model-based

estimators utilize much of the spatial information available in the sample. They weight the sample observations using both the model that is presumed to be generating the data and the configuration of the sample locations. This is the reason they are particularly important for spatial prediction.

(c) Comparative comments

As Brus and de Gruijter (1997) note there has been considerable discussion about appropriate methods of spatial sampling. In the discussion to their paper, Laslett provides historical context within geostatistics and soil science. Model-based spatial sampling has its origins in mining and the development of geostatistics in order to cope with the effects of 'convenience' sampling when the purpose is spatial interpolation. Classical sampling theory applied in this context tends to overestimate large values and underestimate small values. Adopting a model-based approach provides the required regression to the mean. In areas where sampling is more uniform, such as in soil sampling for example, the benefits of the geostatistical approach to spatial interpolation may not be so evident (see section 4.4.2(*v*)).

Other discussions, particularly in a social science context, have drawn attention to the conceptual meaning of basing inference on a superpopulation view (Galtung, 1967). The mathematical model assumes a 'hypothetical universe' of possible realizations that in practice are never observed and a sample of size one that permits valid inference about model parameters only under certain conditions (see section 9.1.2). However Bernard in the discussion to Godambe and Thompson (1971) remarks: 'one is rarely . . . concerned with the finite de facto population of the UK at a given instant of time: one is more concerned with a conceptual population of people like those at present living in the UK'. Notwithstanding this comment the important issue is again whether a model-based strategy provides sounder inference on the properties of interest. In those areas of research where there is inherent randomness in the underlying processes the model-based approach is the appropriate choice.

In summary a model-based sampling strategy should be used for predicting values at particular locations, mapping and for estimating the parameters of the underlying stochastic model (such as the model mean μ) but not quantities like (3.1) or (3.2) unless the model is known. Design-based sampling should be used for estimating global properties of the (realized) population of values such as the population mean (3.1) or (3.2) but not for estimating individual values or mapping. In the next chapter, in the context of data completeness and estimating missing values, the problem of spatial prediction and estimation will be revisited (see section 4.4).

3.2.3 Sampling plans

The main classes of sampling plan for estimating a map property are random, stratified random and systematic and these will be considered first (Ripley, 1981). Sampling may take the form of points or quadrats. For a discussion of quadrat sampling see for example Kershaw (1973).

Under random sampling, *n* sites are selected so that each member of the population has an equal and independent chance of selection (figure 3.2(a)). Under stratified random sampling the population to be sampled is partitioned into areal strata and, within each of these, sites are selected according to the method of random sampling. Figure 3.2(b) shows a stratified random plan based on nine strata with one sample taken from each stratum. Figure 3.2(c) shows one type of systematic sampling, centric systematic sampling, with square strata

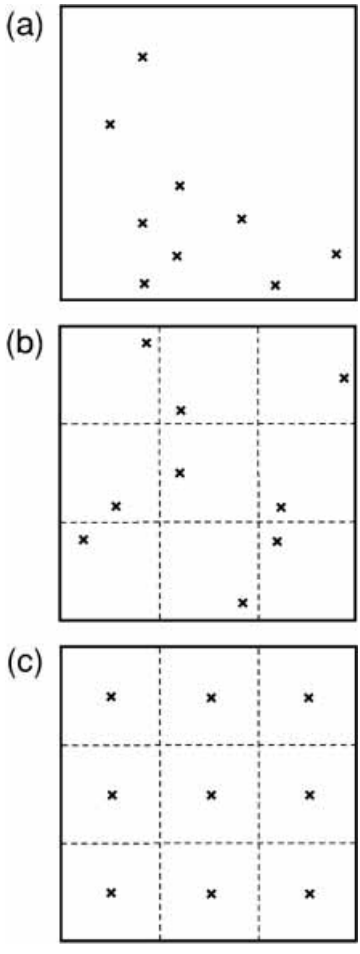


Figure 3.2 Spatial sampling: (a) random, (b) stratified random, (c) systematic

and one point taken from the centre of each stratum. Randomization can be introduced by selecting the site in the first stratum at random. The remaining $(n-1)$ sample points occupy the same relative positions in their respective strata as the first. There are many variants of this form of sampling (Koop, 1990). In the case of both stratified random and systematic sampling, other strata shapes may be adopted, such as hexagonal or triangular strata. In social and economic applications the strata may be chosen to capture different regions, such as administrative areas, so that comparisons can be made between them.

Random sampling does not ensure even coverage of the area to be sampled and in fact often gives rise to relatively large unsampled areas and to groups of sample sites that appear geographically clustered. Stratification reduces both these problems although there can still be some evidence of gaps and clusters in the spatial distribution. Systematic sampling is often easy to implement and provides well-defined directional classes and large numbers of samples separated by specified distances.

Other sampling plans may be appropriate in particular circumstances, although with the proviso that there can be problems with computing sampling variances. In social science research, cluster sampling is often used. If the population is grouped into clusters, rather than trying to sample by one of the above methods the clusters are first sampled. The sample of individuals is then drawn, often at random, from the individuals in the selected clusters. This is similar to stratified random sampling except that there are two stages of randomization and at the first stage some of the strata are randomly removed from the sample. Kahn and Sempos (1989) give the example of a health survey involving a large number of economically similar but geographically dispersed villages. Instead of drawing a sample of size n from the whole population that could involve costly and time-consuming travel to visit all the villages, a subsample of villages is drawn and the sample of size n is drawn at random from within this subset. This method of sampling will give good population estimates if the villages are just random assemblages of members of the population. It will not give good population estimates if the villages are economically more homogeneous than the population as a whole – that is, if there is strong positive correlation between the members of each cluster.

Variation in the population may be associated with different spatial scales or spatial hierarchies and the contribution from each of these scales may be the target of inference. In plant ecology areas are exhaustively divided into small blocks (the smallest scale the analyst is interested in) and then these blocks can be combined to provide a sequence of blocks of increasing size (area). Analysis to detect, for example, vegetation patch size proceeds using this census. For a discussion of the methodology see for example Cressie, 1991, pp. 591–7.

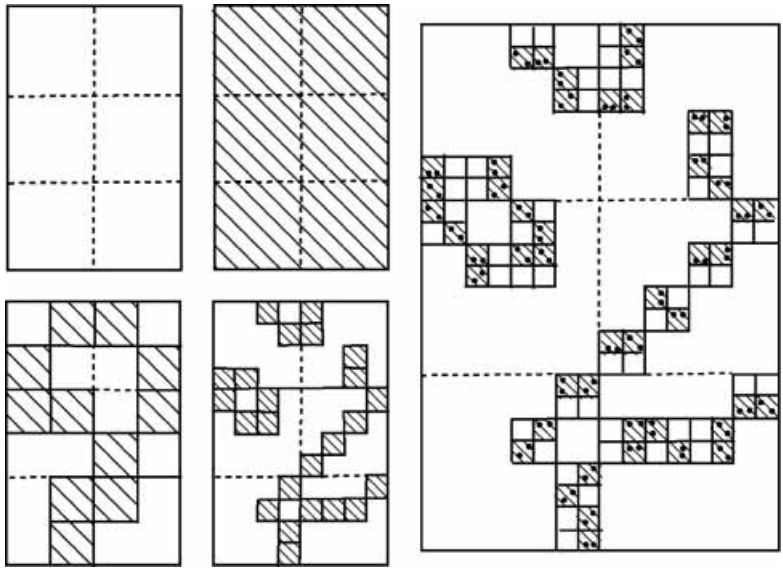


Figure 3.3 A nested sampling scheme

However, if a large range of spatial scales are to be examined necessitating a fine partition of the area then considerable data collection could be involved. One solution is to follow a nested sampling approach.

Nested sampling requires that the population is divided into blocks (level 1) which are then subdivided (level 2) and then level 2 blocks are subdivided into level 3 blocks and so on. Blocks at any level nest within blocks at the higher level. Figure 3.3 shows a form of nested spatial sampling. Each randomly selected datapoint contains variation that derives from each of the levels and the contribution from each level can be estimated by hierarchical analysis of variance. Data requirements are reduced whilst preserving the ability to analyse different spatial scales.

A problem with the scheme in figure 3.3 is that the sampling plan does not control for distance between the samples so scale effects may be masked or at least confounded by the variation in inter-sample distances that exist even at the same scale. Youden and Mehlich (1937) proposed a sampling plan where the distance between pairs is fixed. Primary sampling points are fixed at a specified distance apart. From each of these, further sample sites are randomly chosen at a fixed distance (randomly chosen direction) from the primary sites and from these other sites are selected a fixed distance apart as shown in figure 3.4. This analysis can be used to show at what spatial scale most variation occurs (Webster and Oliver, 2001, pp. 93–4). Miesch (1975) shows how accumulating

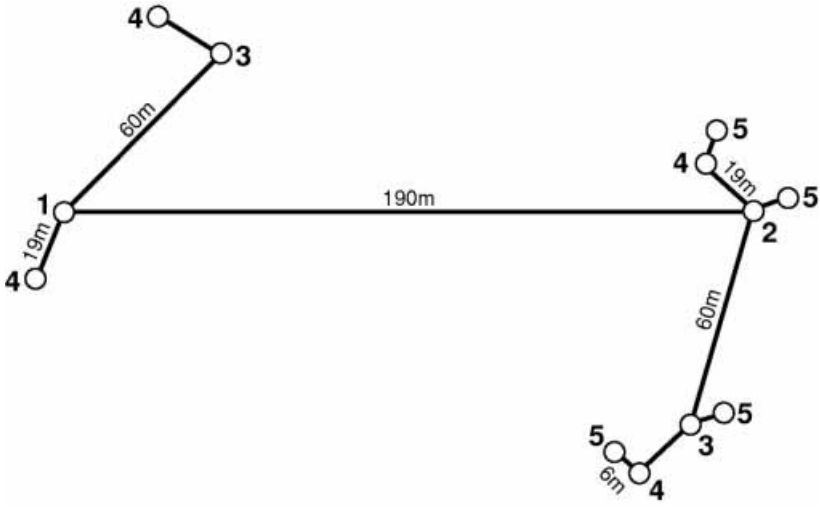


Figure 3.4 Fixed interval sampling

the different components of variance can provide rough estimates of the variogram at different distances. This can then be used to plan a survey to obtain a more careful estimate of the variogram, concentrating sampling effort where the variation occurs (Oliver and Webster, 1986).

Spatial variability in the levels of explanatory variables can be used to explore relationships between explanatory and response variables (see section 1.1). One of the contributions of geographic variation to furthering scientific understanding is to provide the researcher with a natural laboratory through which to look for possible relationships between variables. It will be important therefore to select sampling sites that whilst controlling for other factors achieves good variability on the explanatory variables of interest. This is a form of stratified sampling where the stratification is in terms of the variablity of explantory variables. The Harvard 'six cities' study, exploring the relationship between air pollution and mortality, selected cities that differed markedly in terms of the average levels of air pollution their populations were exposed to (Dockery et al., 1993).

3.2.4 Selected sampling problems

(a) Design-based estimation of the population mean

We consider the problem of estimating (3.1) which we now denote $\bar{Z}$. Results given here are for a single area or region that may be subdivided into strata. These results can be extended to the case where separate samples have

been taken in different areas or regions and the analyst wants to draw comparisons between the areas (Griffith et al., 1994). Significant differences between areas might be identified, for example where confidence intervals do not overlap.

An intuitively simple and robust estimator is given by:

$$(1/n)\Sigma_{k=1,\ldots,n} z(k) = \bar{z} \tag{3.3}$$

This is in fact the design-based estimator under random sampling where the probability of inclusion for an individual in the sample is n/N (Brus and de Gruijter, 1997).

Under random sampling and assuming the individuals are independent, from classical sampling theory, (3.3) is an unbiased estimator of $\bar{Z}$. The error variance of $\bar{z}$ as an estimator of $\bar{Z}$ is $E[(\bar{z} - \bar{Z})^2] = \sigma^2/n$ where σ^2 is the population variance. It can be shown, again from classical sampling theory, that s^2/n, where:

$$s^2 = (1/(n-1))\Sigma_{k=1,\ldots,n}(z(k) - \bar{z})^2 \tag{3.4}$$

provides an unbiased estimator of this error variance. These results can be used to place confidence intervals on the estimates, since, for large n, $\bar{z}$ is normally distributed (see, e.g., Freund, 1992). As sample size (n) increases estimator precision increases.

If the individuals are not independent as would be the usual expectation for a spatial population, then the error variance of $\bar{z}$ as an estimator of $\bar{Z}$ with finite population correction $f = n/N$ is (Dunn and Harrison, 1993, p. 595):

$$((1-f)/n)(N/(N-1))[\sigma^2 - (2/N(N-1))\Sigma_j \Sigma_{k(j<k)} \mathrm{Cov}(z(j), z(k))] \tag{3.5}$$

where the second term inside the square brackets measures the average covariance between all pairs of individuals in the population. For large N, the term in front of the square brackets in (3.5) simplifies to $(1/n)$ and again s^2/n provides an unbiased estimator of this error variance (Haining, 1988, p. 579). For the continuous space analogue of (3.5) where $\bar{z}$ is the estimator of (3.2) see Ripley (1981, p. 32).

The result (3.5) is based on taking expectations both over the positioning of the randomized sample points and over the distribution of the values $\{Z(k)\}$ which have a constant mean (μ) and spatial covariance that depends only on distance separation and possibly direction. Error variances for $\bar{z}$ as an estimator of $\bar{Z}$ can be derived by the same methods also under stratified random and systematic sampling. There are reasons to anticipate that these two sampling methods should outperform random sampling. As noted in section 3.2.3, random

sampling leaves unsampled 'holes' in the sample plan whilst also having clusters of sample points in others (see figure 3.2(a)). This seems inefficient because given the dependency between near neighbour individuals this means that random sampling produces some information redundancy. Adjacent sampled individuals carry 'overlapping' amounts of information about the population. Stratified random sampling goes some way to ensuring a more uniform coverage of the population and reducing information redundancy. Systematic sampling by fixing the interval between sample points in adjacent strata carries this process further.

Results corresponding to (3.5) for stratified and systematic sampling are given by Dunn and Harrison (1993, p. 595) and for the continuous space analogue by Ripley (1981, p. 23). In the case of stratified random sampling the second term inside the square brackets in (3.5) is replaced by the average across all strata of the average covariance between all possible pairs of individuals *within a stratum*. In the case of systematic sampling the error variance is the average covariance between individuals in the systematic sample minus the second term inside the square brackets in (3.5).

From these results it follows that the design-based error variance (using the estimator $\bar{z}$) is minimized in the case of stratified random sampling by taking small strata – in order to maximize the average within-stratum covariance. In the case of systematic sampling the design-based error variance is minimized when samples are taken far enough apart so that the covariance between elements of the same systematic sample are as small as possible. The conclusion is that stratification ensures that both stratified random and systematic sampling will outperform random sampling. Ripley (1981, p. 25) expects systematic sampling to be best in the presence of strong local positive spatial correlation unless there is spatial periodicity in the population. Empirical evidence in Dunn and Harrison (1993) confirm the benefits of these two methods over purely random sampling but show that their relative efficiencies, with respect to one another, are far less clear cut. They suggest that their relative performance 'appear to reflect the complex and varied autocorrelation functions of the data' (p. 600).

Estimates of error variances (s^2/n) for stratified random and systematic sampling can be obtained by applying (3.4) to each stratum and averaging over each of these strata. There must be at least two sample points in each stratum if (3.4) is to be computable. If there is only one sample per strata then one option is to impose larger strata (post hoc stratification) of size two for the purpose of estimating s^2 (Ripley, 1981, pp. 26–7). These methods do better than treating the stratified sample as if it were a random sample and computing (3.4) once on the whole data set. A 95% confidence interval is obtained by taking the square root

of the estimated error variance ($s/n^{1/2}$) and multiplying by 1.96, the value from the standard normal tables.

The estimator (3.3) can be used for estimating the proportion of the population which has some property, that is when $z(k)$ is 0 or 1. In this case s^2 defined by (3.4) simplifies to:

$$(n/(n-1))\,[p(1-p)] \tag{3.6}$$

where p is the proportion of 1s in the sample of size n so that $\Sigma_{k=1,\ldots,n} z(k) = \Sigma_{k=1,\ldots,n} z(k)^2 = np$. The study by Dunn and Harrison (1993) evaluates post hoc stratification (of size 2) for binary data. Equation (3.6) is computed for each stratum – so that p is computed separately for each stratum – and then averaged over the strata. This quantity is then divided by n to obtain the estimated error variance. They show that this method overestimates the true sampling error. However this estimator does better than treating the sample as if it were a random sample of size n and computing $p(1-p)/(n-1)$ as the estimated error variance. They conclude however that further evaluation of methods is needed.

(b) Model-based estimation of means

There are now two kinds of means to consider – the mean of the realized values ($\bar{Z}$) and the mean of the underlying model (μ).

The best linear unbiased predictor (BLUP) of Z under the model-based approach to spatial sampling is given by kriging theory (Cressie, 1991, p. 173). It is a predictor of the form:

$$\Sigma_{k=1,\ldots,n} a(k) z(k) \tag{3.7}$$

where the $\{a(k)\}$ depend on the *known* spatial stochastic model and hence on the spatial properties of the underlying model. Another way to look at (3.7) is as follows. The population of realized values is $z(1), \ldots, z(N)$ and suppose the sites have been indexed so that $z(1), \ldots, z(n)$ denote the values in the sample. The BLUP of $\bar{Z}$ can be expressed in the form:

$$(1/N)\,[\Sigma_{k=1,\ldots,n}\, z(k) + \Sigma_{k=n+1,\ldots,N}\, \tilde{z}(k)] \tag{3.8}$$

where $\tilde{z}(k)$ is the BLUP of an unsampled $z(k)$. The BLUP of $z(k)$, $\tilde{z}(k)$ and its prediction error will be discussed further in section 3.2.4(c). However, there may be two reasons to prefer (3.3) and the sampling theory developed in the previous section to an estimator such as (3.7). First the prediction errors associated with (3.7) will be too large if the idea of other realizations is controversial in the application. This is because the prediction errors are calculated over all

possible realizations and not just with respect to the one outcome that has in fact occurred (Isaaks and Srivastava, 1989, pp. 506–13). Second, in practice the model which determines the $\{a(k)\}$ in (3.7) is rarely known. By contrast in the design-based strategy $a(k) = 1/n$ for all k and follows from the probability of inclusion in the sample which *is* known.

The estimator for μ in the case of a normal population will be discussed in chapter 8. At this point, note that (3.3) is an unbiased estimator of μ. The problem with this estimator is that s^2/n underestimates the sampling error if observations are spatially correlated. If a large amount of data are available, one solution is to sample an independent subset using stratification and use (3.3) together with s^2/n. This will avoid the need to identify a model for the spatial dependence, be easier to implement, and if n is chosen large enough provide an estimator with the desired precision for the application.

(c) Spatial prediction

The problem considered here is the selection of a sampling plan for an attribute of interest so that good predictions can be made for unsampled sites and a good map of the attribute can be drawn. By a 'good' map is meant that 'best' predictors are chosen (they have the property that they minimize the mean squared prediction error over the class of linear unbiased estimators). Sampling plans are selected so that the prediction error at any point on the map that has not been directly observed is less than some specified amount. The sampling intensity and the position of sample points are critical to obtaining good predictions and for constructing maps from sample data. Such problems arise in many areas of application including mining, hydrology and precision agriculture and provide the context for the development of geostatistics (Webster and Oliver, 2001).

The selection of a sampling plan that will meet these objectives is based on the application of the theory of kriging (optimal spatial prediction). The methodology depends on being able to specify a model (superpopulation) for spatial variation in the attribute to be mapped. This then allows the identification of optimal weights for local prediction, even if for the application it is difficult to defend the idea of multiple realizations. The predictor is the expected value of the attribute Z at a given site, o, given the available data $z(1), \ldots, z(n)$: $E[Z(o) \mid z(1), \ldots, z(n)]$. The theory of kriging will be discussed in more detail in chapter 4 so the reader may wish to return to the following description at a later stage.

Kriging theory shows that the largest prediction errors will usually be found along the boundary of the study region (if sampling is sparse there) and at the centres of the largest gaps between the sampled points in the interior of the

study region. A systematic sampling plan (with sample sites coming close to the boundary) will be better than an irregular sampling plan, particularly one that creates clusters of sampled sites. In the case of a rectangular grid of sample points, the maximum prediction error will be in the centre of any rectangle of sampled points. For a given sampling density, triangular grids give lower prediction errors but for practical reasons rectangular grids are preferred and of these centric aligned systematic grids are best (Burgess and Webster, 1980; Webster and Burgess, 1981). Burgess, Webster and McBratney (1981) show how a plot of the maximum value of the minimized prediction error, taken from the site at the centre of any rectangle of sampled points, plotted against sample size can be used to select sample size to achieve a required level of prediction error.

A map can be constructed to show how prediction errors vary over the sampled area. This information can then be used to suggest where new sampling points might be added. Conversely by experimentation, deleting sampling points or small subsets of sampling points, it is possible to examine the impact of such a deletion process on prediction error across the map (see section 3.2.1 for examples).

These predictors and estimates of prediction error are based on the superpopulation or model-based strategy. The analyst has to decide if the idea of there being 'other realizations' (at other points in time or other locations in space) has any meaning because if it does not then these prediction errors for a predictor of the expected value of a population of possible values of $Z(o)$ may be too large. Prediction error in this case should be with reference to $z(o)$ (since no other values are possible) and based on the idea of re-sampling the single data set with a similar sampling plan.

In the case of an environmental pollutant the superpopulation view is usually reasonable because a single source of pollution might be responsible for a number of polluted sites or it might be reasonable to consider the same process being replicated through time at the same place. Even in a geological context, if the study area is large, there may be several subregions with similar statistical properties so that prediction errors can refer to this collection of areas (Isaaks and Srivastava, 1989, p. 507).

(d) Sampling to identify extreme values or detect rare events

The researcher has reason to believe that an area contains extreme values of some attribute, that is values above some critical value (z_{crit}). The attribute of interest might be soil contamination arising from industrial processes and the critical value has been set by government. The aim is to identify where such sites are located and the total area above some critical value. If an exhaustive data set has been acquired by small quadrats that partition the area,

so there is no sampling problem, the next step is to map the indicator function ($I(.)$):

$$I(z(k)) = 1 \quad \text{if } z(k) > z_{\text{crit}}$$
$$= 0 \quad \text{if } z(k) \leq z_{\text{crit}}$$

Quadrats (k), above z_{crit} can be coloured black, those less than or equal to z_{crit}, white. The total area above the critical threshold can then be calculated by summing the areas of the quadrats coloured black. Such information may be useful in assessing the costs of any clean-up operation or deciding whether the area is safe on health grounds for a particular change of use. Critical values may also be defined in terms of the distribution of observed attribute values. The indicator function might be used to identify all areas that have an attribute value more than (say) three standard deviations above the mean, or in the top 10% of values.

When no data are available a sampling strategy is needed. If a sample has been taken the problem may be to convert the sample data into useful information on the two questions of 'where' and 'how much'. We consider first the situation where there are no data.

Selecting regions to sample If no sample data are available then the problem is to devise a sampling plan. Neither random, stratified random nor systematic sampling, described in section 3.2.3, are generally used in these circumstances. They tend to discard too much information the analyst may have about where the critical values are likely to be found. 'Judgement' sampling, in which the sampler draws on experience and knowledge about where the most contaminated sites are likely to be found, is more commonly used (Brus and de Gruijter, 1997). This question of where to look (and when to look elsewhere) also arises in archaeological searches for rare artefacts (Switzer, 2000).

Even using judgement sampling the sampler may be confronted by a number of possible areas that experience suggests could have rare artefacts or levels of an attribute above z_{crit}. Suppose for simplicity there are two areas where there might be high levels of a contaminate. The analysts prior belief is that area 1 is more likely to be where the high levels are to be found rather than area 2 so that area 1 is the best place to start looking. This prior belief may be based on the types of industrial activities that went on in the two areas. So, judgement suggests to start sampling area 1 first, but if sampling fails to throw up sample values above z_{crit} is there a point (after taking a certain number of samples) at which sampling should switch to area 2 and when is that point reached?

This is a problem which can be formalized in Bayesian terms as a process in which prior beliefs are progressively modified in the light of new data. At some

point the posterior distribution (the prior distribution modified by the data) may be such that when combined with a switching rule leads the sampler to move areas in the search for high levels of the attribute.

Area 1, the area where the sampler believes extreme values are most likely to be found, is subdivided into n quadrats and $\theta(1)$ denotes the probability of finding an extreme value in any given quadrat. This probability is based on the samplers judgement in the light of what is known about the areas history – perhaps in relation to experience from elsewhere. For convenience in this example, assume that the prior distribution for $\theta(1)$ is the beta distribution with parameters $a(1)$ and $b(1)$. That is:

$$\theta(1) \sim \text{beta}\,(a(1), b(1)) \tag{3.9}$$

where $\sim$ denotes 'has the probability density'. The beta distribution parameters imply that the expected value and variance of $\theta(1)$ is (Gelman et al., 1995, pp. 476–7):

$$E\,[\theta(1)] = a(1)/(a(1) + b(1)) \tag{3.10}$$

$$\text{Var}[\theta(1)] = a(1)b(1)/\{[a(1) + b(1)]^2[a(1) + b(1) + 1)]\} \tag{3.11}$$

Suppose the same prior density is specified for $\theta(2)$, for area 2, but with beta distribution parameters, $a(2)$ and $b(2)$. Since the sampler has chosen to start in area 1 rather than 2, these parameters are assumed to satisfy the relationship:

$$a(1)/(a(1) + b(1)) > a(2)/(a(2) + b(2)) \tag{3.12}$$

The sampler draws n samples from area 1 of which y have values that exceed z_{crit}. We assume that the likelihood for this sampling process is the independent binomial model for the number of areas with an attribute value greater than z_{crit}. So:

$$\{y \mid \theta(1)\} \sim \text{binomial}(n, \theta(1)) \tag{3.13}$$

The assumption of independence derives from the sampling process adopted by the sampler. The fact that adjacent quadrats are likely to be spatially correlated if quadrat size is less than the scale of pollution patches is not important here. (If the sampler were to adopt a sampling strategy in which the choice of the kth site to sample was dependent on the location and attribute values in $r (r \geq 1)$ preceding samples, the assumption of independence would be invalid.)

The posterior density for $\theta(1)$ given the data (y extreme values from a sample of size n) can de derived using Bayes rule:

$$\{\theta(1) \mid y\} \sim \text{beta}(a(1) + y, b(1) + n - y) \tag{3.14}$$

(Gelman et al., 1995, pp. 35–7). The posterior mean of (3.14) is:

$$E[\theta(1) \mid y] = (a(1) + y)/(a(1) + b(1) + n) \tag{3.15}$$

A possible sampling rule would be: if the first n samples yield no values above the critical level ($y = 0$) switch the search area from 1 to 2 when:

$$(a(1))/(a(1) + b(1) + n) < a(2)/(a(2) + b(2)) \tag{3.16}$$

(Switzer, 2000).

Consider another example. The Poisson model arises naturally in counting the number of cases of a rare disease. Suppose the problem is to detect clusters of cases of the rare disease in regions with poor medical records. We assume the same situation as in the previous example with the sampler starting to take samples in the area where cases are expected (area 1) but needing a decision rule to help decide when to look elsewhere. Let $\phi(1)$ denote the disease rate in area 1 and on the basis of previous experience or other information about the area, the prior distribution is assumed to be gamma distributed with parameters $\alpha(1)$ and $\beta(1)$:

$$\phi(1) \sim \text{gamma}(\alpha(1), \beta(1))$$

So, the expected value of $\phi(1)$ is (Gelman et al., 1995, pp. 44–5):

$$E[\phi(1)] = \alpha(1)/\beta(1) \tag{3.17}$$

The probability of finding y cases in area 1 with a population at risk of $n(1)$ when the underlying rate is $\phi(1)$ is Poisson distributed:

$$\{y \mid \phi(1)\} \sim \text{Poisson}(\phi(1)n(1)) \tag{3.18}$$

The posterior density for $\phi(1)$ given new data (y cases are found in a sample of n individuals) is:

$$\{\phi(1) \mid y\} \sim \text{gamma}(\alpha(1) + y, \beta(1) + n) \tag{3.19}$$

(Gelman et al., 1995, pp. 48–9). The posterior mean of (3.19) is:

$$E[\phi(1) \mid y] = (\alpha(1) + y)/(\beta(1) + n) \tag{3.20}$$

A possible sampling rule would be: if the first n individuals yield no cases of the disease ($y = 0$) switch the search area from 1 to 2 if and when:

$$(\alpha(1))/(\beta(1) + n) < \alpha(2)/\beta(2) \tag{3.21}$$

Mapping areas with extreme values We now turn to the case of a continuous surface of values, where sample data have been collected according to a sampling plan. The analyst now wishes to draw a map showing where the attribute value exceeds a critical value and compute areas. The approach to be discussed assumes a superpopulation or model-based view of the data.

Kriging (section 3.2.4(c) and 4.4.2) provides a means of predicting values on a surface given sample data. Confidence intervals can be attached to each prediction using the prediction standard errors. By supplementing sample data with predicted values particularly in areas where sampling is sparse, the surface can be quantified at various sites and contour lines drawn to give a continuous representation of the surface (Ripley, 1981, pp. 75–7). The extent to which the resultant maps show spatial detail will depend on sample intensity as well as sample positions and detail is lost as sampling becomes less intensive.

But for identifying areas with extreme values there are two problems with adopting the methodology of section 3.2.4(c). First, prediction standard errors derived from sampling cannot be used to assess the variability of non-linear functions of the predicted surface such as line lengths or areas (Ripley, 1981, p. 64; Chilès and Delfiner, 1999, pp. 449–51). Second, describing a distribution using the results of a limited sample is not as informative as examining model properties and in particular looking at the distribution that derives from the model (Switzer, 2000, p. 629). Kriging predictors are *expected values* of the random variable $Z(o)$ given the sample data. Kriging smooths the variation in the map as a whole and this smoothing is not uniform across the map. There is overestimation of small values and underestimation of large. This is an undesirable property if interest focuses on the variability in the surface or if the analyst is particularly interested in identifying areas with extreme values. In order to get a better picture of the distribution it is necessary to examine not just the central tendency of the distribution but also its variability and especially its tail properties. *Conditional* simulation can be used to achieve this. Conditional simulation produces representations consistent with known data values and 'aims to retain the overall texture of the variation of the statistics of the original data in the simulated values' (Frogbrook and Oliver, 2000, p. 226). The next section describes unconditional and conditional simulation.

Once multiple conditional realizations of a probability model have been obtained, there are several options for presenting results. Sample simulations can be shown (e.g., Bloom and Kentwell, 1998; Frogbrook and Oliver, 2000). To provide a summary, areas of the map that exceed the threshold for an extreme value in more than say 90% or 95% of simulations could be highlighted or maps that show the percentage of simulations where a specified extreme value

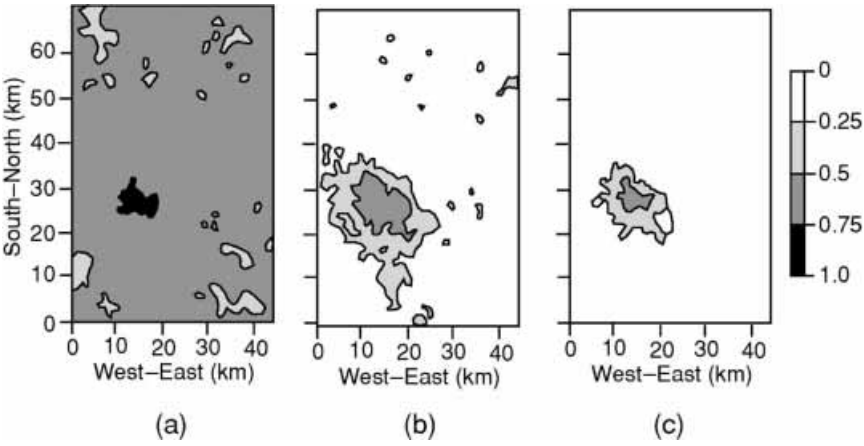


Figure 3.5 Gaussian simulations of Strontium 90: maps of probability of exceeding (a) 0.2, (b) 0.5, (c) 0.7 Ci/km (Savelieva et al., 1998, p. 463)

is exceeded can also be generated (e.g., Savelieva et al., 1998). Figure 3.5 shows simulated maps of the probability of extreme levels of Strontium 90 contamination arising from the Chernobyl fallout. Showing maps of the average across the set of independent simulations is not recommended as it will be similar to the map that would be obtained by kriging and so for the purposes described here will be too smooth. The variance of these simulated maps will tend to the kriging variance.

3.3 Maps through simulation

If a probability model can be specified for data then a better understanding of spatial variability can be obtained by generating multiple realizations from the model. *Multiple* realizations are needed because any simulation is but one of a large number of representations of the specified probability model. The following procedures can be used to generate multiple realizations of a normal random field with a specified mean and covariance or semi-variogram structure. The discussion is in two parts: first *unconditional* and then *conditional* simulation. Because this section uses results that will not be discussed until later in the text (chapters 4 and 9) the reader may prefer to return to this later.

The aim of *unconditional* simulation is to obtain $\iota = 1, \ldots, M$ realizations of a spatial model at n locations where the attribute Z is a normal random variable with a specified mean $\boldsymbol{\mu}$ (of length n) and specified $n \times n$ variance–covariance matrix $\boldsymbol{\Sigma}$. Valid models for spatial data will be discussed in chapter 9.

Unconditional simulation yields realizations that are consistent with the probability model but simulated data values are not required to correspond with known data values.

A valid simulation procedure is as follows. First, the Cholesky decomposition of the $n \times n$ matrix $\boldsymbol{\Sigma}$ is obtained. The Cholesky decomposition states there is a lower triangular n by n matrix $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma}$. This decomposition should be obtained analytically if possible, otherwise numerically. A numerically efficient method is to decompose $\boldsymbol{\Sigma}$ into the matrix product $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$ where the columns of $\mathbf{Q}$ are the n eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$ is a n by n diagonal matrix of corresponding eigenvalues. It follows that $\mathbf{L} = \mathbf{Q}\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^T$. Note that although this aspect of the simulation is a 'one-off' operation it can be a computationally demanding operation if n is large (>1000). If n is large it may be necessary to partition the area into large overlapping neighbourhoods and keep watch for spurious discontinuities on the final simulated surface (Chilès and Delfiner, 1999, p. 468).

Next, for each of the M simulations that are required, obtain a vector of length n of uncorrelated normal random variables with mean zero and unit variance, $\boldsymbol{\xi}(\iota)\,(\iota = 1, \ldots, M)$. Third, compute $\mathbf{L}\boldsymbol{\xi}(\iota)\,(\iota = 1, \ldots, M)$ to obtain the M realizations of the variance–covariance part of the field. Each vector can now be added to the mean vector to give M *unconditional* realizations of the field:

$$\mathbf{z}^*(\iota) = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\xi}(\iota) \quad \iota = 1, \ldots, M \tag{3.22}$$

In the case of *conditional* simulation, realizations are from a normal probability model where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are specified but at the locations where data values are known the realization matches these. This is achieved by transforming an unconditional simulation of the probability model. The principle underlying conditional simulation is due to Matheron (1976) and is as follows. Let $z(\mathbf{s})$ denote the true but unknown value at location $\mathbf{s}$. Let $\hat{z}(\mathbf{s})$ denote the kriging predictor of $z(\mathbf{s})$ based on data at a set of sample points (see (4.37), where $\mathbf{y}$ is used instead of $\mathbf{z}$). Now, clearly:

$$z(\mathbf{s}) = \hat{z}(\mathbf{s}) + [z(\mathbf{s}) - \hat{z}(\mathbf{s})] \tag{3.23}$$

The second term on the right-hand side of (3.23) is the kriging error however this is also unknown. An estimator of the kriging error is required. This is obtained as follows. The same expression as (3.23) for the output of an unconditional simulation is:

$$z^*(\mathbf{s}) = \hat{z}^*(\mathbf{s}) + [z^*(\mathbf{s}) - \hat{z}^*(\mathbf{s})] \tag{3.24}$$

where $z^*(\mathbf{s})$ is the unconditional simulation value at $\mathbf{s}$ (3.22). Now $\hat{z}^*(\mathbf{s})$ is the kriging predictor obtained using the data taken from the unconditional

simulation at the same locations used to calculate $\hat{z}(\mathbf{s})$. Now if $z^+(\mathbf{s})$ denotes the conditional simulation value at $\mathbf{s}$, then:

$$z^+(\mathbf{s}) = \hat{z}(\mathbf{s}) + [z^*(\mathbf{s}) - \hat{z}^*(\mathbf{s})] \tag{3.25}$$

The data value at $\mathbf{s}$ is honoured because kriging is an exact interpolator. There is further discussion together with a demonstration that (3.25) yields map output with the required properties providing there is no systematic measurement error (normal $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ model; $z^+(\mathbf{s}) = z(\mathbf{s})$ at those $\mathbf{s}$ where there are sample data values) in Chilès and Delfiner (1999, pp. 465–8).

The argument can be summarized as follows. If $z^+(\mathbf{s}; \iota)$ denotes the realized value at location $\mathbf{s}$ in conditional simulation ι then (Cressie, 1991, p. 208):

$$\begin{aligned} z^+(\mathbf{s}; \iota) &= \hat{z}(\mathbf{s}) + (z^*(\mathbf{s}; \iota) - \hat{z}^*(\mathbf{s}; \iota)) \\ &= z^*(\mathbf{s}; \iota) + \mathbf{c}^T \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \mathbf{z}^*(\iota)) \end{aligned} \tag{3.26}$$

where $z^*(\mathbf{s}; \iota)$ is the realized value at location $\mathbf{s}$ in simulation ι from an unconditional simulation (3.22). $\hat{z}^*(\mathbf{s}; \iota)$ is the simple kriging predictor as defined by (4.37) except the data vector $\mathbf{z}$ ($\mathbf{y}$ in (4.37)) is replaced with $\mathbf{z}^*(\iota)$ given by (3.22). The unconditional simulations are only involved through the estimation of the kriging errors (the second term on the right-hand side of (3.26)). Chilès and Delfiner (1999, p. 468) say, of conditional simulation, that it '"vibrates" in between the datapoints within an envelope defined by the kriging standard error'. Maps obtained by conditional simulation are useful qualitatively because they provide realistic pictures of the spatial variability based on the evidence in the data; they are useful quantitatively because they allow the analyst to assess the impact of spatial uncertainty on outcomes (Chilès and Delfiner, 1999, p. 453). The methodology can be extended using co-kriging to the multivariate case (see, e.g., Savelieva et al., 1998).

There is discussion of simulation methods in Ripley (1981, pp. 16–18, 64–72), Cross and Jain (1983), Haining et al. (1983), Cressie (1991, pp. 200–9), Goovaerts (1997) and Chilès and Delfiner (1999, chapter 7). Simulation methods, particularly conditional simulation methods, along with interpolation methods (see chapter 4) are used for downscaling data, that is transfering data from larger to smaller scales (Bierkens et al., 2000, pp. 111–44). Switzer (2000) describes a method for efficiently sampling from the model distribution rather than adopting unrestricted random sampling as a result of which many realizations might be generated that are similar to one another whilst leaving unsampled other areas of the space of realizations.

4

Data quality: implications for spatial data analysis

This chapter is concerned with examining the implications of different aspects of data quality for the conduct of spatial data analysis. It was noted at the end of chapter 2 how particular aspects of data quality may have an impact on particular stages of spatial data analysis. Whilst some quality issues impact on the data collection and data preparation stages prior to undertaking analysis, other quality issues impact more on the form and conduct of the statistical analysis or on how results can be interpreted.

The first section deals with error models and the implications of different types of error for data analysis. Section 4.2 considers various problems associated with the spatial resolution of data. The problems discussed include: the impact of varying levels of precision across a map divided into areas; the change of support problem (moving from one spatial framework to another); the problems associated with ecological analyses including aggregation bias and the modifiable areal units problem. Sections 4.3 and 4.4 deal with consistency and completeness problems which include the missing data problem. Some of the results in these later sections use data models which are discussed in chapter 9.

4.1 Errors in data and spatial data analysis

4.1.1 Models for measurement error

All data contain error as a consequence of the inaccuracies inherent in the process of taking measurements. Error models are important in data analysis. An error model allows quantification of the probability that the true value lies within a given range of the measured value. Valid error models allow exploration of the effects of error propagation where arithmetic or other

operations are performed on one or more variables that individually contain error. Specification of an appropriate model for the errors is an important element of regression modelling.

(a) Independent error models

The independent and identically distributed (iid) normal model, $N(\mu, \sigma^2)$, where μ denotes the mean and σ^2 the variance of the distribution, is widely used as a model for measurement error. This is because in many situations the errors that occur are a compounding of many small independent sources of random error (Mikhail, 1976). Suppose a quantity is measured that has a true value X. Suppose also that each source of error deflects the measurement up or down by a quantity ε and that these two possibilities occur with equal probability. If there are n sources of error then the measured value could range from $(X + n\varepsilon)$ to $(X - n\varepsilon)$. If v of the n sources give positive errors and $(n-v)$ give negative errors then the measured value will be: $(X+(2v-n)\varepsilon)$. Let the probability of v positive errors out of n sources be given by the binomial probability distribution with parameters n and $p = 1/2$. As n increases and $\varepsilon \to 0$ this distribution converges to a normal distribution about the true value X. Providing $\varepsilon \downarrow 0$ and $n \uparrow \infty$ such that $(\varepsilon\sqrt{n})$ remains fixed, then the standard deviation of the normal distribution is given by $\sigma = \varepsilon\sqrt{n}$ (Taylor, 1982, pp. 197–9).

The mean (μ) represents any systematic bias in the measurement process and the variance (σ^2) measures the dispersion around the mean. The square of the RMSE provides an estimate of σ^2. The model can be used to compute the probability that the error will lie between two values (a, b) of X.

The $N(\mu, \sigma^2)$ model may be appropriate for the errors associated with the measurement of a point location in one dimension. For a point location on a two-dimensional surface the bivariate normal density will allow for error in both dimensions.

The independent normal model may be plausible for the errors associated with the individual sample points along a boundary but not for the errors associated with the measurement of the boundary itself. Nor is it likely to be plausible for the attributes of the area feature (such as its size or boundary length) that are derived from the representation. This is because these errors also depend on the positioning and density of sample points in relation to the shape and overall complexity of the boundary.

The assumption of normality may be reasonable for errors in measuring attribute values although in the case of synthetic aperture radar data, theoretical arguments have suggested a Rayleigh distribution (Besag, 1986). The independence assumption may also be justifiable for attribute measurement error.

However, as illustrated in section 2.4, there are circumstances where models are needed that allow for spatial dependence amongst the errors – particularly for errors at neighbouring locations.

(b) Spatially correlated error models

A general model for dependent normal measurement error on a variable X across a set of n locations (points or areas) is provided by the multivariate normal distribution, MVN($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) $\boldsymbol{\mu}^T = (\mu(1), \ldots, \mu(n))$ is the vector of means representing the bias in the measurements at each location and:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \sigma_{n,3} & \cdots & \sigma_n^2 \end{bmatrix}$$

The term σ_1^2 is the variance for the measurement error at i and $\sigma_{i,j}$ is the covariance in the errors between locations i and j.

When there is spatial correlation in attribute errors it may be reasonable to assume that it is local in scale (see 2.3.2). So, for any location i let:

$$\sigma_{i,j} \begin{cases} \neq 0 & \text{if } j \in N(i) \\ = 0 & \text{if } j \notin N(i) \end{cases}$$

where $N(i)$ denotes a low-order neighbour of location i, that is area i and j share a common boundary or are perhaps at most one step removed (see section 2.4). There are a number of possible models that meet this requirement whilst also satisfying the condition of positive definiteness for $\boldsymbol{\Sigma}$ which is a requirement for a valid covariance function (Morrison, 1967, p. 60).

A possible model with a local covariance structure is given by the moving average model (Haining, 1978) where:

$$u(i) = \nu \Sigma_{j \in N(i)} w(i, j)\, e(j) + e(i) \tag{4.1}$$

where ν is a constant, $\{e(i)\}$ denote independent drawings from a $N(0, \sigma^2)$ distribution. It follows that the variance of $u(i)$ is:

$$\text{Var}(u(i)) = \sigma_i^2 = \left(1 + \nu^2 \Sigma_{j \in N(i)}\, w(i, j)^2\right)\sigma^2 \tag{4.2}$$

The covariance between $u(i)$ and $u(j)$ where $j \in N(i)$ is given by:

$$\begin{aligned} \text{Cov}(u(i), u(j)) = \sigma_{i,j} = &\left(\nu(w(i, j) + w(j, i))\right. \\ &\left. + \nu^2 \Sigma_{k \in \Im(N(i), N(j))} w(i, k) w(j, k)\right)\sigma^2 \end{aligned} \tag{4.3}$$

where $\Im(N(i), N(j))$ denotes the set of sites that are neighbours of both i and j. The covariance between $u(i)$ and $u(k)$ where $j \in N(i)$, $k \in N(j)$, but $k \notin N(i)$ is given by:

$$\text{Cov}(u(i), u(k)) = \sigma_{i,k} = \left(\nu^2 \Sigma_{j \in N(i)} w(i, j)\, w(k, j)\right) \sigma^2 \tag{4.4}$$

All other covariances are zero. Cliff and Ord (1981, p. 150) define a moving average model where covariances are zero after the border adjacency, as do Kiefer and Wynn (1981). These are not the only local error models. Ripley (1981, p. 55) describes a process due to Zubrzycki that can generate a covariance function for a spatial surface that decays steeply to 0.

Correct model specification includes specifying a valid model for the errors. To illustrate, take the case of a simple bivariate regression model where the response variable is Y and the explanatory variable is X and:

$$Y(i) = \beta_0 + \beta_1 X(i) + e(i)$$

where β_0 and β_1 are the intercept and slope or regression parameters respectively. The model is to be fit to data values $\{y(i), x(i)\}$ associated with n spatial objects. The term $e(i)$ is the error term which in a correctly specified model (in terms of Y and X) and with no error in measuring values of X would account for measurement error on Y. Ordinary least squares fitting assumes the $\{e(i)\}$ are independent and identically distributed with a mean of 0 and variance σ^2. However there will be situations arising in spatial data analysis where this assumption may not be supported by the data and a more general model needed that allows for spatial dependence amongst the errors. This has implications for model fitting.

4.1.2 Gross errors

(a) Distributional outliers

Data values which are extreme with respect to the overall distribution of values, may not be wrong but amongst such *distributional outliers* is one area to start looking to detect gross errors. The detection of extreme values on a variable is important because their presence is likely to give rise to a number of consequences. Particularly in small samples, non-resistant descriptive statistics like the mean or standard deviation are affected by the presence of such data values. Where the presence of such gross error is suspected, or if the distribution of values is known to contain some extreme but valid data values, statistics like the median and the inter-quartile range (the difference between the upper and lower quartile) provide descriptions of the centre and spread of a set of data values that are resistant to their presence. In the case of

regression modelling, extreme data values in the X or explanatory variable give rise to large leverage effects whilst in the Y or response variable give rise to large residuals or outliers (see figure 4.1). The presence of large leverages and/or outliers can have a disproportionate influence on regression parameter estimates and model predictions. Regression diagnostics are provided in statistical packages like MINITAB and SPSS for example to assess the influence of individual observations by deleting cases (Belsley, Kuh and Welsch, 1980).

Methods for identifying outliers in a data set usually draw on a combination of graphical and numerical techniques. Some extreme forms of location error can be detected just by overlaying the data cases on a map of the study area and using local knowledge. Location error may also come to light by examining attribute values by area. Extreme attribute values may be caused by location error in the data.

Attribute error in the set of values on a variable Z may be suspected from a histogram plot in which data intervals on the horizontal axis are specified in terms of fractions of standard deviations. Cases lying more than three standard deviations above or below the mean may be considered extreme and flagged for closer investigation. The above method using the mean and standard deviation is equivalent to basing the identification of an outlier on the mean shift outlier model (Weisberg, 1985, p. 114). The model specification is:

$$z(i) = \beta_0 + \delta u(i) + e(i) \tag{4.5}$$

where β_0 and δ are parameters, the $\{e(i)\}$ are independent and identically distributed errors with a mean of 0 and variance σ^2 and $u(i) = 0, (i \neq j)$ and $u(j) = 1$. The jth case is under suspicion and is deemed an extreme value if $\delta \neq 0$ because its mean value is $\beta_0 + \delta$ whilst for all other values $(i \neq j)$ the mean is β_0. Gross measurement error on the jth case may be responsible for $\delta \neq 0$. The above test is equivalent to regressing Z on U and testing $\delta = 0$ against the two-sided alternative $(\delta \neq 0)$. The t-test with $n - 2$ degrees of freedom can be used if the $\{e(i)\}$ are iid $N(0, \sigma^2)$.

The D statistic is a resistant technique for checking for the presence of extreme values. It is defined:

$$D = n^{1/2}(\bar{z} - \text{med}(z))/(0.7555\hat{\sigma}) \tag{4.6}$$

where $\hat{\sigma} = (U - L)/1.349$ and U and L are the upper and lower quartiles respectively. $(U - L)$ is called the inter-quartile range, med(z) is the median value, $\bar{z}$ is the mean and n is the number of cases. This criterion can be used on any symmetric distribution of values (e.g. a normal distribution) to flag the presence of atypical values. Their presence is suspected if $|D| > 3.0$. A resistant approach to isolating particular cases is associated with the boxplot. An extreme value

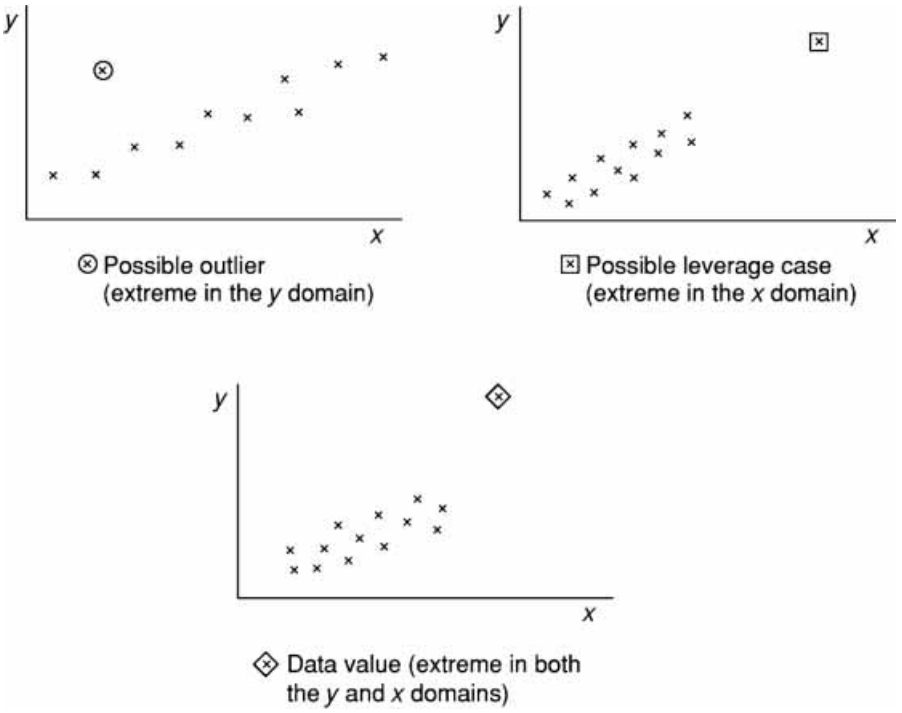


Figure 4.1 Graphical identification of cases with extreme values

criterion for any individual value, $z(i)$, is given by:

$$z(i) > U + \phi(U - L) \quad \text{or} \quad z(i) < L - \phi(U - L) \tag{4.7}$$

where $(U - L)$ denotes the inter-quartile range and ϕ is usually set to 1.5. Since these quartile statistics are resistant measures they will not be much affected if there are some extreme values, unlike the mean shift outlier model.

An error may only be apparent when examined in the context of other variables. For example an error may be made in recording a disease rate for an area, $z(i)$, that is not apparent from looking at the distribution of $\{z(i)\}$. The given value is not necessarily a distributional outlier. In figure 4.1 the outlier is not a distributional outlier on the set of y values. An extreme value may only be suspected when it is analysed in the context of another characteristic, X, for example deprivation. The evidence for a possible error may only be apparent when a scatterplot of $\{z(i)\}$ against $\{x(i)\}$ is constructed or after fitting a regression model of disease rate on deprivation and examining outlier diagnostics (Weisberg, 1985). The test is again based on Weisberg's mean shift outlier model (4.5), although Z is now regressed on both U and X:

$$z(i) = \beta_0 + \delta u(i) + \beta x(i) + e(i) \tag{4.8}$$

The jth case is an extreme value if $\delta \neq 0$ because the mean of $z(j)$ is $\beta_0 + \delta + \beta x(j)$. This test can be generalized to more than one explanatory variable.

(b) Spatial outliers

Attribute values may be extreme given their position on the map. Such attribute values are termed 'spatial outliers' because their values are extreme relative to the set of neighbouring values on the map. It is possible for a data value to be a spatial outlier without being extreme in the distributional sense which is why methods other than those described above are needed. A map of the data may help flag up possible spatial outliers. Further visual evidence can be obtained from a scatterplot of $\{z(i)\}$ on the vertical axis against their corresponding values in $\{\mathbf{W}^*\mathbf{Z}(i)\}$ on the horizontal axis where the row sums of $\mathbf{W}^*$ equal 1 and $w(i, i) = 0$. $\mathbf{W}^*\mathbf{Z}(i)$ is the value in the ith entry of the product of the matrix $\mathbf{W}^*$ with the vector of values on Z. So, this plot displays the attribute value against the average of its neighbours.

The plot may highlight individual cases if they are sufficiently different from their neighbours, however it would be useful to have criteria for deciding when to flag particular cases for further investigation. If the data are on a square lattice, Cressie (1984, 1991, pp. 38–40) suggests computing the D statistic (4.6) for each row and column. These tests can be adapted to non-lattice data by assigning data values to row and column *bands*. Cressie (1991, pp. 396–8) assigned the 100 counties of North Carolina to the nodes of a 9 by 24 square lattice. In non-lattice cases, however, results could be sensitive to the chosen allocation procedure.

Fitting the model:

$$z(i) = \delta u(i) + \beta_1 \mathbf{W}^*\mathbf{Z}(i) + e(i) \tag{4.9}$$

where $u(i) = 0\,(i \neq j), u(j) = 1$ offers a method for testing whether the jth value is extreme given its neighbouring values. The test can be used if there is no spatial trend or after removing the trend. The method is based on Weisberg's (1985, pp. 115–16) mean shift outlier test described above (4.8).

Figure 4.2 shows a plot of $z(i)$ against $\mathbf{W}^*\mathbf{Z}(i)$ where $z(i)$ is the standardized mortality rate for accidents in Glasgow in community medicine area i (Haining, 1990, pp. 199–200). The circled case is an outlier at the 1% level based on fitting the model:

$$z(i) = \beta_0 + \beta_1 \mathbf{W}^*\mathbf{Z}(i) + e(i) \tag{4.10}$$

by ordinary least squares and flagging cases with a standardized residual exceeding |3.0|. This case is probably not an error. It refers to the city centre and the extreme value (relative to the average of the neighbouring rates) is

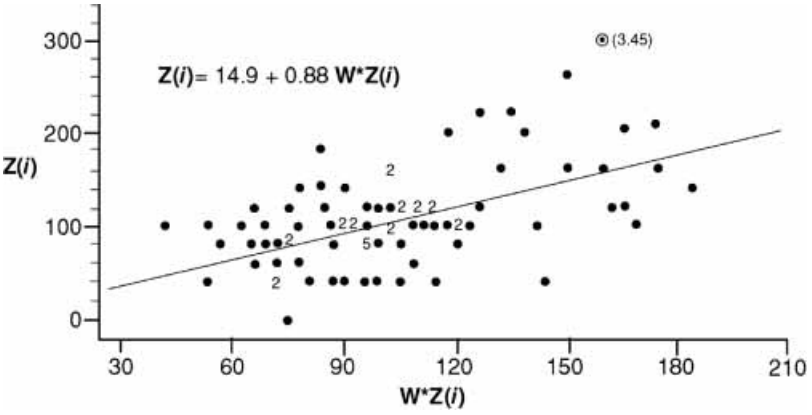


Figure 4.2 Model-based identification of a spatial outlier

probably a consequence of using population as the choice of denominator. Such areas have low resident populations but high daytime and early evening populations and high levels of traffic.

It should be noted though that ordinary least squares fitting of either (4.9) or (4.10) is likely to overestimate the number and seriousness of the outlier problem. This is because ordinary least squares fitting in the case where $\beta_1 \neq 0$ underestimates the residual standard error. However the purpose here is simply to flag cases for closer investigation.

(c) Testing for outliers in large data sets

The previous tests seem reasonable if the analyst is using them to check a particular value that has been suspected in advance of reviewing the data. In the case of very large data sets it may be important to draw on automated methods that scan through the entire data set. But automated methods that pass through the entire data set in search of possible errors are performing n (the number of cases) significance tests. The probability of finding at least one outlier as a result of performing n independent significance tests approaches 1 as n increases. Any testing procedure therefore should be modified to reflect this multiple testing. Weisberg (1985, p. 116) proposes resetting the critical value based on the Bonferroni inequality. For an overall α significance level of 0.05, a level of $0.05/n$ is chosen for each test. Choosing the $(\alpha/n) \times 100\%$ point of the t-distribution for each of the n tests gives an overall significance level of α. However, if the data set is very large dividing by n will probably result in a very conservative test (underestimating the number of possible extreme values). The application of this adjustment in the case of the model defined by (4.9) raises further problems however because the n tests are not independent. Not only do the n tests use overlapping subsets of the data but if $\beta_1 \neq 0$ then the

underlying data values are spatially correlated. If the test is not to be conservative even for moderate sized data sets n needs to be reduced to reflect the 'effective' or 'equivalent' number of independent tests. In the light of these problems a method based on simply ranking the extreme values and identifying breaks in the distribution would seem easiest and most practical.

The methods described above are appropriate for detecting single extreme values but when there are many it is possible that they could mask one another, making their detection difficult. Spatial clusters of errors can arise too, for example through error propagation effects as described in section 4.1.3. One approach to the detection of multiple extreme values is to develop methods that search all subsets of cases for outlying subsets (Hawkins et al., 1984). These methods are not discussed further except to note that techniques based on the principal of analysing all possible *spatial* subsets have been developed to test for spatial clusters of disease and these will be considered in chapter 7.

It is important to emphasize that extreme data values are not evidence in themselves of gross errors and in some geographical problems extreme values arise because of the top heavy or 'primate' structure of the underlying system (Cox and Jones, 1981). Capital cities and economic enclaves in developing countries and core areas of developed economies may appear as outliers or leverage cases in plots and analyses of regional socio-economic data. In other circumstances finding extreme values is the objective of the analysis as in the case of finding disease or crime hot spots or areas with special environmental or geological characteristics not found elsewhere. Thus in some cases extreme values are providing evidence of what is sometimes termed 'spatial heterogeneity'. If, however, an extreme value is the result of error the data value may be correctable or it may have to be discarded. The latter course of action can lead to other problems associated with the consequences of data incompleteness (see section 4.4).

4.1.3 Error propagation

Error propagation results from carrying out arithmetic operations on interval and ratio data that contain error. Errors can also propagate as a result of performing logical operations – for example, identifying areas or point sites that simultaneously satisfy conditions associated with two or more variable values ($x(i)>x$.AND.$y(i)<y$); at least one of the variables satisfy a specified condition ($x(i)>x$.OR.$y(i)<y$) and so on (see Arbia et al., 1998, p. 150 for examples.) Error propagation in this case has implications for identifying regions on a map and for the statistical analysis of nominal- and ordinal-level data.

In analysing error propagation effects, the size and geography of the errors is important. Where errors possess spatial continuity, blend in with map structure, and appear visually plausible this has implications both for the detection of errors and for the interpretation of the results of statistical analysis. 'Small-scale, visually plausible patterns of error are perhaps more likely to escape detection' (Haining and Arbia, 1993, p. 294). Carter (1992) and Lee et al. (1992) describe how spatial correlation in elevation estimates combined with known levels of precision in digital elevation models impacts on slope estimates and drainage basin estimates.

The Geman and Geman (1984) 'corruption model' has been used to explore error propagation in raster data sets. A special case of this model assumes that n ground truth values $\{t(i)\}$ are corrupted as a result of location error and spatially correlated measurement error into a set of n observed values $\{z(i)\}$. In particular:

$$z(i) = \Sigma_{j \in N(i)}\, w^{+}(i, j)\, t(j) + u(i) \tag{4.11}$$

where again $N(i)$ denotes the set of pixels that are 'adjacent' to pixel i. For all i and j, $w^{+}(i, j) \geq 0$ and $\Sigma_{j \in N(i)}\, w^{+}(i, j) = 1.0$. Total error $(z(i) - t(i))$ can be partitioned:

$$z(i) - t(i) = (w^{+}(i, i) - 1)\, t(i) + \Sigma_{j \neq i}\, w^{+}(i, j)\, t(j) + u(i) \tag{4.12}$$

The first two terms on the right-hand side represent a form of location error arising from the fact that if $w^{+}(i, i) < 1.0$, what is recorded for pixel i and which is supposed to represent ground truth at the corresponding ith parcel of land in fact represents ground truth at an average of one or more nearby parcels of land. Attribute error includes a further component arising from reflectance picked up by the sensor which acts as a filter, observing ground truth for any parcel of land i as a weighted function of ground truth in the adjacent areas (Forster, 1980). The second term, $\{u(i)\}$, can represent this additional source of error and can be modelled as a sample from a MVN($\mathbf{0}$, Σ) distribution. Spatial correlation in the measurement process is specified using the matrix Σ. This component of error is treated as spatially localized so that only the off-diagonal values in Σ representing the near neighbours of each pixel i, are non-zero.

A model similar to (4.11) was analysed using Taylor series methods by Heuvelink et al. (1989) and Heuvelink (1993). In a series of papers, Haining and Arbia (1993), and Arbia, Griffith and Haining (1998, 1999) analyse and visualize in map form error propagation arising from various arithmetic and logical operations using Geman and Geman's model. Analytical as well as Monte Carlo simulation methods are employed. The covariance structure in Σ was based on the filter identified by Forster (1980). Monte Carlo methods allow properties to

be identified by examining multiple simulated realizations of a process. Properties of the propagated error are given under different map operations as well as the contribution of different sources of error to both aspatial and spatial error properties. The effect of location error is least when there are high levels of spatial correlation in ground truth values (e.g. the landscape is broadly uniform in character). It increases as ground truth shows less and less spatial correlation relative to the scale of the location error (e.g. a highly mountainous area). The formation of 'error regions' – that is contiguous pixels on the map with particularly high levels of error – is noted in the case of the ratio operation. This arises particularly where ground truth is spatially correlated, there is location error and there are high levels of spatial correlation in the attribute measurement process (as described by the matrix Σ). Similar but less-striking spatial structure seems to emerge in the case of the other arithmetic and logical operations.

Much of this work in the geographical and environmental sciences has been motivated by the need to understand the reliability of the output from geographic information systems which bring together data sets with different origins, generated at different scales often of varying quality and which are then subject to various map overlay operations (Unwin, 1995). One approach to this problem is to try to derive properties of the propagated error starting from model-based assumptions about the errors in the source maps, as in the examples above. An alternative approach is to perturb the final maps to see by how much they must be changed before associations and relationships identified from the observed map start to break down (Moran and Bui, 2000). If the perturbations need to be severe before any marked changes in conclusions are uncovered, then the analyst may feel confident in the results based on the observed map; where small perturbations of the data lead to marked changes, caution is called for. Even following this empirical approach some limits and some structure need to be imposed on the perturbation process so as to be consistent with what is known about the uncertainty in the data – and for this a model is useful.

The methodology of Monte Carlo simulation is widely used for analysing error propagation – see, for example, Goodchild et al. (1992), Veregin (1994, 1995). Models are often relevant to quite specific circumstances and assume spatial continuity and spatial homogeneity of error, whereas in reality error is likely to vary across the map. In some types of surveys there may be discontinuities arising from parcelling up an area into blocks and assigning different data collectors to different blocks (Milne, 1959). Goodchild (1995) expresses the view that 'as we learn more about the nature of errors, and why they occur, it will be possible to produce more refined models of error that take . . . spatial

heterogeneity into account' (p. 74). However, the generality of findings arising from this area of research has often been questioned as it is difficult to disentangle general findings from the specificity associated with particular case studies. At the present time the practical use of much of this work for spatial data analysts is probably limited to raising awareness of the way error can corrupt statistics. It helps to qualify the results of statistical analysis and in some cases helps to decide which of several competing statistics might be the more robust to likely errors in the data. This latter concern applies to maps of vegetation indices where there are several to choose from, including those based on image differencing, orthogonalizing transformations such as Gram-Schmidt (Kauth and Thomas, 1976) and principal components analysis (Richards, 1986). The paper by Arbia et al. (2003) offers some suggestions. Health (Jarman, 1993; Townsend et al., 1988) and mortality (Friedman, 1994) indices also manipulate spatial data values through arithmetic operations. Error propagation effects on such maps again need to be recognized.

4.2 Data resolution and spatial data analysis

For a given study area, the results of any analysis of aggregated spatial data will depend on the choice of scale and partition in the case of vector data, and the scale, orientation and origin of the grid in the case of raster data. Space has no natural origin or partitioning. Even re-orienting a grid or taking a new origin will result in a different aggregation, although where spatial correlation is strong in all directions in the variable of interest up to the scale of the pixel size neither of these should raise serious problems. Land-use categories that occupy small scattered areas may be lost if the scale of resolution is large so that their presence is underestimated, whilst that of land-use types appearing in large contiguous areas is overestimated. Arbia et al. (1996) analyse the effect of grid resolution on image misclassification and have shown how error increases as a consequence of moving from fine to coarse resolutions, although with an effect that is moderated by the level of spatial correlation in the underlying surface.

The scale of a grid in the case of remotely sensed data, or the size of the areal units in the case of recording population characteristics, could be altered. For a given study region partitioned into subareas, terms such as 'size of the data set' or 'number of cases' are scale-dependent quantities. As sample size increases as a result of using a finer and finer resolution, any null hypothesis is likely to be rejected if a sufficiently fine resolution is selected. The usefulness of undertaking classical hypothesis testing, particularly when data volumes are large so that any simple null hypothesis will almost certainly be rejected in favour of

a general alternative, continues to provoke debate that ranges from adjusting significance levels to reflect the volume of data (Leamer, 1978) to abandonment of classical hypothesis testing altogether (Nester, 1996).

Where the scale of variation of the phenomenon is greater than the resolution of the spatial unit through which its attributes are recorded then data values for adjacent spatial units will be spatially dependent. If there is a further reduction in the size of the areal unit leading to, say, a doubling of the number of 'cases' this is similar to augmenting the data file by simply duplicating some of the observations already in the file. The 'effective' sample size, in the sense of the amount of independent information carried in the data file about the process or the spatial surface, has not increased by the same amount as the increase in the number of cases. The 'effective' sample size for the purpose of statistical inference in the case of spatially dependent data is less than the number of cases because neighbouring data cases carry overlapping amounts of information (Clifford et al., 1985). The concept of the 'effective sample size' is one way to quantify the effect of spatial dependence in data values for the purpose of carrying out statistical testing (see section 3.2.4(a) and chapter 8).

4.2.1 Variable precision and tests of significance

The number of crimes committed or cases of a particular disease in an area are often viewed as a sample from a random process (see section 2.1.4). What is then computed for each area, such as the rate of the disease (number of cases divided by the population at risk) provides only an *estimate* of the true underlying risk. Suppose the underlying process is homogeneous, that is the same random process is operating across all members of the population at risk. Intuitively we would expect the precision of the estimator to be higher (the variance of the estimator smaller) the larger the number of individuals used in any aggregation and the more common the event that is being recorded.

These remarks lead to the conclusion that any map of rare disease rates, for example, needs to be interpreted cautiously if it is used to infer variation in risk across the map. The variance of the estimator for any area is greater the smaller the underlying population – the denominator used in the calculation of the rate. It follows that maps of rates where the areal framework has produced large differences in the denominator is likely to show evidence of rate variation that may be a statistical artefact of the spatial framework rather than any intrinsic variation across the region in the true underlying disease rate. This gives rise to two specific consequences (see, e.g., Mollie, 1996). First, most extreme rates are found in administrative units with small populations.

Second, most of the standardized rates, where the observed rate is significantly greater than the expected rate under the null hypothesis of a random allocation of cases across the map, are found in administrative units with large populations. This latter remark follows because as the denominator increases estimator variance decreases so that it only takes relatively small departures of the observed count from the expected count to generate a statistically significant difference. The methodology for constructing 'reliable' maps, will be discussed in more detail in chapter 7 as part of exploratory spatial data analysis and chapter 10 as part of univariate modelling.

Size–variance relationships arise in other contexts (Haining, 1990, p. 49). For example, suppose $Y(i)$ is any continuous valued variable that measures the average of $n(i)$ equally variable observations (with variance σ^2) in spatial unit i. It follows that the variance of $Y(i)$, $\text{Var}(Y(i)) = \sigma^2/n(i)$. If $Y(i)$ is the response variable in a regression model then this property would violate one of the assumptions of ordinary least squares regression that error variances must be constant (homoscedasticity). Violation of this assumption would be checked by constructing a scatterplot of the square of each residual obtained from the ordinary least squares fit of Y on the set of explanatory variables against the corresponding $1/n(i)$. If the scatterplot is upward sloping or 'wedge-shaped' to the right this provides evidence of a size–variance relationship (heteroscedasticity). The remedial action is to refit using weighted least squares estimation with weights given by $n(i)$, thereby downweighting the contribution of datapoints with large variances. Note that if $Y(i)$ is a total of $n(i)$ observations then $\text{Var}(Y(i)) = n(i)\sigma^2$ and a similar argument can be constructed for this situation (Weisberg, 1985, p. 83).

4.2.2 The change of support problem

This section looks at two different change of support problems, that is situations where data are collected with respect to one spatial framework but need changing or transforming to a new framework. In geostatistics a commonly encountered change of support problem arises when the analyst wants to make inferences about the arithmetic average of a variable for an area on the basis of point data, that is the transfer of data from smaller to larger scales or 'upscaling' (Bierkens et al., 2000). In the case of regional data, data are collected on one spatial framework but need transforming to another. In the geographical literature this is called the areal interpolation problem.

(a) Change of support in geostatistics

Suppose samples have been taken from a continuous surface at locations $s(1), \ldots, s(n)$. These locations are considered here as *points* or very small

areas. The n observed values on the measured variable Y, might be the yield of some mineral or a measure of soil contamination or air pollution. These data values are denoted $\mathbf{y} = (y(1), \ldots, y(n))$ and they are assumed to be realizations of a random process. The analyst needs to predict the average value of the variable Y for an area A, $\bar{Y}(A)$. In an agricultural context samples may refer to levels of a nutrient in a small volume of soil and A is the field they are taken from. In mining the purpose might be to assess the commercial viability of mineral exploitation in A; in the case of soil contamination the problem might be to assess whether children, eating a particular volume of soil, are likely to put themselves at risk of poisoning (Heuvelink et al., 1999). Point sample data are used to make predictions for areas. This is known as a change of support problem because the observed data are with respect to point supports $(s(1), \ldots, s(n))$ and the prediction is with respect to an area support (A).

For simplicity assume a constant mean so that for any point $E[Y(i)] = \mu$. It follows that $E[Y(A)] = \mu$. Intuitively it might be expected that the sample mean:

$$\bar{Y} = (1/n)\Sigma_{i=1,\ldots,n}\, y(i) \tag{4.13}$$

would make a good predictor of $Y(A)$. However a predictor that gives equal weighting to each observation is ignoring the geographical distribution of the sites relative to the location where the prediction is needed. In statistical terms this means ignoring the spatial correlation or dependence in Y – both in the sample data and between the sample datapoints and the location where the prediction is required. Furthermore (4.13) does not take into account the difference in the variances of $Y(s(i))$ and $Y(A)$. Computing an average over a large area has the effect of reducing the variance of the point sample data, although the reduction occurs less rapidly the more continuous, the more spatially correlated, the data (see, e.g., Isaaks and Srivastava, 1989, p. 462). It can be shown that $\mathrm{Var}(Y(A)) < \mathrm{Var}(\bar{Y})$ (see, e.g., Cressie, 1996, pp. 163–4). The sample mean is therefore an inefficient predictor (it has a large sampling variance) and this can lead to significant over or under prediction of $Y(A)$. In the context of commercial mining for example, where sampling is used to predict areas or blocks of commercially viable deposits, this could be a costly mistake.

If the surface mean is not known but can be assumed constant, from standard results in geostatistics (Cressie, 1996, p. 163), the best linear unbiased predictor of $Y(A)$, which means that of all linear unbiased predictors it has the minimum mean squared prediction error, is given by:

$$\hat{Y}(A) = [\mathbf{c}^T \boldsymbol{\Sigma}^{-1} + (\mathbf{1m})^T \boldsymbol{\Sigma}^{-1}]\mathbf{y} \tag{4.14}$$

where:

$$\mathbf{m} = (\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1})^{-1}(1 - \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{c}).$$

$\mathbf{1}$ is a column vector of *1*s; $\mathbf{c}^T = [\text{cov}(Y(A), Y(1)), \ldots, \text{cov}(Y(A), Y(n))]$ where cov denotes covariance. $\boldsymbol{\Sigma}$ is the $n \times n$ symmetric variance–covariance matrix. The (i, j)th element of $\boldsymbol{\Sigma}$ is $\text{cov}(Y(i), Y(j))$ and the ith diagonal element is the variance of $Y(i)$ which is also denoted $\text{Var}(Y(i))$. The predictor (4.14), called the *ordinary* block kriging predictor, addresses the shortcomings of (4.13). The inclusion of $\boldsymbol{\Sigma}^{-1}$ in the estimator adjusts for the spatial dependency between the sample data and the inclusion of $\mathbf{c}$ exploits the spatial dependency between the sample points and the area to be predicted.

All these quantities including the elements in $\mathbf{c}$ can be estimated from the point support data (Isaaks and Srivastava, 1989, pp. 324–7). There will be further, more detailed discussion of kriging in section 4.4.2(*v*). At this point however note that the prediction error for (4.14) is:

$$\text{Cov}(Y(o), Y(o)) - \mathbf{c}^T \boldsymbol{\Sigma}^{-1} \mathbf{c} + \mathbf{m}[1 - (\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{c})] - \text{Cov}(Y(A), Y(A)) \quad (4.15)$$

where $\text{Cov}(Y(o), Y(o))$ is the average point variance in the sample and $\text{Cov}(Y(A), Y(A))$ is the average within block variance. In due course the reader should compare (4.15) with (4.42). The prediction error measured by (4.15) is reduced relative to (4.42) by the final term, the average within block variance. In general the larger the support A the smaller the variance of Y. Equation (4.15) shows that the prediction error for $Y(A)$ is also smaller than the prediction error for any point prediction, say $Y(o)$. Webster and Oliver (2001, pp. 162–4) provide illustrative examples.

(b) Areal interpolation

Areal interpolation is undertaken in order to make attribute values recorded on one spatial framework available on another spatial framework. This may be required when attempting to make comparisons over time – for example across two or more censuses when it is likely that some census tract boundaries will have been altered. In another context an analyst needs to link socio-economic, environmental and health data in a single database on a common spatial framework. However, environmental data on air pollution may be recorded by grid squares from a model, land use may be recorded using area classes where boundaries denote changes of land use; socio-economic data may be from the census and based on enumeration districts; health data may be recorded by postcodes. Enumeration districts may be used for the common spatial framework. Health data records may be linked to enumeration districts using an ED/postcode directory (see Collins et al., 1998 for a discussion of the errors this can introduce). Kelsall and Diggle (1995) in an analysis of disease

rates attach numerator and denominator data to a common grid framework and construct a relative risk surface. Markoff and Shapiro (1973) provide some examples from historical research.

Mrozinski and Cromley (1999) classify areal interpolation problems into one of four types. However their first type, the missing data problem, will be considered in the context of data incompleteness (see section 4.4). In identifying the three other types, they distinguish between area-class map data and choropleth map data. Area-class data are data where the spatial unit boundaries reflect breaks or changes in the attribute level of a continuous or field variable – such as a land-use map. Attribute values are typically at the nominal or ordinal level of measurement. Choropleth data arise where the spatial unit boundaries are independent of the recorded attribute – such as when socio-economic and demographic data are recorded by census tract. Attribute data associated with choropleth maps are typically at the interval or ratio level. They may refer either to counts such as total population or total income (spatially extensive) or rates or averages such as per capita income or population density (spatially intensive) (Goodchild and Lam, 1980).

Type (1): choropleth data are available on one areal partition but need to be interpolated to another areal partition that may be a spatial framework that derives from an area-class framework or another, different, choropleth framework.

Type (2): data are available on two variables, one a choropleth map and the other an area-class map. Values need to be interpolated to the intersections created by overlaying them.

Type (3): data are available on two variables, both choropleth maps and values need to be interpolated to the intersections created by overlaying them.

In the case of type (1) problems there is a single set of 'source zones' that are to be interpolated to a different set of 'target zones'. In types (2) and (3) there are two sets of source zones which when overlayed create a set of target zones arising from their intersection. It is these new zones ('resels' or resolution elements) created by the intersection for which interpolated values are required. In type (1) and (2) problems the known 'volume' associated with the source zone data must be preserved on the new target zones on just one variable. In type 3 problems this volume preserving (or pycnophylactic) property needs to be honoured for both variables.

The methodology of areal interpolation divides into cartographic methods and statistical methods. Cartographic methods exploit information on which areas overlap and sometimes adjacency information which is used for smoothing to avoid sharp discontinuities on the interpolated surface and/or

preserve volume. Statistical methods typically involve fitting models to the source data drawing on a range of ancillary data to explain variation. These methods may be sequential (fit a model to the observed data and then use the model to interpolate) or iterative (fit model then estimate data values then refit the model and so on). We consider cartographic methods first and concentrate on type (1) problems, noting however that solutions to type (1) problems can usually be applied to type (2) problems. There are links between the cartographic methods described here and smoothing methods used in exploratory spatial data analysis which are discussed in section 7.1. The method chosen in any particular case must honour the type of variable, that is whether it is spatially extensive or intensive.

The known values on variable Z for n source zones are denoted $\{z(s(i))\}$ where $s(i)$ now signifies the ith source zone. If the source zone variable is a count (spatially extensive) the true unobserved value for Z in target zone j, $t(j)$, is:

$$z(t(j)) = \Sigma_{s(i)}\, z(s(i) \cap t(j)) \tag{4.16}$$

where $(s(i) \cap t(j))$ is the intersection of the ith source zone with the jth target zone and $z(s(i) \cap t(j))$ denotes the true but unobserved count associated with this area of intersection. If the source zone variable is a ratio (spatially intensive), the true unobserved value for Z in target zone j, $t(j)$ is:

$$z(t(j)) = \Sigma_{s(i)}\, z(s(i) \cap t(j))\, [A(s(i) \cap t(j))]/[A(t(j))] \tag{4.17}$$

where $z(s(i) \cap t(j))$ denotes the true but unobserved rate associated with this area of intersection and where $A(.)$ denotes area.

The problem is to *estimate* values of the variable of interest across the set of k 'target zones' – $\{\hat{z}(t(j))\}$. In the case where the source area $s(i)$ is represented by a single point, such as the centroid of a set of individual point objects that occupy $s(i)$, then $z(s(i) \cap t(j))$ in (4.16) can be estimated by $z(s(i))$ if the representative point falls within $t(j)$, but is 0 otherwise. This is called the point-in-polygon method. This method should work well if the individual objects cluster close to the representative point (Sadahiro, 2000).

A variant of the point-in-polygon method is the kernel method. This distributes the weight according to a probability density function (the kernel) around the representative point (Silverman, 1986; Sadahiro, 1999). Bracken and Martin (1989) used the kernel method to continuously map the spatial distribution of the UK population from enumeration district-level census data that are associated to population weighted centroids. Kelsall and Diggle (1995) use kernel density estimation to assign data from different irregular spatial frameworks to a common framework to compute disease rates. A problem with this is that there is often no basis on which to decide whether the real map meets

these conditions. The resulting map of the attribute on the set of target zones tends to be too smooth.

If the source zone variable is spatially extensive and can be assumed to be uniformly distributed within the source zone an estimator for (4.16) is given by:

$$\hat{z}(t(j)) = \Sigma_{s(i)}\, z(s(i))\,[A(s(i) \cap t(j))]/[A(s(i))] \tag{4.18}$$

If $z(s(i))$ is an area dependent ratio or proportion (spatially intensive) an estimator for $z(t(j))$ is given by substituting $z(s(i))$ for $z(s(i) \cap t(j))$ in (4.17) so:

$$\hat{z}(t(j)) = \Sigma_{s(i)}\, z(s(i))\,[A(s(i) \cap t(j))]/[A(t(j))] \tag{4.19}$$

Again (4.19) will be appropriate providing the ratio is uniform across the source zone. With this and other estimators it may not be straightforward to decide on the area of intersection. If the operations are performed in a geographic information system the analyst needs to guard against errors in recording area boundaries which compound when the two sets of zones are overlayed giving rise to false intersections.

These area weighting estimators (4.18) and (4.19) are based on geometric properties of the source and target zones. Where the assumption of a uniform distribution of the variable across source zones is known to be false, dasymetric mapping excludes from the area calculations those areas where the variable of interest is known to be 0. For example, water areas and industrial areas would usually be excluded in calculating population counts and satellite data overlaid on the source and target zone maps may be useful (Langford et al., 1991). Evidence suggests that this modification can substantially improve area-weighting estimates (Fisher and Langford, 1995; Mrozinski and Cromley, 1999). Brindley et al. (2002) allocate modelled pollution data at a fine spatial scale to enumeration districts using data on where the population are located within each ED. This is to obtain an ED-level measure of pollution that attempts to reflect more closely what the people who live there are exposed to. A representative point in the target zone is used to select the appropriate data value from the source zone.

If population varies as a function of some category like land use or rock type the dasymetric method needs to reflect this. Providing it is possible to obtain estimates on how Z varies as a function of category and providing the geographic distribution of the category within each target zone is known then this should improve on binary dasymetric area weighting. These are examples of so-called 'intelligent' interpolation (Fisher and Langford, 1995). Goodchild et al. (1993) use data from a third set of spatial units they refer to as control units and within which the spatially extensive variable Z is uniform. The control unit

boundaries need not be congruent with either the target or source zones. Let $c(k)$ denote control zone k then:

$$z(s(i)) = \Sigma_{c(k)}\, z(c(k))\,[A(s(i) \cap c(k))]/A(c(k)) \tag{4.20}$$

Now if $z(c(k))$ can be estimated, $\hat{z}(c(k))$, then:

$$\hat{z}(t(j)) = \Sigma_{c(k)}\, \hat{z}(c(k))\,[A(t(j) \cap c(k))]/A(c(k)) \tag{4.21}$$

Goodchild et al. (1993) discuss various statistical methods of estimating $\{z(c(k))\}$ using $\{z(s(i))\}$ including Poisson regression and constrained least squares.

Mrozinski and Cromley (1999) place areal interpolation for count data within a matrix framework also found in spatial interaction modelling. An intersection matrix can be defined comprising S rows (corresponding to the source zones) and T columns (corresponding to the target zones). The matrix consists of 1s where zones i and j intersect, 0 otherwise. The problem is to estimate the variable values associated with the non-zero cell entries in the intersection matrix. Variable values associated with the other cells in the matrix are known to be 0. In type (1) and (2) problems only row sums are known (a singly constrained model) whilst in type (3) problems both row and column sums are known (a doubly constrained model). Once cell values have been estimated target zone estimates can be obtained by aggregation. Iterative operations are performed on cell entries until convergence takes place. The operations of polygon smoothing and dasymetric polygon smoothing (for type (1) and (2) problems) which Mrozinski and Cromley describe are extensions to Tobler's (1979) original pycnophylactic method. The methodology involves assigning to any intersection zone a density given by the source zone it sits within, averaging this density with values from the contiguous intersection zones and iterating until convergence occurs. A volume-preserving constraint is introduced and counts are obtained at the end by multiplying the density values with the areas of the intersection zones. Polygon smoothing performed a little better than area weighting in their test data but when the dasymetric adaptation was introduced there was no clear winner in terms of overall root mean squared error. These methods raise a number of questions about the existence and uniqueness of solutions under different conditions including the effect of increasing the number of intersections as well as how to represent the interaction between the source and target zones and what types of ancillary data might be useful.

The other main approach to areal interpolation draws on statistical modelling. Flowerdew and Green (1989) propose a statistical modelling approach to areal interpolation which can be used if ancillary data are available in the form of area class or choropleth data. Their approach is an extension of the

'intelligent' method of interpolation for spatially extensive data and is similar to Goodchild et al. (1993). They apply their method to estimate population counts for target zones where population varies according to whether land type is grassland (1) or woodland (2). The full data set $\{z(s(i))\}$ together with data for each source zone on the area under each land-use type, $\{A(s(i); 1)\}$ and $\{A(s(i); 2)\}$, are used to estimate the parameters of a Poisson regression model (with identity link and no constant term). These estimated parameters ($\hat{\lambda}_1$, $\hat{\lambda}_2$) are the expected population counts per unit area for each land-use type. Providing the area under each land-use type in each target zone is known ($\{A(t(j); 1)\}$, $\{A(t(j); 2)\}$) then:

$$\hat{z}(t(j)) = \hat{\lambda}_1 A(t(j); 1) + \hat{\lambda}_2 A(t(j); 2)$$

Estimates for smaller zones formed by the intersection of the population map and the land-use map can also be determined by this method. The counts will need scaling to ensure the target zone total matches the source zone total. Langford et al. (1991) develop a similar approach where the ancillary datum is a land-cover classification derived from remotely sensed data.

Flowerdew, Green and Kehris (1991) generalize the model-based approach using the intersection matrix described by Mrozinski and Cromley (1999). In their generalization, variable values are estimated using the EM algorithm which is a statistical technique for dealing with estimation problems where there are missing data (see section 4.4). The algorithm computes expected (E) values of the 'missing data' (the variable values in the non-zero entries of the intersection matrix corresponding to the target zones) given the model and the observed data. At the first iteration these values could be derived from the areal weighting approach. The completed data set is then used to fit the model by maximum likelihood (M). The algorithm then iterates between the E and M steps until convergence occurs. The example by Flowerdew et al. (1991) again uses the Poisson model but others can be used in the algorithm.

There are problems with their model-based approach as Flowerdew et al. (1991) observe. The method is computationally intensive and processing speed is much slower than the point-in-polygon method. This is an important consideration when processing large spatial databases like national census data (Sadahiro, 2000). Model goodness of fit is no guarantee as to how well the target zone values will be estimated by the model. Parameter estimates in the Poisson regression model are global estimates and do not take into account possible spatial heterogeneity in the association between, say, population and land-use type.

The above factors may help to explain why area weighting and simple dasymetric methods based on localized mapping often perform better than more

elaborate methods in comparative trials (Fisher and Langford, 1995). Cockings et al. (1997) compare interpolation methods using Monte Carlo methods and show that, whilst areal weighting method errors are strongly related to geometric properties of the target zones, dasymetric method errors are more correlated with population or attribute parameters. However trying to make links between error properties and characteristics of the zones is made difficult by the complex inter-relationships between the various confounding effects (Cockings et al., 1997, p. 327).

An underlying problem in this area is that the analyst performs these operations in order to transform the data on to a more appropriate spatial framework. No new data are created, rather estimates are obtained based on transforming the original data. If the analyst now treats these estimates as data, these new 'data' now contain *additional* unknown errors. However there are relatively few criteria to help decide which method to use in any particular situation and no diagnostics to assist in evaluating when a chosen method is or is not doing well and no quantitative measure of the precision of the estimates.

Empirically based studies indicate the performance of different methods in particular cases but leave open the question of the generality of findings. This experimental shortcoming can be overcome by Monte Carlo methods (Fisher and Langford, 1995). It can also be handled using the stochastic modelling approach of Sadahiro (1999, 2000). However these analyses at best indicate the relative performance of different methods under different conditions. The presence of spatial heterogeneity means that any one method may perform very differently in different subareas of the map. The situation is analagous to trying to analyse the effects of error propagation (see section 4.1.3).

Areal interpolation is an estimation problem but we are not in a position to attach confidence intervals to the estimates or to map the geography of the likely errors. The success of most estimators appears to depend on satisfying a set of often quite restrictive assumptions. Where there is strong spatial correlation at a scale that exceeds the scale of the areal partition then iterative smoothing methods should do well but perhaps not otherwise. The maps in Mrozinski and Cromley (1999) suggest that it is where there are sudden changes in population density associated with urban places with sharp boundaries that large interpolation errors are found. In these circumstances further ancillary data will be helpful.

This is a very problematic area of spatial analysis and will remain so whilst data are collected on different spatial frameworks. The past focus has been on methods for transforming data from one spatial framework to another but perhaps more attention needs to be given to the *consequences* of having to change spatial frameworks. This means assessing the robustness of findings to induced

errors and exploring the variability in model findings that arise from data uncertainty.

4.2.3 Analysing relationships using aggregate data

Ecological inference is the process whereby grouped data (also known as ecological or aggregate data) are analysed and results used to infer individual-level relationships. It is of interest to those working with spatial data because data aggregated by area is one type of grouped data.

Ecological inference is important because there may be a lack of reliable individual-level data. It may be impractical to collect data at an individual level or if it were collected it would only be available with uncontrollably large imprecision, for example exposure rates to environmental risk factors. Small-area averages of natural radiation levels may actually be a more reliable basis for inference than data based on estimating individual exposure and show greater robustness to measurement error. Some data may only be available for areas, for example income data and electoral data.

Gelman et al. (2001) use the following model to represent the nature of ecological inference. Consider $j = 1, \ldots, m(i)$ individual units in area $i\,(i = 1, \ldots, n)$. Let $y(j, i)$ be the response for individual j in area i, let $x(j, i)$ be the corresponding predictor and let the $e(j, i)$ be independent with mean 0 and also independent of the $x(j, i)$. Define:

$$y(j,i) = \alpha(i) + \beta(i)\,x(j,i) + e(j,i) \tag{4.22}$$

The interest is in estimating $\alpha(i)$ and $\beta(i)$, but no individual level data are available, only averages for each area denoted $\bar{y}(i)$ and $\bar{x}(i)$. Assuming that average errors $\{e(i)\}$ are close to 0 then it follows from (4.22):

$$\bar{y}(i) = \alpha(i) + \beta(i)\,\bar{x}(i) \tag{4.23}$$

Ecological regression estimates the parameters in (4.23) by regressing $\bar{y}(i)$ on $\bar{x}(i)$ with one datapoint per area, that is fitting:

$$\bar{y}(i) = \alpha + \beta\,\bar{x}(i) + u(i) \quad i = 1, \ldots, n \tag{4.24}$$

where $u(i)$ is independent with mean zero and also independent of the $\bar{x}(i)$. Ecological inference involves using the estimates of the aggregate parameters α and β in (4.24) in place of the local regression coefficients $\alpha(i)$ and $\beta(i)$.

Ecological (or aggregation) bias is the difference between estimates of relationships obtained using grouped data (4.24) and those estimates obtained using individual-level data (4.22). The analyst who takes the estimate

obtained from grouped data and uses it to infer an individual-level relationship, without specifying the conditions under which the estimates are reasonable, would be said to be guilty of committing the ecological fallacy.

Simpson (1951) drew attention to the dangers of only analysing the margins of complex (multi-dimensional) contingency tables and there is much evidence showing how correlation and regression estimates yield different results depending on whether individual- or aggregate-level data are analysed. Early examples can be found in Gehlke and Biehl (1934), Neprash (1934) and Robinson (1950). Robinson's study of race and literacy in the USA showed that whilst at the individual level the tetrachoric correlation for the 2×2 table of counts was only 0.203, at the state level the product moment correlation of the percentage figures was 0.773. Assembled evidence in epidemiology and elsewhere suggests that the correlation coefficient is more seriously affected than regression parameters (Firebaugh, 1978 and Morganstern, 1982).

Other forms of aggregation bias can arise in spatial analysis. There are potentially two sources of aggregation bias in modelling the relationship between say house prices and a set of predictor variables that includes distance from the city centre using aggregate data. First the use of areally grouped data for the response and predictor variables, second the use of the area centroid to represent the location of the houses in any unit rather than the true spatial average of the individual houses in each areal unit (Okabe and Tagashira, 1996).

Inferring individual-level relationships from grouped data is important in many scientific areas because the individual is the object of study and hence the target of inference. However groups can act as effect modifiers and hence so can areas. The overall composition of the constituency in a first-past-the-post electoral system could affect whether and how individuals vote. There is evidence that individual propensities to commit acts of vandalism can be influenced by the density of young males in an area. Individual preventative health behaviours may be influenced by norms and attitudes in the immediate environment. The analyst will want to try to distinguish individual-level effects from area-level effects. The converse to the ecological fallacy is the atomistic fallacy (disaggregation bias) which is the error that can arise from identifying associations from individual-level data whilst ignoring area-level or contextual effects. Multi-level modelling (see section 9.2.3) is used to model these types of problems.

In fields such as geography and areas of environmental science the target of inference is often at the area level rather than at the individual level. How do different forms of spatial aggregation affect statistics like correlation and regression which are known *not* to be invariant to aggregation effects? When carrying out comparative work including checking findings in one area by

carrying out a parallel study elsewhere or at a different time in the same place the analyst might be concerned that any differences observed do not reflect real differences. They might, in fact, be an artefact of two different spatial aggregations that have produced different degrees of within-area homogeneity.

In geography the generic name for this is the modifiable areal units problem (MAUP). 'If a statistic is calculated for two different sets of areal units which cover the same population, or sample, a difference will usually be observed even though the same basic data have been used in both analyses. This difference is cited as evidence of the modifiable areal units problem' (Holt, Steel and Tranmer, 1996, p. 181). The term 'modifiable' is used because neither the choice of number of spatial units (the scale of the analysis) nor their particular configuration (the selected partitioning or zoning given the scale of analysis) is fundamental and any one of a number of other choices could have been made. For a brief overview of the extensive work in geography examining the volatility of regression parameters and correlation coefficients see Wong and Amrhein (1996).

If results differ between different scales of analysis this may reflect the operation of scale-dependent processes. Two economic activities might compete for land at a local scale but be found clustered together when their distribution is examined at a larger scale (using larger spatial units) because of input–output linkages or other forms of interaction. If altering the partition whilst holding scale constant produces a different set of results this is because the new partition has introduced a different smoothing of the data and brought about a reconfiguration of the within-area homogeneity.

In fields such as epidemiology (dose–response relationships) and econometrics (quantity–price relationships) ecological inference is undertaken with continuous-valued variables. In political science, ecological inference is used on data where variable values at the individual level are binary (vote–not vote). Ecological inference is undertaken where: (i) there are no reliable individual-level data; (ii) there are individual-level data on the variables of interest, for example, originating from a national survey which has provided some sparse coverage across smaller spatial units; (iii) there is individual-level data available on a set of other variables – so-called 'grouping' variables – that explain the within-area homogeneity that underlies the causes of ecological bias. A distinction can also be made between studies in terms of their target of inference. Is the target of inference the individual-level relationship *within* each of the areas that partition the study region so that there are as many inferences as there are areas in the study region (the conditional approach), or is the target of inference the individual-level relationship *across* all the areas in the study region (the marginal approach)?

We now consider three aspects of the problem of making valid inferences from aggregate data.

(a) Ecological inference: parameter estimation

We review the problem of ecological inference for discrete valued data (on both the response and predictor variables of (4.23)) and draw on King (1997), Holt et al. (1996), Wrigley et al. (1996) and Tranmer and Steel (1998) to identify what underlies aggregation bias. Four methodologies for pursuing ecological inference, and in particular estimating parameters of interest, due to Goodman (1953), Freedman et al. (1991), King (1997) and Tranmer and Steel (1998) are described.

Let $T(i)$ denote the proportion of the voting-age population who turn out to vote in an election in area i – called the turnout in area i. Let $X(i)$ denote the proportion of the voting-age population who are non-white in area i, so that $1 - X(i)$ is the proportion who are white. $\{T(i)\}$ and $\{X(i)\}$ are the (aggregate) data. It follows that if $\beta^{o}(i)$ is the proportion of voting-age non-whites who vote and $\beta^{w}(i)$ is the proportion of voting-age whites who vote in area i then by definition:

$$T(i) = \beta^{o}(i)\,X(i) + \beta^{w}(i)(1 - X(i)) \tag{4.25}$$

The parameters of interest are $\{\beta^{o}(i)\}$ and $\{\beta^{w}(i)\}$ and they are unknown. If individual-level data were available the cells of the 2×2 contingency table (vote/not vote; white/non-white) could be filled in and the parameter values read off. In the absence of such data, if there are n areas then there are n equations but twice as many ($2n$) parameters. The parameters of interest in (4.25) cannot therefore be estimated and this is called the indeterminacy problem. ((4.25) is a reparameterization of (4.23) for discrete-valued variables so that the area averages are proportions. In (4.23) the equivalent problem to that described for estimating (4.25) is that in (4.23) there is one equation and two unknowns.)

One approach to tackling this problem is Freedman et al.'s (1991) neighbourhood model which assumes that at the chosen level of aggregation ethnicity has no influence on voting behaviour so that $\beta^{o}(i) = \beta^{w}(i) = T(i)$ (see Gelman, 2001, p. 105). Goodman's (1953) regression approach makes the constancy assumption that $\beta^{o}(i) = B^{o}$ and $\beta^{w}(i) = B^{w}$. Although voting propensity is allowed to vary between ethnic groups it does not vary by area. The parameters can be estimated by weighted least squares in a regression model of $T(i)$ on $X(i)$ and $(1 - X(i))$ where the intercept coefficient is set at 0. The weights are chosen to reflect the argument that the variance in the dependent variable $T(i)$ is expected to be inversely proportional to the population in the ith area (see

section 4.2.1). King (1997, pp. 56–73) has an extended discussion of the problems associated with estimating Goodman's regression model. These two methods adopt the marginal approach to individual-level inference and provide no evidence on how voting propensities might vary between voting areas.

Consider now the nature of ecological or aggregation bias arising in the Goodman regression approach to estimation. Following King (1997, p. 45) write: $B = B^{o} - B^{w.}$ Write $\hat{B}$ for the difference in the weighted least squares estimates of B^{o} and B^{w}. That is: $\hat{B} = \hat{B}^{o} - \hat{B}^{w}$. Aggregation bias is then given by: $\hat{B} - B$.

Now, let $X(j, i)$ and $T(j, i)$ denote the individual-level data where (j, i) signifies individual j ($j = 1, \ldots, m(i)$) in area i ($i = 1, \ldots, n$). These are binary (0 or 1) variables. Let $\mathbf{X}$ and $\mathbf{T}$ denote the vectors of length $N (= \Sigma_{i=1,\ldots,n}\, m(i))$ that contain the individual-level data. Now let $\bar{\mathbf{T}}$ and $\bar{\mathbf{X}}$ denote the simple average of $\{T(j, i)\}$ and $\{X(j, i)\}$ over all individuals in all areas. Then:

$$[\mathbf{T} - \bar{\mathbf{T}}] = [\mathbf{X} - \bar{\mathbf{X}}]B$$

This is the individual-level expression of the relationship, and follows from the definition (4.25) after introducing Goodman's constancy assumption. The value of B (the individual-level parameter) can then be computed exactly either by least squares or as a cross tabulation:

$$B = [[\mathbf{X} - \bar{\mathbf{X}}]^{T}[\mathbf{X} - \bar{\mathbf{X}}]]^{-1}\,[\mathbf{X} - \bar{\mathbf{X}}]^{T}[\mathbf{T} - \bar{\mathbf{T}}]$$

Now consider the effects of aggregation. Aggregation is introduced via a N by n grouping matrix ($\mathbf{G}$) which assigns individuals to areas so that $g(k, i) = 1$ if individual $k (k = 1, \ldots, N)$ is in area i ($i = 1, \ldots, n$). It can be shown (see King, 1996, p. 47) that $\mathbf{G}(\mathbf{G}^{T}\mathbf{G})^{-1}\mathbf{G}^{T} = \mathbf{H}$, when multiplied with a vector, replaces each individual observation by its area mean. So, the aggregate version of the model is:

$$\mathbf{H}[\mathbf{T} - \bar{\mathbf{T}}] = \mathbf{H}[\mathbf{X} - \bar{\mathbf{X}}]B$$

and B is obtained by least squares of $\mathbf{H}[\mathbf{T} - \bar{\mathbf{T}}]$ on $\mathbf{H}[\mathbf{X} - \bar{\mathbf{X}}]$ or by weighted least squares of $[\mathbf{T} - \bar{\mathbf{T}}]$ on $[\mathbf{X} - \bar{\mathbf{X}}]$:

$$\hat{B} = [(\mathbf{H}[\mathbf{X} - \bar{\mathbf{X}}])^{T}\mathbf{H}[\mathbf{X} - \bar{\mathbf{X}}]]^{-1}(\mathbf{H}[\mathbf{X} - \bar{\mathbf{X}}])^{T}\mathbf{H}[\mathbf{T} - \bar{\mathbf{T}}]$$

Estimation bias induced by the aggregation can now be examined by analysing $(\hat{B} - B)$ but the challenge is to put the difference into a form that sheds light on the source of the discrepancy. King (1997, pp. 46–53) discusses a number of different representations. Following a derivation by Palmquist the

discrepancy can be decomposed into the *product* of two components (see King, 1997, pp. 51–3). The first component is what is termed a 'specification shift' which is described as the effect of using individual data but being forced to include the grouping operator **G** in the regression. King (1997, p. 52) describes this problem as 'reverse omitted variable bias'. This enforced specification shift will have most effect on the estimate of B when **G** (the system of aggregation) causally intervenes between T and X. *Hence any form of homogeneous grouping using either of the two variables X and T or any other variable causally associated with X or T, will introduce specification shift*. The method of aggregation that has the least effect is random aggregation.

The second component is called an 'inflation factor' and is given by:

$$\left[\sigma^2_{(X(j,i))} / \sigma^2_{(X(i))}\right] - 1 \tag{4.26}$$

where $\sigma^2{}_{(X(j,i))}$ is the variance of X over all individuals $(X(j,i))$, and $\sigma^2{}_{(X(i))}$ is the variance over all the group-level values $(X(i))$. *This quantity* (4.26) *is least when aggregates that are homogeneous in X are used (and 0 when all aggregates comprise individuals identical in terms of X) and considerably greater when random aggregates are used*.

This analysis identifies two situations when aggregation bias will be small – either when random aggregates are used (because, although the inflation factor is large, the specification shift is zero) or when pure homogeneous aggregates on X are used (because, although the specification shift is large, the inflation factor is 0). This observation follows because the bias is a product of these two components.

Spatially defined groups will display some degree of homogeneity because of spatial correlation. In the case of census data: 'individuals who live in the same area are exposed to common influences and as a result exhibit similarities . . . individuals with similar characteristics choose to live in the same area' (Tranmer and Steel, 1998, p. 818). This means that grouping by geographical proximity will tend to produce aggregates that are homogeneous in terms of the independent variable (X in the above example) and/or the dependent variable (T in the above example) rather than random aggregates. This introduces specification shift but providing the aggregate is homogeneous so that there is no within-area variation the analyst can expect no aggregation bias because the inflation factor associated with homogeneous aggregates is zero. In practice, of course, areal aggregates are rarely if ever purely homogeneous which is why aggregation bias is always present. Where the analyst has the opportunity to construct his or her own areal aggregates (areas or regions) these findings have implications for the criteria that should be employed in region building. Approaches to region building are considered in the context of exploratory spatial data analysis (see section 6.2.2).

Tranmer and Steel (1998) analyse aggregation bias on variance, covariance and correlation statistics using a model for within-area homogeneity in which:

$$y(j,i) = \mu + \alpha(i) + e(j) \tag{4.27}$$

where μ is a regional mean effect; $\alpha(i)$ is a random variable with 0 mean and variance $\sigma^2(i)$ representing the area effect associated with the ith area; $e(j)$ is a random variable with a mean of 0 and variance σ_e^2 representing a pure individual effect. The area-level effect is common to all individuals from the same area. A similar variance–covariance model is specified for a second variable and covariances between the area effects and the individual effects are defined for the two variables.

This is a multi-level model. Intra-area homogeneity is modelled by the random variable $\alpha(i)$ because all individuals in the same area have the same value of $\alpha(i)$. The $\{\alpha(i)\}$ are independent so area-level effects are independent. In the case of an aggregate such as an enumeration district this means individuals across the street but in different enumeration districts have independent area-level effects but two individuals some distance apart but in the same enumeration district have identical area-level effects. This approach to analysing the effects of intra-area homogeneity has also been used by Arbia (1989). Cliff and Ord (1981, p. 127) constructed a nested hierarchical model for two variables, each with m hierarchical levels together with a common factor at each of the m levels.

In analysing the difference between correlation coefficients computed on individual-level data and on area-level data Tranmer and Steel (1998) identify the difference as a function of the relative sizes of three quantities denoted $\delta_{1,1}$, $\delta_{2,2}$ and $\delta_{1,2}$ (pp. 823–4). Two of these ($\delta_{1,1}$ and $\delta_{2,2}$) are each inversely related to Palmquist's inflation factor and refer to the two variance terms (corresponding to the two variables) in a correlation estimate. The third, $\delta_{1,2}$, is a within area measure of cross-correlation between the two variables and measures the effect of the aggregation on the estimate of the relationship between the two variables. This term corresponds to Palmquists specification shift effect. The size of any aggregation bias in estimating the individual-level correlation coefficient using aggregate data is a function of these three terms and also the average number of individuals per area. This latter term can also have a serious effect on the bias because it multiplies the effects of each term $\delta_{1,1}$, $\delta_{2,2}$ and $\delta_{1,2}$ and is often large – even an enumeration district may have about 500 individuals.

Aggregation bias in estimating regression coefficients follows from their analysis as a function of the same terms except that there is only one variance inflation term which is associated with the independent variable in the regression model (Tranmer and Steel, 1998). Tranmer and Steel note that their results

can be used to disaggregate individual-level and area-level correlation and shed light on the MAUP by showing how different zonings by producing different effects on the three data dependent quantities produce different values of correlation and regression statistics.

King (1997) provides conditional estimates of voting propensities that vary by ethnicity and voting area. These area-level estimates can then be treated as fixed values and mapped (see, e.g., p. 25) or their variation modelled in terms of independent explanatory variables. The deterministic method of bounds (King, 1997, pp. 77–90) constrains estimates of the propensities to lie at least within the interval [0, 1] and possibly a narrower band depending on information provided by the data. The voting propensities $\beta^{o}(i)$ and $\beta^{w}(i)$ are assumed to be randomly drawn from a truncated bivariate normal probability distribution, conditional on $X(i)$, and specified by mean and variance–covariance parameters that are estimated from the data on $\{X(i)\}$ and $\{T(i)\}$. The truncation is set by the method of bounds and the same truncated distribution is used to draw the propensities for each area, although the conditioning (on $x(i)$) will vary.

King's approach treats the voting propensities that would be obtained even were the full table available as noisy estimates of an underlying true voting propensity. The truncated normal model represents a prior distribution for the voting propensities, and its parameters are hyperparameters. Because these hyperparameters are estimated from the full data set for $\{X(i)\}$ and $\{T(i)\}$, the methodology 'borrows strength' from the whole data set in providing estimates of individual-area propensities. A number of extensions to the methodology have been discussed in order to further reduce aggregation bias, including incorporating covariates into the specification of the mean value hyperparameters (King, 1997, p. 170).

The Bayesian approach proposed by King (1997) has generated considerable comment with criticism focusing on the reliability of the standard errors as indicators of the magnitude of actual estimation errors, the subjectivity of the diagnostics and the generality of the method (Freedman et al., 1998, 1999; King, 1999). Diagnostics provide checks on data requirements and statistical assumptions and, at the time of writing, criticism focuses on whether the available diagnostics provide sufficient information to assess when it is safe, and more importantly when it is not safe, to use the method. This area of ecological inference involves estimating parameters but diagnostics in relation to the underlying assumptions of the method are critical if the user is to be able to make a valid assessment of the reliability of the answers. Gelman (2001, pp. 105–7) discusses model checking for ecological regression and identifies what information is available to suggest when a model should be rejected.

An area of concern when analysing spatial data is the effect on King's methodology of spatial correlation in the parameters of interest. King (1997, pp. 164–8, 2000) concludes from his experience based on Monte Carlo simulation that spatial correlation 'does not appear to have major consequences for the validity of inferences from the basic model' (p. 168). However, the following parallels with Bayesian modelling of area-specific relative risk rates are worth noting. A spatial model rather than a model of spatial independence could be used for the prior distribution as employed in some areas of spatial epidemiology (Mollie, 1996). A specification that includes spatial dependence has the effect of restricting the 'borrowing of strength' to just the adjacent areas and on the evidence from the spatial epidemiology literature can for certain models result in considerable smoothing of the final map – which may or may not be appropriate. The expected presence of area-level contextual effects and the operation of social networks inducing contagion into the process of deciding whether to vote and if so for which party further complicates the problem of specifying an appropriate prior model.

The estimation methods discussed to this point assume that no individual-level data are available. Tranmer and Steel (1998) consider the situation where there are some limited individual-level data on other variables though not the variables for which inferences need to be drawn. The context for this work is the availablity in the UK of the 2% Sample of Anonymized Records (SAR) relating to districts. Their work shows that adjustments to ecological-level correlations and regression estimates can be made where appropriate individual-level data are available on a set of other variables called 'grouping variables'. Whereas in (4.27) within-area homogeneity appears as an unobserved area-level effect (as also in King's basic model), the grouping variables are introduced because they are observable and are associated with the within-area homogeneity. King's specification shift or 'reverse omitted variable bias' is addressed (in the context of their model) by specifying the variables that the enforced aggregation has introduced into the individual-level regression. It is these variables which need to be controlled for if a reliable estimate of the relationship (between the two variables of interest) is to be obtained.

It is not necessary for the individual-level data to have a locational reference nor even to come from the same data source, providing they refer to the same population. Their approach, which is also discussed in a different context in Cressie (1996) involves the inclusion of data on these variables, as 'grouping variables', into the multi-level model. Thus the original model (4.27) now becomes (Tranmer and Steel, 1998, p. 827):

$$y(j;\, i) = \tilde{\mu} + \boldsymbol{\beta}^T \mathbf{z}(j) + \tilde{\alpha}(i) + \tilde{e}(j)$$

where the tilde over the original terms denotes that they are now conditional on the grouping variables $\mathbf{Z}$; $\mathbf{z}(j)$ denotes the (column) vector of values of the grouping variables for the ith individual and $\boldsymbol{\beta}$ is the (column) vector of coefficients that relate $y(j;i)$ to the grouping variables $\mathbf{Z}$.

Tranmer and Steel (1998, pp. 829–30) derive adjusted correlation and regression coefficients that they show yield a considerable improvement in the estimation of the individual-level relationships. The estimation bias introduced by the aggregation into the covariance between two variables $Y(1)$ and $Y(2)$, and which is captured by the grouping variables is removed by a term:

$$\bar{\boldsymbol{\beta}}_{Y(1),\mathbf{z}}{}^{T}(\bar{\mathbf{S}}_{\mathbf{z},\mathbf{z}} - \hat{\mathbf{S}}_{\mathbf{z},\mathbf{z}})\bar{\boldsymbol{\beta}}_{Y(2),z}$$

where $\bar{\boldsymbol{\beta}}_{Y(1),\mathbf{z}}$ is a (column) vector of estimates of the regression coefficients of $Y(1)$ on the grouping variables at the aggregate level and similarly for $\bar{\boldsymbol{\beta}}_{Y(2),\mathbf{z}}$. $\bar{\mathbf{S}}_{\mathbf{z},\mathbf{z}}$ is the aggregate level covariance matrix of the grouping variables and $\hat{\mathbf{S}}_{\mathbf{z},\mathbf{z}}$ is the estimate of the individual-level variance–covariance matrix (on the grouping variables). The adjustment for the variances follow and from these quantities the adjusted correlation and regression estimates can be computed. Tranmer and Steel (1998) suggest that individual-level data from the SAR on variables such as age, housing and ethnic group structure may function effectively as grouping variables. Figure 4.3 illustrates the improvement obtained through this method.

(b) Ecological inference in environmental epidemiology: identifying valid hypotheses

In environmental epidemiology ecological inference arises in the analysis of certain types of dose–response relationships. For example 'response' is the rate of a disease by area and 'dose' refers to exposure to an environmental risk factor by area. Unlike the earlier problem, the data may not be discrete valued. An individual may either have the disease or not, but their exposure to the risk factor may be continuous valued. This raises additional problems because more information is lost as a consequence of aggregation than in the discrete-valued case. With discrete data only cell values in a table are lost and in the 2×2 case where marginal sums are available it is sufficient to estimate only one value in order to be able to complete the table. In the case of continous data the mean provides information only on the centre of the distribution of values, nothing on the spread of values is available. Further, in the voting example linearity was a property of the relationship (it did not have to be assumed), but in epidemiology there is no reason to expect dose–response relationships to be linear.

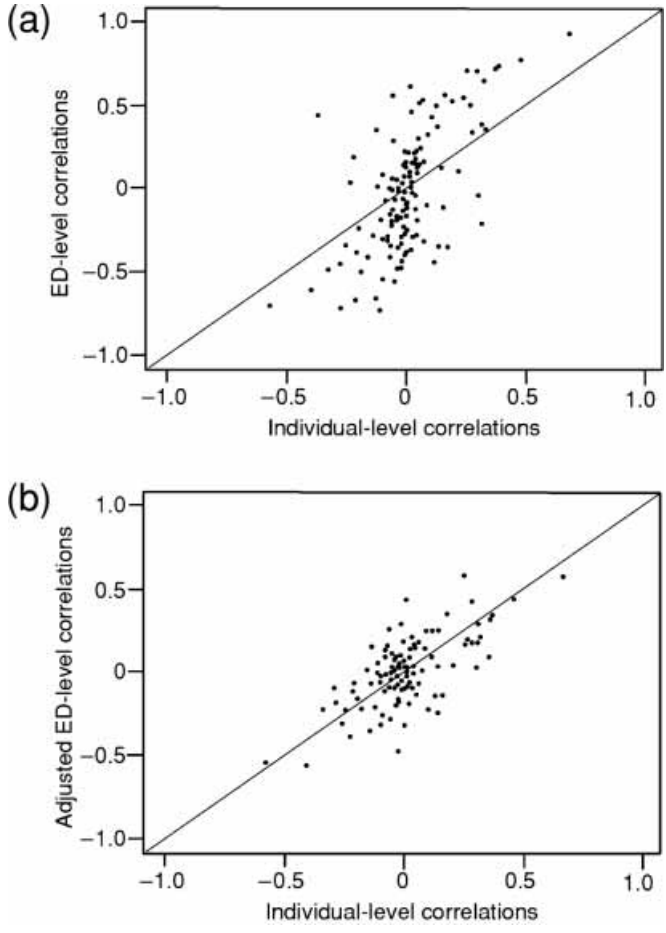


Figure 4.3 ED-level versus individual-level correlations for the Reigate data (a) before and (b) after adjustment (Tranmer and Steel, 1998, pp. 827 and 830)

Richardson (1992) identifies aggregation problems of interest to epidemiologists. (I) What is the difference between individual dose–response relationships and the relationship obtained after aggregating over all individuals in a group? If all individuals have identical parameters in a linear relationship then these parameters can be estimated from the aggregate data – but not if the relationship is non-linear. (II) Assuming linearity of the dose–response relationship within each aggregate, the difference between the estimated slope coefficient from the *N* aggregate values and the average of the *N* within aggregate slope coefficients is the ecological bias. This ecological bias can be decomposed into two elements. The first component is bias due to inter-aggregate variation in the disease rate amongst those *not* exposed to the risk factor (differences in the intercept parameter of the linear relationship across aggregates).

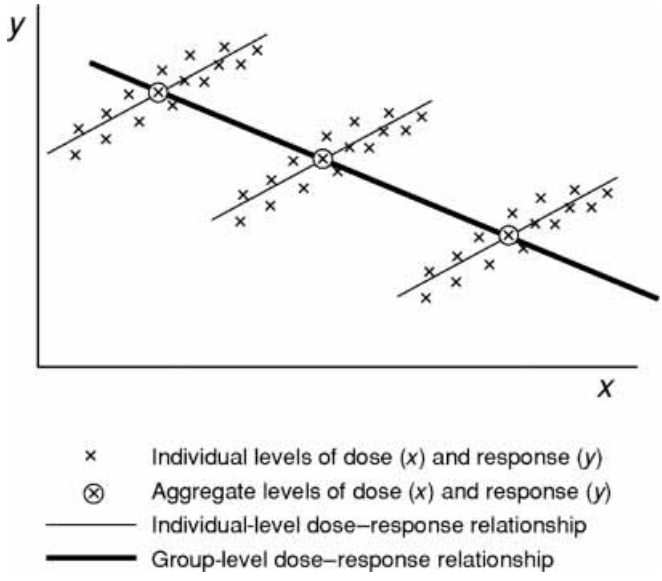


Figure 4.4 Sign reversal in ecological regression

The second is bias due to groups acting as effect modifiers in the dose–response relationship (differences in the slope parameter across aggregates). A further source of bias is the effect due to confounding variables, such as socio-economic variables not allowed for in an analysis (Jolley, Jarman and Elliott, 1992). These biases can be arbitrarily large even resulting in sign reversals as shown in figure 4.4. If the source of the bias can be identified it may be possible to adjust using multiple regression providing the assumptions of the model (e.g. additive joint effects) are satisfied (Greenland and Morganstern, 1989).

The problems associated with ecological inference mean that the analysis of aggregate data in environmental epidemiology is viewed as the weakest form of analysis for establishing exposure–disease relationships (although see Armstrong, 2001, for a defence of ecological analysis). Such analyses are often undertaken in order to generate scientifically valid hypotheses about the existence of an association rather than estimating its strength (Richardson, 1992). These hypotheses may then be followed up by other methods such as cohort studies, particularly if there is supporting evidence from other independent studies. In environmental epidemiology, the emphasis is on design construction to try to counter the problems of ecological analysis that include specification bias, confounding exposure misclassification and effect modification. Richardson (1992, pp. 199–200) lists good practice in the design of geographical studies: allow for heterogeneity of exposure; use well-defined population groups; employ survey data to help identify relevant exposure data; allow for

latency times in the disease; allow for migration effects. The last considerations suggest that time series data on population movements will be important in helping to identify associations. Particular care needs to be taken to try to identify all relevant confounders especially socio-economic and lifestyle factors and other environmental factors – although in practice it is impossible to be sure that all relevant confounders have been allowed for.

(c) The modifiable areal units problem (MAUP)

Tobler (1989) identifies two forms of the MAUP. The first of these is observing the effects of different spatial aggregations and the interpretation of the patterns revealed by different aggregations. Analysis which attempts to measure area-level effects (either because areas are the object of interest or because they are effect modifiers) will need to pay particular attention to the choice of the areal framework to ensure it is meaningful in terms of the underlying processes. The second of Tobler's forms of the MAUP is observing the effects of different aggregations on the behaviour of statistics. Underlying this is the concern that the analyst may be using a statistic that is not appropriate to the problem posed.

The MAUP consists of two distinct effects on the properties of estimators: those associated with the *scale* of the analysis and those associated with the particular *partition* (given the scale of the analysis). By 'scale of analysis' is usually meant the number of subareas a study area is partitioned into because this determines the size of each spatial unit (the areal filter) through which events are observed. Holt et al. (1996) point out that this is imprecise terminology and that it is necessary to specify which property of an estimator is affected – whether it is for example its expected value or its variance. Research in this area, particularly in geography, has tended to focus on the former.

Consider first the issue of scale and in particular the effect of analysing the same set of variables for the same region but repeating the analysis using different sized areal partitions. Measures of association will vary if scale-dependent processes are influencing outcomes. It was noted in the case of ecological inference that measures of association at any given scale of aggregation confound different scales of association. The scale 'problem' is to derive measures of association at any given scale that are pure measures of the association at that scale and not confounded with effects from smaller or larger scales. This calls for multi-level modelling.

In situations where all relationships between variables are a consequence of individual-scale processes so that there are no area-level (contextual) processes, one solution to the partition problem is to select statistics that are invariant to aggregation. This implies discarding statistical techniques like correlation

and regression, indeed all statistics based on computing variances, since they are not invariant to aggregation. Where area-level effects are present, this strategy is inappropriate since such a statistic would by construction be ignoring a potentially important element in the relationship. Another solution, as discussed above, is to develop adjustments to the way the statistics are computed in order to separate area-level from individual-level effects (Holt et al., 1996). Grouping variables may be used as discussed by Cressie (1996) and Tranmer and Steel (1998). Green and Flowerdew (1996) observe that just as estimates of relationships at the individual level are affected by area-level influences, so the area level is likely to be affected by higher, regional-level effects. They suggest adding the regional values of the independent variable into the regression model. These regional values for the *i*th area are averages for a defined area (typically the adjacent neighbours) surrounding the *i*th area. Unlike multi-level modelling which constructs a strict hierarchy these higher-level spatial units overlap (see section 9.2.1 and equation (9.31)).

Where data for small areas are available regionalizations that seek to control for partition effects might be constructed. If analysis involves making comparisons between two regions, partitions might be constructed for the two regions that are similar in terms of their levels of within-area homogeneity for each of the variables (Openshaw, 1996). This is likely to be difficult to achieve in practice and in fact may not give the desired result. This is because the aggregation effect also depends on covariances (not just variances) and because the ecological correlation (for example) compounds individual-level and area-level influences in measuring the relationship. Holt et al. (1996) and Tranmer and Steel (1998, p. 824) are more specific and suggest forming regionalizations that maximize the pure area-level correlation since, they argue, it is this quantity that is of interest geographically rather than obtaining estimates of individual-level relationships. Their work has highlighted the underlying complexity of the relationships responsible for the MAUP and why it is 'not amenable to simple attempts to unravel it' (Holt et al., 1996, p. 198). It is important to analyse the MAUP in the context of a well-defined model that identifies the different scale components (from the individual-level upwards) underlying the behaviour of each of the variables.

4.3 Data consistency and spatial data analysis

Consistency checks are essential to ensure that data values do not fall outside permitted ranges, such as percentages that must lie in the 0% to 100% range or measures of dispersion or distances which must be positive valued. Problems can arise when merging spatial units or moving to a common

spatial framework using interpolation methods. Counts can be summed but in computing new percentages or averages then the analyst should return to the original data. Consistency checks are needed to ensure that error is not introduced into a database as a consequence of undertaking inappropriate or inaccurate manipulations of the data. GIS software for example does not necessarily provide warnings on when inappropriate spatial operations have been performed (Mrozinski and Cromley, 1999, p. 288). Many forms of statistical analysis have to be performed outside a GIS and inconsistencies can enter a database as a result of errors in transferring files, or from creating several copies of a file that may then, inadvertantly, undergo different revisions and updating.

The most important context for carrying out consistency checks is when different databases have to be merged or synchronized particularly if those databases have been collected by different agencies. The problems are likely to be especially acute when the data sets have not been collected at exactly the same scale. When merging data sets that refer to different time periods the analyst needs to be aware of this, report the differences in time period, and consider the possible implications for interpreting findings. Population data derived from a census may provide a poor measure of the appropriate denominator for computing an incidence rate when the health data refer to an intercensus period. In addition to ensuring consistency in data attributes, it is also necessary to ensure consistency when merging spatial objects so that, for example, houses are not located in the middle of bodies of water (Kainz, 1995).

Data inconsistency is a form of data error but is considered apart from data error. Inconsistency errors may be subtle or severe, but in theory, at least, can be avoided by carrying out the appropriate checks on the database both during and after carrying out data operations.

4.4 Data completeness and spatial data analysis

The distinction was drawn in section 2.3 between model and data completeness. If a database is not *model* complete then not all the important variables needed for statistical analysis are available in the database. This can lead to model misspecification resulting in biased estimates of the parameters of those variables that are measured accurately and correctly included in the model. Failure to include a significant variable (say, X_k) means that the estimate of the regression parameter for a variable of interest (say, X_1) on the response variable (Y) measures more than the *direct* effect of X_1 on Y. It also measures all the *indirect* effects associated with the influence of X_1 on Y through X_1's relationship to X_k (see figure 4.5). In practice it is impossible to be sure that all relevant

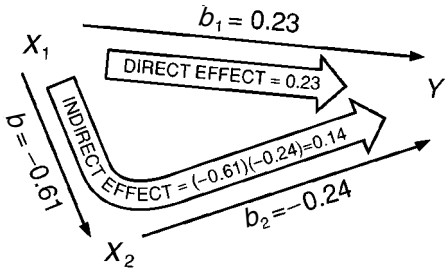


Figure 4.5 Effect of an omitted variable on parameter estimates

factors have been identified but the most important ones should be identified and included in the model. The alternative, not often feasible in observational science, is to randomize the level of X_1 over all cases since this removes the bias of missing variables.

Sections 4.4.1 and 4.4.2 examine methods for handling two aspects of spatial data incompleteness. Section 4.4.1 examines the missing-data problem. The term 'missing data' refers to those situations where there is a true but unrecorded value for a case. Typically the term 'missing data' and the methods of section 4.4.1 are applied to data sets that refer to areas but may be adapted to the case of data that refer to point locations on a spatially continuous surface. The missing-data problem is distinguished here from spatial interpolation and spatial prediction problems. The latter problems arise where data have been recorded at a number of locations on a spatially continuous surface and the analyst now wants to provide an estimate or prediction at one or more *other* locations on the same surface. Methods for spatial interpolation and spatial prediction are the subject matter of section 4.4.2. There is however overlap between the methods of sections 4.4.1 and 4.4.2.

Throughout these sections we refer to either 'predicting' or 'estimating' values. The term estimation is used in statistics when deciding on the value of a parameter which is assumed to have some unknown single value. If it is assumed that there is a single true data value then we speak of *estimating* the value. However the process of measurement or the nature of the underlying mechanism generating the data may suggest thinking of a data value as one of a possible distribution of values from some probability model (section 2.1.4). In this case it is usual to speak of *predicting* a data value and attaching a measure of the prediction error that reflects (i) the inherent variability arising from sampling the *probability* model together with (ii) any *statistical* uncertainty that derives from estimating the unknown parameters of the probability model. There are parallels here with design-based and model-based approaches to spatial sampling (see chapter 3 and in particular section 3.2).

The emphasis throughout the rest of this section is on using local information to make estimates or predictions. However there are circumstances where criteria other than spatial nearness may be introduced to weight observations. There may be a substantive basis for discriminating between near values. In soil surveys the unknown value for a site might be taken to be the mean value of the relevant attribute computed from those nearest values that are in the same soil class as the unknown value.

4.4.1 The missing-data problem

Interest in missing data may originate because the analyst wishes to complete the record and therefore would like to obtain plausible values for the missing data. Data missing at one level of recording leads to an undercount when the datum forms part of a figure for a larger unit of which it is a part. To handle these problems the analyst might want good estimates or predictions of specific missing values on a variable, and an important consideration is whether there are other data that might be used to help provide these values. A second area of interest stems from the need to carry out a statistical analysis such as fitting a regression model and there are gaps in the multivariate data set. Discarding cases with one or more missing values could lead to throwing away a great deal of useful data particularly if missing values are scattered across the data matrix. The analyst is not interested in predicting the missing values per se but rather wants to be able to make valid inferences about the entire target population not just a subset for which the data happens to be complete, making full use of the data that are available.

The mechanism or process responsible for data being missing is important. A missing-data *mechanism* is said to be ignorable in the case of a single variable if the missing data are missing at random. This means the missing observations are a random sample of the sampled units. If the probability of a value being observed on a variable depends on its value (for example all the largest or smallest values have been suppressed) then the missing-data mechanism is not ignorable. Any analysis that does not allow for this will be subject to bias (Little and Rubin, 1987, p. 10). Where there are two variables (Y and X), but only one (Y) is subject to missing values, then the mechanism is ignorable for likelihood- or model-based methods. This holds even if the observed data on Y are not a random sample of the sampled units, providing they are independent of Y and they are random samples of the sampled values within subclasses defined by values of X. If the observed values are independent of both X and Y then the missing-data mechanism is ignorable for sampling-based methods as well. Little and Rubin (1987, pp. 14–18) discuss these issues and give examples.

Most methods for dealing with data sets with missing values assume the missing-data mechanism is ignorable. Some methods are available for non-ignorable missing-data mechanisms, the most direct involving follow-up surveys to try to get some of the information needed (Little and Rubin, 1987, pp. 259–64). In the case of completing Census data at the small-area level this involves follow-up work in each tract and this is expensive.

Spatial data raise some additional issues and we consider three. First, 'missing at random' does not necessarily imply that missing values must be geographically distributed 'at random' but it is an additional consideration. If observations have been deleted at regular intervals this could raise problems if an important scale of spatial variation coincides with that interval. If data values are missing from one area of a study region resulting in a sparse coverage there, or if there are clusters of missing data then the remaining data values will have a large influence on the fit of any model used to describe surface variation. Furthermore, prediction errors will vary over the map. Unwin and Wrigley (1987) illustrate this point with respect to trend surface modelling using the leverage measure (Belsley et al., 1980). Small errors in the remaining data values in the sparsely covered subregion may have a disproportionate influence on the shape and fit of the surface relative to similar-sized errors in data values in areas with a denser coverage. Higher levels of non-response are often clustered in inner-city areas in the case of census data and crime data, and the underlying mechanisms are often non-ignorable because they are linked to poverty or crime levels which are attributes that the surveys may be seeking to measure. In the case of remotely sensed data, sensor failure along a scan line is not usually related to the underlying surface so such a linear structure of missing values need not necessarily violate the missing-at-random assumption. If unemployment data are not recorded for a group of adjacent areas because of local strike action this does not necessarily imply a non-ignorable missing-data problem. These examples also illustrate the need to consider why values are missing before deciding on the approach to take.

The second issue is that, whilst some forms of spatial data (such as Census attributes) might display considerable levels of spatial heterogeneity over quite short distances, there is an underlying continuity even in these data that can be exploited to estimate missing data. More generally the presence of spatial correlation in attribute values means that neighbouring attribute values provide an information source for missing-data prediction. This may also allow some of the difficulties created by the first issue to be overcome by drawing on local subsets of the data close to the area that contains a spatial grouping of missing data.

Finally, estimating or predicting missing values, or identifying values at locations where measurements have not been taken, often assumes particular importance in spatial analysis because the analyst wishes to provide a map of the spatial variability of some characteristic. Although some warning would need to be attached to such areas of the map (such as an additional map showing an estimate of prediction error or sampling density) this might still be preferable to leaving areas blank.

As part of an exploratory investigation of the data, cases with missing values should be mapped to check for spatial bias and also examined to see they are not different with respect to those attributes for which data are available. This will provide some evidence as to whether the missing-data mechanism is ignorable or not. Unwin et al. (1996) have developed the software system MANET that displays properties of datapoints with missing values. In MANET, missing-data information is provided in chart form for each variable to show the proportion of missing values. MANET also adds an extra 'missing data' bar on a histogram, includes missing data as an extra category in a mosaic plot (Friendly, 1995) and in a scatterplot represents cases with missing-data values on either X or Y as a projection on to the Y or X axes respectively of the scatterplot.

(a) Approaches to analysis when data are missing

Approaches to the analysis of data sets with missing values fall broadly into three types. We illustrate with reference to the case where the data are needed for regression modelling. In the first approach the analyst uses only those data records that are complete. In regression analysis this means discarding any case where the response value is unknown and/or one or more explanatory variables have missing values. With small amounts of missing data this may be a reasonable strategy. However, the presence of missing values in a data matrix can have effects out of proportion to the number of missing values. For example if missing values are scattered randomly through a data matrix and if every case with at least one missing value has to be discarded from an analysis, a relatively small proportion of missing values can lead to the discarding of a large amount of valid data. This leads to inflated variance estimates, a loss of power in hypothesis testing and a loss of precision in deriving confidence intervals. If the analyst intends to fit several different regression models to the data and the pattern of missing data means that different subsets are used for different models, this may create problems of comparability.

The second and third approaches produce *single* quantities that provide an estimate or prediction of the missing value and, in the case where the missing value is a drawing from an underlying probability model that has been specified, an estimate of its prediction error. These approaches can be used whether

the analyst's interest is in the missing values themselves or in further statistical analysis.

The second approach is based on *imputation* in which the missing data are predicted, the data matrix is filled in and then analysis proceeds by standard methods. Four of the methods of imputation described by Little and Rubin (1987, pp. 43–7, 60–7) are relevant here. *Mean imputation* substitutes the mean of the responding units for the missing values. *Hot deck imputation* involves a variety of different practices based on constructing an empirical distribution of values usually based on the set of observed values. The substitute for each missing value is drawn from this empirical distribution. In the case of *regression imputation*, a regression model for the variable with missing values is estimated based on data cases with a full set of observed values (for the response and all the explanatory variables). Each missing value is then predicted using the model by substituting values for the observed variables into the prediction equation values (see, e.g., Buck, 1960). *Stochastic regression imputation* adds random noise to each imputed value to reflect the uncertainty associated with each prediction. To reflect the regression-based nature of the approach the noise might be drawn from a normal distribution with a mean of zero and a variance estimated from the regression residuals.

This approach does not draw on any underlying model of the data and these methods are only suited to the situation where the missing-data problem is 'sufficiently minor' (Dempster and Rubin, 1983). Inference in the regression model is based on the number of degrees of freedom (the amount of independent information available in the data set) which is a function of sample size (n). However, missing-data values have been imputed using other data in the database so cannot be considered as data in the sense of providing independent observations. Inferences about variables that have a complete data record can be based on n, but inferences relating to variables where incomplete records have been used in the estimation (such as the regression parameters) should be based on $(n - m)$ where m is the number of cases with missing values.

There are two extended forms of imputation: multiple and iterative imputation. The *single* imputation methods, described above, aim to arrive at a single valid inference of the missing value which then allows standard complete-data methods to be used. *Multiple* imputation constructs M 'possible' data matrices. These M data matrices, that may be obtained by bootstrapping for example, can be used to fit the regression model M times from which the variability associated with parameter estimates can be computed (Little and Rubin, 1987, pp. 255–9). A claimed advantage of multiple imputation is that it can reflect sampling variability if only one model is specified for the missing data, and it can reflect the variability associated with choice of model if more than one

model is specified. This is sometimes cited as an advantage of multiple imputation in the case of non-ignorable missing-data mechanisms which can arise with Census data, because it means the analyst can assess the sensitivity of findings to different assumptions about the mechanism. Under any one model, variability can be decomposed into average within-imputation variance and between-imputation variance (Little and Rubin, 1987, pp. 256–7).

Iterative imputation involves replacing missing values with predicted values, estimating parameters using the filled data matrix, obtaining a new set of predicted missing values assuming the parameter estimates are correct, re-estimating parameters and so on until convergence. This extended form of imputation provides the background to the third, *model- or distribution-based*, approach to missing-data prediction. This method specifies a model for the full data (observed and missing data) and uses the EM algorithm to fit the model and 'estimate' missing values. Little and Rubin (1987, pp. 129–30) discuss the background and development of the EM algorithm which generalizes and formalizes the iterative imputation approach to handling missing-data problems (see also Orchard and Woodbury's (1972) 'missing information principle'). The *M* step is a maximum likelihood estimation of the parameters given the completed – observed plus imputed – data set. The E step finds the conditional expectation of the 'missing data' given the observed data and the current estimates of the parameters and then substitutes these for the 'missing data'. It is this step that distinguishes the EM approach from iterative imputation. The 'missing-data' values are not actual 'estimates' of the individual missing values but rather functions of the missing values which are the missing sufficient statistics.

Little and Rubin (1987, pp. 152–7) describe different versions of this procedure for regression modelling. Values may be missing from the response variable (Y) and/or from one or more of the explanatory variables. In the case where it is only the response that suffers from missing data, the E step involves finding an expected value given the current estimates of the regression parameters and the corresponding levels of the explanatory variables. So if $y(i)$ is missing and $\mathbf{y}_{\text{obs}}$ is the vector of observed values on the response variable, at the tth iteration of the algorithm, the E step gives:

$$E\left[y(i) \mid \mathbf{X}, \mathbf{y}_{\text{obs}}, \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\sigma}}^{2(t)}\right] = \hat{\boldsymbol{\beta}}^{(t)}\mathbf{X}(i)$$

where $E[\,|\,]$ denotes conditional expectation, $\mathbf{X}$ is the matrix of data values on the explanatory variables (including the constant term) and $\mathbf{X}(i)$ is the (column) vector of values on the explanatory variables for case i. The row vector $\hat{\boldsymbol{\beta}}^{(t)}$ denotes the current estimate of the intercept and slope parameters and $\hat{\boldsymbol{\sigma}}^{2(t)}$ the current estimate of the error variance. Each are based on the updated data matrix using the values of the 'missing values' from the previous iteration.

(b) Approaches to analysis when spatial data are missing

The problem addressed here is as follows. Data on k variables have been collected for n areas. Some cells of the $n \times k$ data matrix are empty. In order to undertake further analysis missing values need to be estimated to complete the data matrix. Alternatively the analyst wants to fit a model making the best use of the available data.

Each of the methods of imputation reviewed in the preceding section have spatial analogues in the sense that there is an equivalent methodology which utilizes the extra information provided by the spatial distribution of values. However there may be particular circumstances to take account of with spatial data. Values may be missing for some spatial units but the sum across all the areas is known or even for subgroups of areas. County-level data values are missing or suppressed but state-level totals are known. In these cases missing value imputation should preserve 'volume' by, for example, scaling the imputed values. Since the spatial distribution of values will be utilized, what are the implications if missing-data values occur in geographical clusters?

(i) Spatial mean imputation with equal weights assigned to each included data value. These methods substitute the missing value with the arithmetic mean of the data values within some search neighbourhood or spatial window defined around the area with the missing value. A robust version of this would be to use the median rather than the mean but in either case the analyst needs to consider the size of the window. The areas that share a common border with the area with the missing value could be used (Bernstein et al., 1984).

The choice of window size is clearly a matter of importance – whether to just use the adjacent neighbours or to include higher-order neighbours. There may be no strong reason to prefer one sized window over another, other than the fact that larger windows have a greater smoothing effect. There may be advantages in taking several and examining how missing-data imputations are affected.

(ii) Spatial mean imputation with unequal weights assigned to each included data value. The previous methods are vulnerable to various clustering effects in the distribution of irregular shaped areas. To avoid some of these effects a weighted mean can be employed where the weight $a(i, j)$ might be the proportion of the total border of area i occupied by area j (Kennedy and Tobler, 1983; Tobler and Kennedy, 1985). Thus if $y(i)$ is missing and $N(i)$ denotes the set of neighbours of i:

$$\hat{y}(i) = \left[\Sigma_{j \in N(i)}\, a(i, j)\, y(j)\right] / \Sigma_{j \in N(i)}\, a(i, j) \tag{4.28}$$

There are problems with an estimating equation like (4.28). Some large nearby areas may only share a narrow border with the area with the missing

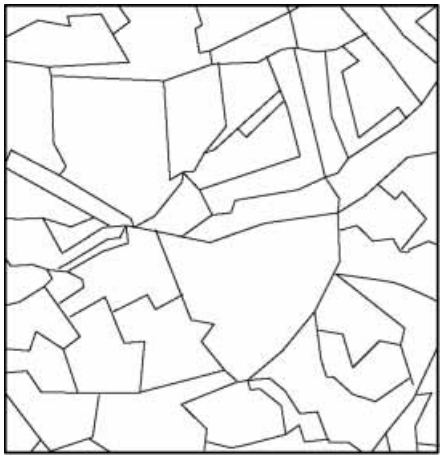


Figure 4.6 ED boundaries from an area of Sheffield to illustrate the range of different types of boundary adjacencies

value or even be 'hidden', whilst some smaller areas may occupy a large proportion of the boundary (see figure 4.6). Missing-data prediction using these methods appears to perform best when the prediction equation concentrates on the near neighbours and when spatial autocorrelation is strong (Upton, 1985). Spatial heterogeneity in the form of a spatially varying autocorrelation structure can also undermine the performance of these methods unless allowance is made by using weights that vary across the map to reflect that heterogeneity. Upton (1991) draws attention to the need to recognize that treating all rates and ratios as equally robust, even though precision may be a function of the population size, can be a source of missing-data imputation error. He recommends incorporating a measure of area size into the estimating equation in order to reflect different levels of precision.

Missing data on an areal partition raises the problem of what to do if the missing values are themselves clustered so that values are not available at some (or all) near neighbours of the partition. In Kennedy and Tobler (1983) averaging methods are adapted to this situation through a process of iteration (see also Tobler and Lau, 1978). A weighted least squares function of the difference between area values (observed and missing) is minimized where the weights reflect the lengths of shared common boundaries.

An equation such as (4.28) has the property that missing-data values cannot exceed the maximum value nor be less than the minimum value of the set of observed values so missing-data predictions cannot follow gradients. A solution to this is to include an estimate of trend (linear, quadratic or higher) in $\hat{y}(i)$. This trend could be extracted first for the whole data set (see chapter 9). An estimator such as (4.28) is then applied to the residuals from

the trend and then the trend re-introduced at the end. Ripley (1981, p. 37) describes the method of distance weighted least squares of Pelto et al. (1968) and McLain (1974) which fit linear or higher-order trends to individual points. The methodology is the univariate version of what in the geographical literature has been referred to as geographically weighted regression (Fotheringham et al., 2000).

Before leaving these methods note that if each area in the partition is represented by its area or population-weighted centroid then missing-data imputation could proceed using the methods to be described below in section 4.4.2.

(iii) Spatial hot deck imputation. This approach constructs the empirical distribution from a spatial window of values. Instead of computing the mean as in spatial mean imputation a value is drawn from the empirical distribution. If imputation is required for households in a Census tract, the empirical distribution is constructed from the set of all observed values in the same Census tract and then particular values imputed by sampling this distribution. Values can be drawn for each missing value by sampling from the set of observed values without going through the process of fitting a distribution to the observed values. This approach is likely to be attractive if there is local spatial heterogeneity so that it is safer to use data values from a local area within the Census tract close to the missing value rather than the entire tract. An extreme form of this type of approach is simply to substitute the nearest neighbour observed value for the missing value as described above. Such a method does not qualify as hot deck imputation since there is no sampling (Little and Rubin, 1987, p. 60 call this '*substitution*') but illustrates how different methods merge into one another. Sampling from the observed data values again constrains the imputed value not to be more than the local maximum value nor less than the local minimum. This is not a constraint if samples are from a distribution fitted to the data. Hot deck methods in common with all methods that impute on the basis of taking a single observed value may produce discontinuities in the spatial distribution of values. Such discontinuities are an artefact of the method not necessarily any real property of the spatial distribution of values.

(iv) Spatial regression imputation. This approach extends regression imputation as described by Buck (1960) to the case where neighbouring values of the variable (X) without missing values appear in the model specification. For example a model of the form:

$$Y(i) = \beta_0 + \beta_1 X(i) + \beta_2 \mathbf{W}^*\mathbf{X}(i) + \mathrm{e}(i)$$

is fit to the data cases with complete records and then used for missing-data prediction on Y. The variable $\mathbf{W}^*\mathbf{X}(i)$ is a spatial average of $X(i)$ as defined in

section 2.4 and the approach can be generalized to include more predictors and more spatial average terms. This type of spatial regression model will be discussed in chapter 9. It does not raise special parameter estimation problems, unlike some forms of spatial regression model. Once the data values on the variable $\mathbf{W}^*\mathbf{X}$ have been constructed the model can be fit by least squares using standard statistical software.

(v) A maximum likelihood approach. Martin (1984, 1989), Haining et al. (1984) and Griffith et al. (1989) developed a maximum likelihood approach to the problem of missing spatial data. The approach involves the iterative estimation of model parameters and *prediction* of missing values and is similar to the EM algorithm. The approach has similarities with the method of *simple* kriging, in that the weights are not required to sum to 1, and also *universal* kriging, in the sense that the mean is not assumed known and the prediction equations have similarities. Kriging is discussed in section 4.4.2. The full approach is described and then an approximation suggested which is similar to proposals found in geostatistics.

Suppose a region has n areas and observations are missing on a variable Y on k of these. Y is assumed multivariate normal with mean $\mathbf{A\theta}$ and variance–covariance matrix $\sigma^2\mathbf{V}$. The elements of the matrix $\mathbf{A}$ are the co-ordinates and powers of those co-ordinates for each datapoint for a pre-specified order of trend surface. So there are n rows and as many columns as necessary for the order of trend surface. If the trend surface is linear, $\mathbf{A}$ has three columns, if quadratic, six, and so on (see section 9.1.1 for details). The vector $\boldsymbol{\theta}$ denotes the parameters of the trend surface to be estimated. The scale parameter σ^2 is to be estimated and $\mathbf{V}$ depends on a set of unknown 'spatial' parameters that also need to be estimated. Permissable models for $\mathbf{V}$ will be discussed in chapter 9.

The column vector of observed values $\mathbf{y}$ may be partitioned, after suitable permutation so that $\mathbf{y}^T = [\mathbf{y}_o^T \mid \mathbf{y}_m^T]$ where $\mathbf{y}_o$ denotes the $(n-k)$ dimensional column vector of observed values and $\mathbf{y}_m$ denotes the k dimensional column vector of missing values.

Let the matrix $\mathbf{V}$ be partitioned after permutation so that:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{oo} & \mathbf{V}_{om} \\ \mathbf{V}_{mo} & \mathbf{V}_{mm} \end{bmatrix} \qquad \mathbf{V}^{-1} = \begin{bmatrix} \mathbf{V}^{oo} & \mathbf{V}^{om} \\ \mathbf{V}^{mo} & \mathbf{V}^{mm} \end{bmatrix}$$

The function to be minimized which is minus twice the log likelihood function is (Martin, 1984):

$$\begin{aligned} \ln L(\boldsymbol{\theta},\boldsymbol{\sigma},\mathbf{V},\mathbf{y}_m \mid \mathbf{y}_o) = {} & (n-k)\ln(2\pi) + (n-k)\ln(\sigma^2) + \ln(|\mathbf{V}_{oo}|) \\ & + \sigma^{-2}(\mathbf{y}-\mathbf{A\theta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{A\theta}) \end{aligned} \tag{4.29}$$

Note that an alternative expression for $|\mathbf{V}_{oo}|$ is $|\mathbf{V}^{mm}|/|\mathbf{V}^{-1}|$. The predictor for the set of missing values is:

$$\hat{\mathbf{y}}_m = \mathbf{A}_m\hat{\boldsymbol{\theta}} + \hat{\mathbf{V}}_{mo}(\hat{\mathbf{V}}_{oo})^{-1}(\mathbf{y}_o - \mathbf{A}_o\hat{\boldsymbol{\theta}}) \tag{4.30}$$

where $\mathbf{A}_m$ and $\mathbf{A}_o$ are submatrices of $\mathbf{A}$ referring to the missing and observed segments of the data respectively. An alternative expression for $\mathbf{V}_{mo}\,(\mathbf{V}_{oo})^{-1}$ is $-(\mathbf{V}^{mm})^{-1}\,\mathbf{V}^{mo}$.

The maximum likelihood estimators for the unknown parameters are:

$$\hat{\boldsymbol{\theta}} = (\mathbf{A}^T\hat{\mathbf{V}}^{-1}\mathbf{A})^{-1}(\mathbf{A}^T\hat{\mathbf{V}}^{-1}\mathbf{y}) \tag{4.31}$$

$$\hat{\sigma}^2 = (\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\theta}})^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\theta}})/(n-k) \tag{4.32}$$

where $\mathbf{y}^T = [\mathbf{y}_o^T \mid \hat{\mathbf{y}}_m^T]$. The parameters of $\mathbf{V}$ are obtained by minimizing:

$$|\mathbf{V}_{oo}|^{1/(n-k)}(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\theta}})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\theta}}) \tag{4.33}$$

At the first iteration the missing values $\mathbf{y}_m$ might be computed using one of the methods of imputation discussed earlier. Setting $\mathbf{V} = \mathbf{I}$ the parameter vector $\boldsymbol{\theta}$ is estimated and then the parameters of $\mathbf{V}$ and σ^2. At the second iteration the missing values are predicted using (4.30), parameters are re-estimated and the cycle continues until convergence – which can be slow if there are many missing values.

The matrix of prediction errors for the missing values is given by:

$$\boldsymbol{\Psi}_{m,m} = \hat{\mathbf{V}}_{mm} - \hat{\mathbf{V}}_{mo}(\hat{\mathbf{V}}_{oo})^{-1}\hat{\mathbf{V}}_{om} \tag{4.34}$$

and the (95%) prediction interval for the ith missing value $y_{m(i)}$ is:

$$\hat{y}_{m(i)} \pm 1.96[\boldsymbol{\Psi}_{m,m}(i,i)]^{1/2} \tag{4.35}$$

where $\Psi_{m,m}(i,i)$ denotes the ith diagonal entry in the matrix $\boldsymbol{\Psi}_{m,m}$ corresponding to $\hat{y}_{m(i)}$. Note that the prediction interval for a missing value will tend to be an underestimate because it ignores the statistical error associated with the estimation of the unknown parameters.

There are complications with the methodology. If the mean is not constant the order of trend surface has to be specified. $\mathbf{V}$ needs to be modelled to ensure a valid covariance function. In practice therefore there are additional specification issues in implementing this approach and there is a further complication which is a circularity problem in estimating $\mathbf{A}\boldsymbol{\theta}$ and $\mathbf{V}$ (Cressie, 1991, pp. 165–9).

The maximum likelihood predictor (4.30) uses weights defined by $\mathbf{V}_{mo}\,(\mathbf{V}_{oo})^{-1}$. Because of the similarities with kriging, discussion of these weights is deferred until section 4.4.2(v). Martin (1984, 1989), Krug and

Martin (1990) and results in Haining et al. (1989) investigate the effect of the pattern of missing data in an areal system. Whether the pattern is scattered or clustered and if clustered the form of the configuration has an effect on the estimation of model parameters and the performance of missing-data prediction. The proximity of the missing data to the study area boundary can also affect parameter estimation and hence missing-data prediction.

If the mean $\mathbf{A\theta}$ is a constant ($\boldsymbol{\mu}$) and known, and the parameters in $\mathbf{V}$ are known then (4.30) is the best linear unbiased predictor of the missing values and the method of prediction is equivalent to *simple* kriging (Cressie, 1991, pp. 109–10). The predictor for $\mathbf{y}_m$ is a weighted sum of the observed values but weights are not constrained to sum to 1.

The full methodology is undoubtedly cumbersome. When the mean is not known there are parallels with ordinary and universal kriging except that weights do not sum to 1. Following the practice in geostatistics and suggested for example in Isaaks and Srivastava (1989, p. 532), a more practical alternative to implementing the full methodology seems to be the following. First fit a trend ($\mathbf{A\theta}$) to the data. Use the residuals from the trend surface to estimate $\mathbf{V}$ – that is obtain the spatial covariances and fit a model to the covariances. Substitute ($\mathbf{A\theta}$) and $\mathbf{V}$ into (4.30), (4.34) and (4.35) to predict the missing values and to estimate the prediction error.

Note that the maximum likelihood approach can be extended to the case where the mean $\mathbf{A\theta} = \mathbf{X\beta}$. So the mean now comprises a set of explanatory variables ($\mathbf{X}$) and their associated coefficients ($\boldsymbol{\beta}$). The model may be the regression model of interest to the analyst who wants to fit it using as much of the data as are available. However the method applies when missing values are only associated with the response variable. No methods appear to have been developed in the spatial literature to deal with missing values in the explanatory variables or the explanatory and the response variables. A possible direction would seem to be an adaptation of the method in Little and Rubin (1987, pp. 142–5, 153–5) treating the data on $(Y, X(1), \ldots, X(k))$ as samples from a joint multivariate distribution.

4.4.2 Spatial interpolation, spatial prediction

There are four requirements any method for estimating or predicting values on a spatially continuous surface should seek to satisfy. First, the method should exploit the spatial structure in the surface and give the most weight to data at locations close to the site where the prediction is needed. Second, observed values that are spatially close together duplicate information so that without some form of weighting for members of a cluster there is a danger that any estimate could be unduly influenced by measurements from one part

of the map. Third, the method should give some indication of the likely error associated with the prediction. Fourth, the method should honour the known properties of the sample and in particular any method, when used to estimate a measured value should return that value – the closer the better. These requirements also underlie some of the (univariate) approaches to the missing-data problem. Smoothing methods (methods for reducing noise in mapped data for the purpose of identifying spatial patterns in mapped data) also share some of these requirements. Smoothing methods will be discussed in chapter 7 where the methods to be described now will again be relevant.

The term spatial interpolation is sometimes used to refer to methods that provide estimates of data values where no underlying probability model is assumed. Interpolation methods tend to be based on exploiting geometric attributes of the locations where there are observed values and in a general sense satisfy the first requirement and sometimes the second as well. Spatial or geostatistical prediction, draws on a probability model for the surface. Kriging comprises a group of methods for predicting data values. They address the weaknesses associated with geometric approaches as well as providing prediction errors. Kriging has been described as 'the logical conclusion' (Webster and Oliver, 2001, p. 37) of this area of methodological development.

In section 4.4.1 observations referred to areas, and areas can be represented by points (e.g. the area centroid). In this section observations refer to point locations but it is possible to construct areas around points using Dirichlet polygons for example (see chapter 2). There is potential therefore for methods to cross over between these two types of spatial data. What is important is that methods are applied critically with an awareness of their properties, linking choice of method to valid assumptions about the nature of the data and the underlying spatial variation in the specific data set.

(i) The Dirichlet partition. The simplest method is to assign to any location the same attribute value as its nearest observed neighbour. This is equivalent to constructing a Dirichlet partition on the sample sites. The imputed value is then given by the sample value associated with the Dirichlet polygon the site falls within. This rule gives rise to discontinuity in the estimation. The surface is implicitly made up of a set of plateaux or plates that meet along lines of discontinuity. To overcome this and problems associated with clustering, the methods reviewed in section 4.4.1 under (i), (ii) and (iii) could be used.

(ii) Cell declustering. This method divides an area about the location to be interpolated into quadrants (for example). The mean is computed for each quadrant and the four means then averaged to yield the estimate. In reported

trials cell declustering did not perform well and not as well as (4.28) (Isaaks and Srivastava, 1989, pp. 241–3).

(iii) Triangulation. The method of triangulation is also a weighted interpolator. Three nearby sites to the site to be predicted are chosen after performing a Delaunay triangulation on the set of observed sites (see Isaaks and Srivastava, 1989, pp. 251–6). The three sites, which are from the nearest neighbours in the Dirichlet partition, are chosen so they form a triangle that encloses the site to be estimated. The next step is to solve for the three unknown coefficients (a, b, c) of the equation of a plane:

$$y(j) = aE(j) + bN(j) + c$$

This is done by substitution where $y(j)$ is the observed value and $E(j)$ and $N(j)$ are the easting and northing co-ordinates of point j ($j = 1, 2, 3$). So:

$$\hat{y}(i) = \hat{a}E(i) + \hat{b}N(i) + \hat{c}$$

where $\hat{a}, \hat{b}$ and $\hat{c}$ are the estimates of the coefficients. This estimator produces a spatially smoother set of values than some of the earlier methods described, but still produces abrupt changes of gradient at the margins of the triangle (Webster and Oliver, 2001, p. 39). Sibson's (1981) 'natural neighbour interpolator' which is an extension of the triangulation method also generates discontinuities (Webster and Oliver, 2001, pp. 39–40).

(iv) Inverse distance weighting. Interpolation methods that employ distance weighting give differential weights to observations based on their proximity to the missing value. Distance weighting is introduced to capture the idea that attribute values close in distance terms tend to be similar but that the similarity weakens as distance separation increases. So it is the nearest sites that should be given most weight in any imputation. The interpolated value for $y(i)$ is:

$$\hat{y}(i) = \Sigma_{j=1,\ldots,n}\, \lambda(i, j)\, y(j) \tag{4.36}$$

where for example $\lambda(i, j) = d(i, j)^{-\alpha}$, α is a positive constant and $d(i, j)$ is the distance between sites i and j. Other choices for $\lambda(i, j)$ include $\exp(-\alpha d(i, j))$ and $\exp(-\alpha d^2(i, j))$ and $\lambda(i, j)$ is often set to zero when it exceeds some chosen distance. The coefficients are scaled so that $\Sigma_{j=1,\ldots,n}\, \lambda(i, j) = 1$. The choice for α influences the contribution made by data from different distances and the smoothness properties of the surface (Ripley, 1981, p. 36). A large value for α means that distant sites play a small role in the imputation which tends to produce a very spiky interpolated map. A small value of α has the effect of giving equal weight over longer distances and gives rise to a smooth interpolated map. Isaaks and Srivastava (1989, pp. 258–9) illustrate the effect of varying

α on the contribution made by sites at different distances from the site to be estimated.

Distance-weighting methods like (4.36) are affected by clusters in the datapoints. A combination of methods might be appropriate. The cell declustering method could be employed in which the quadrant mean is attached to the centroid of the observed sites and then the imputation uses a distance weighting of the means as in (4.36).

Isaaks and Srivastava (1989, pp. 266–77) compare several of the above methods through a series of case studies. They remark: 'the method which is "best" depends on the yardstick we choose' (p. 272). Some methods, like (4.36) with a relatively large decay parameter (e.g. $\alpha = 2.0$) are to be favoured if the objective is to minimize the *largest* errors of estimation whilst triangulation methods have relatively good *average* levels of error. They suggest that incorporating more nearby sites improves estimates but that if the sites are clustered this can offset the benefits of taking more sites (p. 276). The best methods, they conclude, use all the nearby sites and account for the possibility of clustering. These properties will be present in the kriging predictors to be discussed below which utilize information on the spatial covariance properties of the data. By contrast the estimators described above depend only on geometric relationships between the sites and the estimates they yield are dependent on neighbourhood size and how the weights are specified. They do not provide an estimate of the possible error associated with the imputation.

(v) Kriging. It is now assumed that the data $y(1), \ldots, y(n)$ are the realization of a (weakly stationary) stochastic model with mean, $\mu(.)$, and (symmetric) variance–covariance matrix, $\boldsymbol{\Sigma}$. Given a sample of size n, the BLUP of any unsampled point on the surface can be obtained by *simple kriging*. To predict the attribute value at site o, (o), which is not included in the sample compute (Cressie, 1991, p. 110):

$$\tilde{y}(o) = \mu(o) + \mathbf{c}^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \tag{4.37}$$

where $\mathbf{c}^T = (\text{cov}(Y(0), Y(1)), \ldots, \text{cov}(Y(o), Y(n))$. $\boldsymbol{\Sigma}$, as noted, is an $n \times n$ symmetric matrix with (i, j)th element equal to $\text{cov}(Y(i), Y(j))$, $\mathbf{y} = (y(1), \ldots, y(n))^T$, $\boldsymbol{\mu} = (\mu(1), \ldots, \mu(n))^T$ and $\mu(o)$ is the mean evaluated at site o. The second term in (4.37) identifies the simple kriging weights, $\mathbf{c}^T\boldsymbol{\Sigma}^{-1}$, assigned to each datapoint, that yields the BLUP of the unknown attribute value.

The weights in (4.37), $\mathbf{c}^T\boldsymbol{\Sigma}^{-1}$, reflect spatial covariance properties in the data, or '*statistical* distance weighting', as Isaaks and Srivastava (1989, p. 300) call it, rather than arbitrary definitions of spatial relationships. The vector $\mathbf{c}$ introduces a form of distance weighting into the prediction, because covariances tend to decrease with increasing distance so that more remote observed

sites contribute less to the prediction. The presence of Σ incorporates the covariances between each observed data value and every other. Observed values that are close together have large values in Σ, those far apart, small values. The multiplication of $\mathbf{c}$ by Σ^{-1} 'adjusts the raw inverse statistical distance weights (in $\mathbf{c}$) to account for possible redundancies between the observed values' (Isaaks and Srivastava, 1989, p. 300). The presence of Σ^{-1} therefore handles the effect of clustering in the observed values on prediction. This is achieved by downweighting the contribution from sample sites that are members of clusters. If data clustering is not a problem (as in the case where data are from a systematic spatial sample or grid of pixels) it is not surprising that simpler distance weighting or even nearest neighbour methods that approximate the elements of $\mathbf{c}$ but do not incorporate a term corresponding to Σ^{-1} have done nearly as well in some trials (Haining et al., 1989).

The prediction error for (4.37) is:

$$\sigma_{sk}^2 = \text{Cov}(Y(o), Y(o)) - \mathbf{c}^T \Sigma^{-1} \mathbf{c} \tag{4.38}$$

The first term, $\text{Cov}(Y(o), Y(o)) = \text{Var}(Y(o))$ measures the variability of the attribute Y. The second term is the weighted sum of the covariances between the samples and the value to be predicted ($\mathbf{c}$) where the weights are given by the simple kriging weights ($\mathbf{c}^T \Sigma^{-1}$). Prediction intervals can be constructed from (4.38). In the case of 95% intervals the interval is given by $y(o) \pm 1.96\, \sigma_{sk}$. The reader should compare (4.37) and (4.38) with (4.30) and (4.34) recalling though that in section 4.4.1(b)(*v*) the mean and variance–covariance matrix are to be estimated and the notation $\Sigma = \sigma^2 \mathbf{V}$ was used.

In practice the mean is not usually known and for mapping and interpolation the weights are usually required to sum to one. This last condition ensures uniform unbiasedness (Cressie, 1991, p. 120). Assume the mean is constant and the covariance matrix Σ is known at least up to a scalar σ^2. Optimal prediction in the sense of minimizing the prediction error is achieved through *ordinary kriging*. Ordinary kriging provides the prediction in one step and does not require any explicit identification of the mean. If the mean is unknown but follows some order of trend surface (see section 9.1.1), optimal prediction is achieved through *universal kriging*. This too is a one-step prediction which does not require explicit identification of the mean although it does require specification of the order of the trend surface. The following description assumes weak stationarity. Expressions for (4.39) and (4.41) assuming only intrinsic rather than weak stationarity (see section 9.1.2(a)) and expressed in terms of the semi-variogram are given in Cressie (1991, pp. 122 and 153–4).

Ordinary and universal kriging predictors under weak stationarity are given by (Cressie, 1991, pp. 123 and 154):

$$\tilde{y}(o) = [\mathbf{c}^T \Sigma^{-1} + (\mathbf{Am})^T \Sigma^{-1}]\mathbf{y} \tag{4.39}$$

where:

$$\mathbf{m} = (\mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A})^{-1} (\mathbf{a} - \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{c}) \tag{4.40}$$

A is an $n \times p$ matrix where p denotes the number of trend surface parameters which is determined by the order (q) of the trend surface (see below section 9.1.1). The p dimensional column vector **a** identifies the corresponding spatial co-ordinates for site o. For the constant mean case (ordinary kriging) $p = 1, q = 1$; **A** is a column vector of 1s and **a** is equal to 1. For a first- or higher-order trend surface (universal kriging) the ith row of **A** refers to sample point i. For a second-order trend surface for example, $q = 2$, $p = 6$, and the ith row of **A** is:

$$(1, s_1(i), s_2(i), s_1(i)^2, s_1(i)^2, s_1(i)s_2(i))$$

where $(s_1(i), s_2(i))$ denotes the co-ordinate position of the ith point. The p dimensional column vector **m** contains the p Lagrange multipliers that ensure the weights sum to 1.

The kriging prediction error is (Cressie, 1991, pp. 123 and 155):

$$\sigma^2 = \text{Cov}(Y(o), Y(o)) - \mathbf{c}^T \boldsymbol{\Sigma}^{-1} \mathbf{c} + \mathbf{m}(\mathbf{a} - \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{c}) \tag{4.41}$$

Now 95% prediction intervals can be constructed using: $\tilde{y}(o) \pm 1.96\,\sigma$. The first two terms of (4.41) also appear in (4.38). The third term is new and represents the variance arising from the estimate of the mean, however it is usually very small (Webster and Oliver, 2001, p. 179). To further consider (4.41) we take the case of ordinary kriging ($p = 1$), then (4.41) becomes:

$$\text{Cov}(Y(o), Y(o)) - \mathbf{c}^T \boldsymbol{\Sigma}^{-1} \mathbf{c} + \mathbf{m}[1 - (\mathbf{1}^T \Sigma^{-1} \mathbf{c})] \tag{4.42}$$

where:

$$\mathbf{m} = (\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1})^{-1} [1 - \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{c}]$$

First, prediction error will increase the further site o is from sample data. If there are few nearby sample observations, prediction errors will be larger than if there are several. The closer the elements in the vector **c** are to 0 the smaller the second term in (4.42) and the larger the prediction error. This effect is also present in the third term in (4.42) where **c** again appears. However consider now the term in square brackets, $[1 - (\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{c})]$, and also the term $(\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1})^{-1}$. The term in square brackets is the difference between the sum of the simple kriging weights and 1. If the simple kriging weights are close to 1 then this term (which appears twice) will be close to 0 and so the third term in (4.42) will be close to 0. If sample points are clustered this will yield higher prediction errors than if the sample points are spread out. This is because with spatially

correlated phenomena, two sample points that are very close together are not much better than one in predicting values at a nearby site – they carry similar information for the purposes of predicting the value at any unsampled location. In the context of (4.42) this effect is measured through the sum of the covariances between all the sample pairs: $(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}$. The bigger the spatial covariance in the sample the bigger $(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}$. However its effect in (4.42) depends on the size of the discrepancy between the sum of the simple kriging weights and 1. These points are illustrated for the case of ordinary kriging ($q = 1$), using a different perspective, by Isaaks and Srivastava (1989, pp. 497–9). Webster and Oliver (2001, chapter 8) provide further illustrations. Irregularly spaced data generate patches with larger prediction errors (i.e. in areas with few sample points), than regularly spaced data for a given density of sample points and given spatial covariance structure.

Ordinary kriging appears to be a more widely employed method of spatial prediction in geostatistics than universal kriging. Spatial prediction is based on local information and within such a window the assumption of a constant mean may be appropriate. Isaaks and Srivastava (1989, p. 532) suggest an alternative to universal kriging which they suggest is more practical for geostatistical interpolation where $\boldsymbol{\Sigma}$ is usually not known. First fit a trend ($\mathbf{A}\boldsymbol{\theta}$) to the data. Next remove the trend and use the residuals from the trend to estimate $\boldsymbol{\Sigma}$. Apply the method of ordinary kriging to the residuals adding the trend back in at the end in making the final set of predictions.

There are other forms of kriging to apply in situations where the data are non-Gaussian such as categorical or count data and for which non-linear geostatistics is needed. See Journel (1983) for indicator kriging and Matheron (1976) for disjunctive kriging both of which are reviewed in Cressie (1991, pp. 278–84).

An irregular distribution of seven sites is shown in figure 4.7 together with the Dirichlet polygon for site o. It is the value at site o that is to be predicted. Tables 4.1(a)(a) to (d) give the data values and intersite distances, the upper triangular elements (including the diagonal) of the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ and the elements of the vector $\mathbf{c}$. These covariances are based on an assumed isotropic covariance model of the form:

$$C(h) = 5.0\exp(-0.2|h|)$$

In practice the full data set (of which the area shown is a subarea) would be used to estimate the spatial covariances (or semi-variogram) at various distances and then a model fitted to the plot.

Table 4.2(a) shows the results of a selection of different methods described in sections 4.4.1 and 4.4.2 for predicting the value at the site marked o. In the

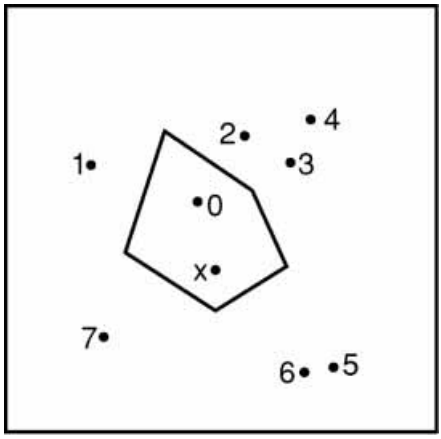


Figure 4.7 Pattern of sites for the worked examples in tables 4.1 and 4.2

Table 4.1(a) *Distances (|h|) between sites in figure 4.7; data values recorded at the seven sample sites*

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	0.0	2.0	1.5	2.0	2.5	4.0	3.7	3.0
1		0.0	2.9	3.7	4.2	5.7	5.5	3.2
2			0.0	1.0	1.3	4.5	4.5	4.6
3				0.0	0.8	3.8	3.9	4.8
4					0.0	4.6	4.7	5.5
5						0.0	0.5	4.3
6							0.0	3.7
Values		55	45	41	49	75	78	80

Table 4.1(b) *Variance–covariance matrix between the sample (1–7):* Σ

5.00	2.799	2.385	2.158	1.599	1.664	2.636
2.799	5.00	4.093	3.855	2.032	2.032	1.992
2.385	4.093	5.00	4.260	2.338	2.292	1.914
2.158	3.855	4.260	5.00	1.992	1.953	1.664
1.599	2.032	2.338	1.992	5.00	4.524	2.115
1.664	2.032	2.292	1.953	4.524	5.00	2.385
2.636	1.992	1.914	1.664	2.115	2.385	5.00

case of the kriging predictions note that the severest downweighting in Σ^{-1} to compensate for the effect of clustering is associated with sites 5 and 6 (–0.998), 3 and 4 (–0.524) and 2 and 3 (–0.365). Sites 4 and 5 both receive negative kriging weights that reflect the fact that 4 is 'hidden' by 2 and 3 and 5 is 'hidden'

Table 4.1(c) *Inverse of variance–covariance matrix* $\boldsymbol{\Sigma}$ *:* $\boldsymbol{\Sigma}^{-1}$

0.347	−0.148	−0.005	0.010	−0.010	0.013	−0.127
−0.148	0.725	−0.365	−0.180	0.011	−0.013	−0.009
−0.005	−0.365	1.002	−0.523	−0.084	−0.020	−0.015
0.010	−0.180	−0.523	0.777	−0.002	0.005	0.006
−0.010	0.011	−0.084	−0.002	1.129	−0.998	0.033
0.013	−0.013	−0.020	0.005	−0.998	1.175	−0.134
−0.127	−0.009	−0.015	0.006	0.033	−0.134	0.324

Table 4.1(d) *Covariances (***c***) between sample points and prediction site o (*$\mathbf{c}^T$*)*

	1	2	3	4	5	6	7
covariances	3.351	3.704	3.352	3.033	2.247	2.386	2.744

Table 4.1(e) *Distances and covariances (***c***) to sample points and prediction site x*

	1	2	3	4	5	6	7
distance to *x*	3.0	2.5	2.5	3.4	2.9	2.5	2.4
covariances	2.744	3.032	3.032	2.533	2.799	3.032	3.093

Table 4.2(a) *Weights associated with different interpolation methods for site o*

Site	a	b	c	d	e	f	g
1	0.143	0.2	0.24	0.250	0.17	0.19	0.16
2	0.143	0.2	0.22	0.083	0.23	0.32	0.18
3	0.143	0.2	0.16	0.083	0.17	0.19	0.16
4	0.143	0	0	0.083	0.14	0.12	0.15
5	0.143	0	0	0.125	0.09	0.05	0.11
6	0.143	0.2	0.16	0.125	0.09	0.05	0.11
7	0.143	0.2	0.22	0.250	0.11	0.08	0.13
Interpolation	60.43	59.8	59.74	64.125	56.1	52.57	58.04

Notes: **a**: Arithmetic mean. **b**: Dirichlet neighbours. **c**: Dirichlet neighbours weighted by length of shared common border. **d**: Cell declustering (N–S/E–W axes used for quadrant borders). **e**: Inverse distance weighting (4.36: $\lambda(i, j) = |h|^{-1}$). **f**: Inverse distance weighting (4.36: $\lambda(i, j) = |h|^{-2}$). **g**: Negative exponential weighting (4.36: $\lambda(i, j) = \exp(-0.2\,|h|)$.

Table 4.2(b) *Triangulation method applied to site o*
(i) Co-ordinates of sites

	0	1	2	3	6	7
E–W	35	16	44	53	55	18
N–S	43	49	55	50	12	17

(ii) Coefficient estimates and interpolated values

Triangulation sites	**a** (E–W)	**b** (N–S)	**c**	Interpolated value
(1,2,6)	−0.182	−0.814	97.816	56.416
(1,2,7)	−0.187	−0.792	96.850	56.249
Average				56.332

4.2(c) *Weights associated with simple kriging and ordinary kriging (sites o and x)*

	Simple kriging		Ordinary kriging	
Site	site (*o*)	site (*x*)	site (*o*)	site (*x*)
1	0.286	0.129	0.278	0.125
2	0.387	0.190	0.385	0.189
3	0.119	0.244	0.120	0.244
4	−0.004	−0.107	−0.013	−0.112
5	−0.041	−0.024	−0.048	−0.028
6	0.136	0.302	0.133	0.301
7	0.151	0.282	0.144	0.278
Sum	1.04	1.02	1.00	1.00

4.2(d) *Predictions and prediction errors for simple and ordinary kriging (sites o and x)*

	Simple kriging		Ordinary kriging	
	(*o*)	(*x*)	(*o*)	(*x*)
Predicted value	57.523	64.875	55.25	63.675
Prediction error	1.571	1.872	1.574	1.873
Lagrange multiplier (*m*)			−0.10	−0.05

Note: The predicted value for simple kriging is obtained from (4.37) for ordinary kriging (4.39). The prediction error for simple kriging is obtained from (4.38) for ordinary kriging (4.42).

by 6. Although both sites 1 and 3 are at equal distances from the site to be predicted the weighting for site 1 is the larger since there are no other sites close to 1. The ordinary kriging error is only slightly larger than the simple kriging prediction error because the simple kriging weights are only slightly greater than 1.0. To illustrate how prediction error is affected by the distribution of sample sites in relation to the site where the prediction is needed, a second site (x) has been evaluated. The distances (from site x to the sample sites) and the corresponding elements of the vector **c** are shown in table 4.1(e). Note the increase in the prediction error at x relative to o (table 4.2(d)). The closest neighbour of x is 2.4 units, compared to 1.5 for o, which has three neighbours less than or equal to 2.0 units away.

Apart from inverse distance weighting ($\alpha = 2$) where the weights decay very rapidly with increasing distance all the methods provide predictions that are larger than the prediction provided by ordinary kriging which should be considered the 'gold standard'. Note this remark also applies to the distance weighting method using the exponential function ($\exp(-0.2|h|)$). The cell-declustering method (based on a north–south/east–west partition of the area) gives the largest value which differs most from the 'gold standard'. Two triangulations are examined and the average of the two computed because there seems nothing to choose between them. This table of results should not be taken as indicative of which methods come closest to approximating the gold standard since with other data sets other rank orderings might arise. The purpose is to show the range of predictions arising from different methods.

Figure 4.8 summarizes the approaches to missing-data estimation and spatial interpolation and prediction.

4.4.3 Boundaries, weights matrices and data completeness

In conclusion we briefly draw attention to some other aspects of model completeness that ought to be considered for spatial data analysis. The application of some spatial statistical techniques, because they draw on data from spatial neighbourhoods around each case, may require data on variables where the data values refer to spatial units beyond the boundary of the study area. If these boundary data are not available this represents a form of data incompleteness. Figure 4.9 classifies sites. An *interior site* lies inside the study area so its value is observed as are the values for all its neighbours. An *observed boundary site* lies inside the study area so its value is observed but the values for at least one of its neighbours lies outside the study area and so has not been observed. An *unobserved boundary site* may be a neighbour of an observed boundary site but

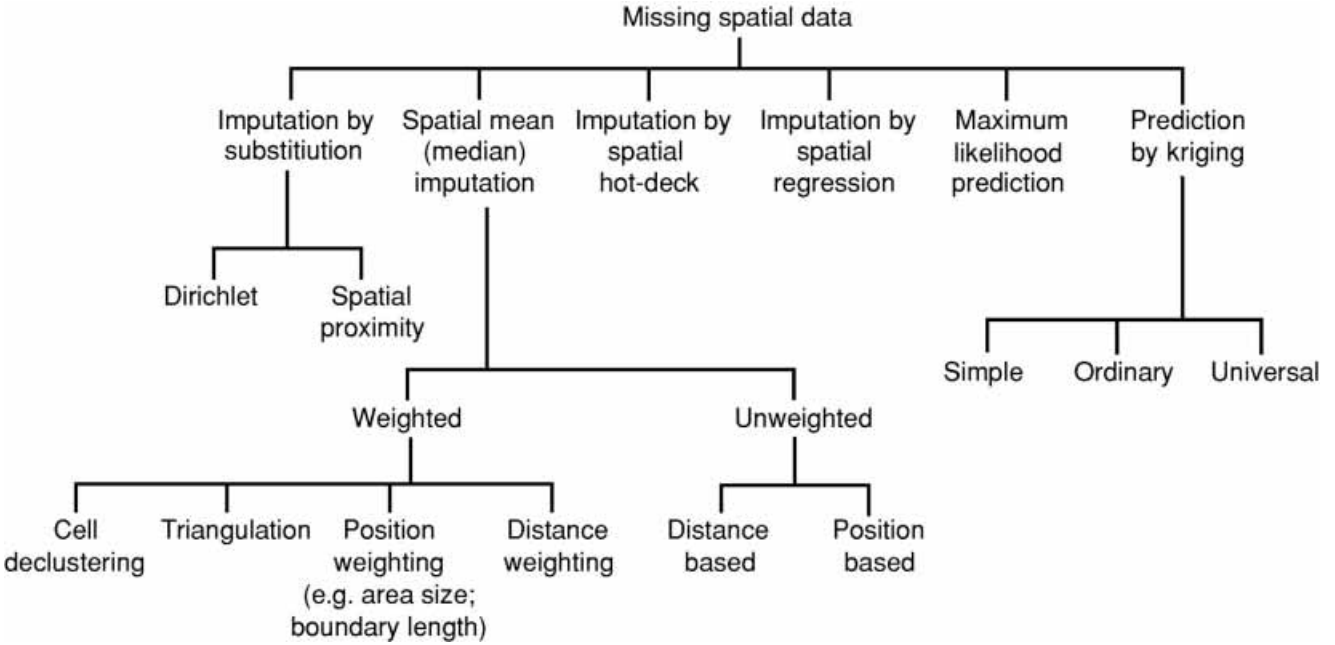


Figure 4.8 Approaches to missing-data estimation, interpolation and prediction

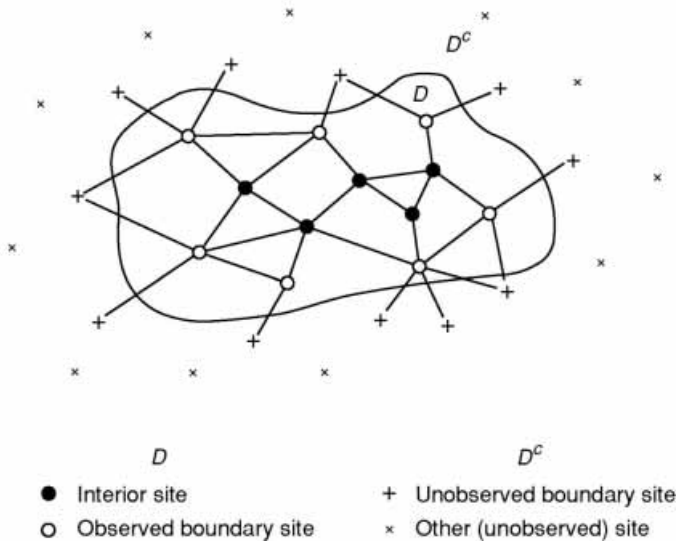


Figure 4.9 Classification of sites according to their relationship to the boundary and the definition of their neighbours

because it lies outside the study area its value has not been observed. The classification of sites depends on the graph or connectivity structure that has been assumed to represent spatial relationships between the sites (see 2.4).

Martin (1987) defines two aspects of the boundary value problem: the effect of different boundary formulations on the variance–covariance matrix ($\mathbf{V}$) of the process; the effect of boundary forms on the properties of estimators of model parameters. The effect of the boundary on the conduct of data analysis depends on how the boundary is handled and this depends on how the underlying process is conceptualized. One approach is to treat the process observed within the study region as a subset of a process that is operating over a wider area. The model covariance, for example, between an observed and unobserved boundary site is the same function of distance as between any pair of observed values and may be estimated accordingly using the data that are available. A second approach is to define one model for the interior sites and another model for the boundary. This would seem appropriate where there are different processes operating on either side of the boundary. This could arise if the boundary is natural in process terms such as in the case of an island where the coastal boundary affects the spread of some disease. It could also arise if the boundary is a political boundary with different policies or regimes operating on either side of the boundary. Let $\mathbf{y} = (y(1), \ldots, y(n))$ denote the observed values (interior and boundary sites) whilst $(y(n+1), \ldots, y(n+h))$ denotes the unobserved boundary values (D^c). Let $\mathbf{y}^* = (y(1), \ldots, y(n+h))$ and $\mathbf{V}_{\mathbf{y}^*}$ denote the variance–covariance

matrix for $\mathbf{y}^*$. The variance covariance matrix for $(y(1), \ldots, y(n))$, denoted $\mathbf{V}_\mathbf{y}$ is obtained by deleting the rows and columns associated with $(y(n+1), \ldots, y(n+h))$ in $\mathbf{V}_{\mathbf{y}^*}$. From the theory of partitioned matrices if:

$$\mathbf{V}_{\mathbf{y}^*}^{-1} = \begin{bmatrix} \mathbf{V}^{oo} & \mathbf{V}^{ob} \\ \mathbf{V}^{bo} & \mathbf{V}^{bb} \end{bmatrix}$$

where o refers to the observed sites (interior and boundary) and b refers to the unobserved boundary sites in D^c then:

$$\mathbf{V}_\mathbf{y}^{-1} = \mathbf{V}^{oo} - \mathbf{V}^{bo}(\mathbf{V}^{bb})^{-1}\mathbf{V}^{ob} = \mathbf{V}^{oo} - \boldsymbol{\Gamma}$$

(see Martin, 1987). The matrix $\boldsymbol{\Gamma} = 0$ if, for example, the process outside the region is 'disconnected' from the process inside the study region so that $\mathbf{V}^{bo} = \mathbf{V}^{ob} = 0$. Martin gives examples. Different boundary assumptions can be modelled through the specification of $\boldsymbol{\Gamma}$. The contrasting cases mirror the 'stochastic' and 'fixed' boundary value assumptions used in time series analysis with the added complication that now the boundary encircles the study region. Now boundary assumptions will influence a larger proportion of data values and boundary effects only dissipate towards the interior of the study region providing the study region has a large number of interior sites that are distant from the boundary.

The specification of connectivity relationships between sites or areas (the elements of the matrix $\mathbf{W}$) is a modelling assumption but it can be informed by relevant data on interactions between sites or areas. The objective of analysis may be to explain spatial variation in per capita income levels between regions within a country. But such regional variation is unlikely to be independent and a minimal specification will need to allow for spatial dependencies between the regions. Ideally factor flow data are needed either to build explicitly into the specification of the model or to inform the specification of the spatial relationships, described by the matrix $\mathbf{W}$. The absence of interaction data in these circumstances may be seen either as a form of model incompleteness or data incompleteness. It might be seen as a form of data incompleteness because, whilst it may be possible to specify a plausible connectivity structure from other information about the geography of the regions, the data necessary to undertake this carefully cannot be considered complete if real interaction data are not available.

4.5 Concluding remarks

This chapter has considered the implication of spatial data quality for the conduct of spatial data analysis. Many of the techniques of spatial data

Table 4.3 *The dimensions of spatial data quality in relation to stages of spatial data analysis*

	Accuracy	Resolution	Consistency	Completeness
Data collection and preparation of final database	Concerns about the presence of gross errors	Creating a common spatial framework for data collected on different frameworks	Incompatible data values (e.g. disease cases reported in areas without population)	Presence of missing data; need to interpolate or predict
Form and conduct of statistical analysis	Choice of error model. Need for robust and resistant statistical methods of analysis	Differences in variable precision across spatial units. Sensitivity of results to different methods of areal interpolation		Modelling in the presence of missing data.
Interpretation of results		Ecological bias. Forming invalid individual level hypotheses from aggregate analyses		Concerns about spatial variation in undercounting. Model misspecification due to the effects of missing variables

analysis have evolved in response to data quality issues that regularly arise in handling spatial data. The purpose of this chapter has been to show how various techniques can be ordered and classified in relation to this need. In conclusion we return to the earlier remark about how different facets of spatial data quality may impact on particular stages of spatial data analysis. Table 4.3 provides a cross-classification by data quality dimension and stage of analysis with some indicative examples.