9

Environmental Sampling Design

Dale L. Zimmerman
University of Iowa

Stephen T. Buckland
University of St Andrews, Scotland, UK

CONTENTS

9.1 Introduction ................................................................. 181
9.2 Sampling Design for Environmental Monitoring .................. 182
  9.2.1 Design framework .................................................... 182
  9.2.2 Model-based design ................................................... 183
      9.2.2.1 Covariance estimation-based criteria .......................... 183
      9.2.2.2 Prediction-based criteria ...................................... 184
      9.2.2.3 Mean estimation-based criteria ............................... 186
      9.2.2.4 Multi-objective and entropy-based criteria ................ 187
  9.2.3 Probability-based spatial design ................................... 188
      9.2.3.1 Simple random sampling ..................................... 188
      9.2.3.2 Systematic random sampling .................................. 189
      9.2.3.3 Stratified random sampling ................................... 189
      9.2.3.4 Variable probability sampling ................................. 190
  9.2.4 Space-filling designs ............................................... 190
  9.2.5 Design for multivariate data and stream networks .......... 193
  9.2.6 Space-time designs .................................................. 195
  9.2.7 Discussion ............................................................ 198
9.3 Sampling for Estimation of Abundance .............................. 198
  9.3.1 Distance sampling ................................................... 199
      9.3.1.1 Standard probability-based designs ......................... 199
      9.3.1.2 Adaptive distance sampling designs ........................ 201
      9.3.1.3 Designed distance sampling experiments ................... 203
  9.3.2 Capture-recapture ................................................... 205
      9.3.2.1 Standard capture-recapture ................................. 205
      9.3.2.2 Spatial capture-recapture .................................... 205
  9.3.3 Discussion ............................................................ 206
Bibliography ................................................................. 206

9.1 Introduction

Environmental and ecological data occur in space and time, and precisely where and when they occur may affect what can be inferred from the data. For example, if precipitation is measured at gauges widely separated in space, average precipitation amounts over large
regions may be estimated quite well but little may be inferred about the local or small-scale spatial variability of those amounts. Consequently, considerable effort has been put into the development of sampling designs, i.e., selections of spatial locations and times, that maximize or otherwise enhance the quality of inferences that can be made from sampling the environment. This chapter considers such designs. In particular, it reviews designs used to monitor an environmental variable (such as precipitation, or the ambient level of an atmospheric pollutant) over space and time, and designs used to estimate the abundance, or population size, of an organism in the environment.

Among the designs we will describe are those that are probability-based, i.e., that involve the selection of samples according to a random mechanism that can be described by a probability distribution. Other designs to be considered are not probability-based; nevertheless, they are chosen to meet certain well-defined statistical or geometric objectives. This sets them apart from other types of non-probability-based designs, such as those that arise from haphazard sampling, in which samples are chosen purely out of convenience, or judgment sampling, in which samples are chosen with a goal of being “representative” in some sense.

### 9.2 Sampling Design for Environmental Monitoring

#### 9.2.1 Design framework

We describe sampling design for environmental monitoring within the following framework for spatio-temporal data. Assume that the environmental variable of interest, $Z$, is continuous (or discrete with so many levels that it can be approximated as such) and varies over a region $S$ and a time interval $T$ of interest. Let $Z(s,t)$ denote the value of $Z$ at point site $s \in S$ and time point $t \in T$. Let $D$ denote the design space, i.e., the set of site-time pairs where it is possible to take observations on $Z$, and suppose that the sample size $n$, i.e., the number of site-time pairs at which $Z$ will be observed, is predetermined. The sampling design problem is then to select $n$ site-time pairs from $D$ in such a way as to best meet some objective or combination of objectives. A wide variety of objectives are possible, including such things as estimating a spatial average or temporal trend, detecting noncompliance with regulatory limits, determining the environmental impact of an event, measuring the effects of environmental mitigation, and detecting extreme levels (e.g., floods or smog) so that public alerts can be issued.

Although $D$ could in principle be a continuum coinciding with $S \times T$, administrative, practical, and economic considerations may restrict $D$ to a finite subset of points in $S \times T$. Even in cases where there are no reasons such as these for restricting $D$, for some approaches to sampling design it may be necessary to discretize $D$ in order to simplify the search for the best, or at least a good, design.

It is possible for the costs of sampling to vary across $D$; for example, some sites may be more expensive to travel to than others. However, variable costs are often difficult to incorporate formally into the design criteria to be reviewed here, and we will not attempt to do so. In practice, a design of one of the types described here, which are all obtained via statistical considerations, may need to be modified to meet budgetary or other non-statistical constraints.

In some situations where a good sampling design is sought, data have already been collected at some sites and times, perhaps using haphazard or judgment sampling, and the job of the designer is to select sites where additional observations should be taken (augmentation) or sites that, perhaps due to reduced budgets, must be deleted from the
network (contraction). In other situations, however, the designer has the opportunity to create the design de novo. Designs of the latter type may be constructed all at once or built in stages, using parameter estimates or other information from earlier stages to guide the selection of sites and times for subsequent stages.

Some of the sampling design approaches to be described require the use of a design criterion function and a computational algorithm to optimize that function. A variety of algorithms have been proposed. For relatively small augmentation and contraction problems, enumeration of all possible designs may be feasible. In problems too large for complete enumeration, greedy algorithms, exchange algorithms, branch and bound algorithms, simulated annealing, and gradient-based methods may be used.

In what follows, we break the design problem for environmental monitoring into two parts. The first part, spatial sampling design, deals only with where to sample; the temporal aspect is completely ignored. Sections 9.2.2, 9.2.3, and 9.2.4 in turn review model-based, probability-based, and space-filling approaches to spatial sampling design, and Section 9.2.5 describes extensions of the approaches to multivariate and stream network data. Because attention in those sections is limited to the spatial aspect of sampling design, we suppress the time index in \( Z(s,t) \) therein. The second part of the problem, space-time sampling design, considers not only where but when to sample and is the topic of Section 9.2.6. The time index will reappear in that section.

9.2.2 Model-based design

Model-based sampling design, as its name suggests, applies to situations in which a statistical model is assumed for \( Z(s) \). The model specifies that the observations on \( Z \) are a spatially incomplete sample of one realization of a random field, or stochastic process, \( \{Z(s) : s \in S\} \). Some assumptions typically are made about the first- and second-order moment structure of the random field. The mean function of the random field is assumed to exist and be linear in its parameters, i.e., \( E[Z(s)] = [x(s)]'\beta \) where \( x(s) \) is a vector of known functions of the spatial coordinates or other covariates observable at \( s \) and \( \beta \) is a vector of unknown parameters. Also, the covariance function, \( \text{cov}[Z(s), Z(u)] = C(s, u; \theta) \) is assumed to exist, where \( C(\cdot) \) is a known positive definite function of pairs of location vectors and \( \theta \) is a vector of unknown parameters; it may be assumed further that \( C(s, u; \theta) = C(\|s-u\|; \theta) \), i.e., that \( \{Z(s) - E[Z(s)] : s \in S\} \) is second-order stationary. The problem is to choose a design to optimize a criterion that measures how suitable the design is for making precise inferences about either the parameters of the model or predictions of unobserved values of \( Z(\cdot) \). The design criterion may thus depend on the object(s) of primary inferential interest, how the chosen object is estimated/predicted, and how to measure the quality of the chosen estimator/predictor.

We describe model-based design with respect to criteria focused respectively on covariance parameter estimation, spatial prediction, mean parameter estimation, and spatial entropy. Further details may be found in the book of Müller (2007). Several quantities play an important role in our description. For a given design \( D = \{s_1, \ldots, s_n\} \), let \( X \) denote the matrix whose \( r \)th row is \( [x(s_r)]' \), let \( \Sigma \) denote the matrix whose \( ij \)th element is \( C(s_i, s_j; \theta) \), and let \( z \) denote the vector with \( r \)th element \( Z(s_r) \). Each of \( X, \Sigma, \) and \( z \) depend on the design \( D \), but this is not explicitly indicated by the notation. It is assumed that both \( \Sigma \) and \( X'\Sigma^{-1}X \) are invertible, which may slightly restrict the set of candidate designs.

9.2.2.1 Covariance estimation-based criteria

Sometimes understanding the second-order spatial dependence, or covariance, structure of a model is the primary inferential objective. Even when it is not, and estimation of mean
parameters or spatial prediction of unobserved values of \( Z \) is the primary inferential goal, the success of these other goals may depend, in part, on how well the covariance parameters are estimated. So in either case it would seem appropriate to pay some attention to the quality of the design for estimating covariance parameters. Relevant early work on this topic considered design criteria for good estimation of the semivariogram, a close cousin of the covariance function. Russo (1984) considered a design to be optimal if it minimized the dispersion of intersite distances within bins used for the classical method-of-moments estimator of the semivariogram, while Warrick and Myers (1987) proposed a criterion that measured how well the intersite distribution corresponding to a design conforms to a user-specified distribution. Neither of these criteria are directly related to estimation quality of semivariogram or covariance parameters, however.

Criteria that are more relevant to estimation quality were introduced by Müller and Zimmerman (1999), Zhu and Stein (2005), and Zimmerman (2006). Müller and Zimmerman’s approach was to maximize the determinant of the information matrix of the weighted or generalized least squares estimators of semivariogram parameters. Zhu and Stein’s and Zimmerman’s approaches were similar, except that it was the determinant of the information matrix of maximum likelihood (ML) or residual maximum likelihood (REML) estimators of covariance parameters that was maximized, under an assumption of a Gaussian random field model. For example, for REML estimation, the criterion to be maximized is

\[
\phi_{CPE}(\theta; D) = |I(\theta, D)|
\]

where subscripts “CPE” denote covariance parameter estimation, the \( ij \)th element of the information matrix \( I(\theta, D) \) is given by \((1/2)\text{tr}[P(\partial \Sigma/\partial \theta_i)P(\partial \Sigma/\partial \theta_j)]\), and

\[
P = \Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}.
\]

Designs that are optimal with respect to this criterion and others like it generally consist of several small clusters or linear strands of sites. Many of the clusters lie along the periphery of \( S \), with the remainder spaced more or less evenly across \( D \). The leftmost panel in Figure 9.1 displays a design of size 5 that maximizes \( \phi_{CPE}(\cdot; \theta) \), where \( \Sigma = \Sigma^{(s0)} \) is a 5 \( \times \) 5 square grid and the underlying random field has mean 0 and isotropic exponential covariance function \( C(s, u) = 0.5\|s-u\| \). Such a design produces relatively many small lags and large lags, and fewer intermediate lags, than most other designs, which facilitates more precise estimation of covariance parameters.

One practical difficulty in choosing a single best design for covariance parameter estimation is that the optimal design (corresponding to a given estimator and a given measure of the utility of the design for optimizing the quality of that estimator) may depend on the unknown parameter vector \( \theta \). In other words, a globally optimal design may not exist, and we may have to settle for a locally optimal design (a design that is optimal for a specific value of \( \theta \) and presumably for values in a neighborhood of it). There are various ways to circumvent this difficulty. One is to use a two-stage adaptive sampling approach wherein some sites are sampled without regard to their optimality in order to provide an initial estimate of \( \theta \) and then subsequent sites are chosen to yield the locally optimal design corresponding to this estimate. Alternatively, a maximin approach chooses the design that maximizes the minimum value of \( \phi_{CPE}(D, \theta) \) across all values of \( \theta \). Still another approach is Bayesian, in which the uncertainty about \( \theta \) in the design criterion is incorporated into the model via a prior distribution, about which we will say more in the next subsection.

### 9.2.2.2 Prediction-based criteria

Let \( Z(s_0) \) denote the realized but unobserved value of \( Z \) at an arbitrary point \( s_0 \in S \). If the covariance parameter vector \( \theta \) is known, the best (in the sense of minimizing the prediction error variance) linear unbiased predictor (BLUP) of \( Z(s_0) \) is given by

\[
\hat{Z}(s_0; \theta, D) = [c + X(X'\Sigma^{-1}X)^{-1}(X_0 - X'\Sigma^{-1}c)]'\Sigma^{-1}z.
\]
FIGURE 9.1
Optimal model-based designs of size 5 for the examples described in Section 9.2.2. First (leftmost) panel, optimal design for covariance estimation; second panel, optimal design for prediction with known covariance parameters; third panel, optimal design for empirical prediction; fourth panel, optimal design for mean estimation. Candidate design points within each grid are represented by open circles, and points actually belonging to the design are represented by closed circles.

Here, \( \mathbf{c} \) is the vector whose \( i \)th element is \( C(s_i, s_0; \theta) \) and \( \mathbf{x}_0 = \mathbf{x}(s_0) \). The prediction error variance, or kriging variance, associated with \( \hat{Z}(s_0; \theta, D) \) is given by

\[
\sigma_k^2(s_0; \theta, D) = \text{var}_{\theta}[Z(s_0; \theta, D) - Z(s_0)] \\
= C(s_0, s_0; \theta) - \mathbf{c}'\Sigma^{-1}\mathbf{c} \\
+ (\mathbf{x}_0 - \mathbf{X}'\Sigma^{-1}\mathbf{c})(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}(\mathbf{x}_0 - \mathbf{X}'\Sigma^{-1}\mathbf{c}).
\]

Observe that the kriging variance depends on neither the mean parameters \( \beta \) nor the observed data \( \mathbf{z} \); it does, however, depend on the prediction site \( s_0 \), the covariance parameter \( \theta \), and the design \( D \). A common measure of the global performance of a design for prediction under a model with known covariance parameters is the maximum prediction error variance over a set \( \mathcal{P} \) where prediction is desired, i.e.,

\[
\phi_K(D; \theta) = \max_{s \in \mathcal{P}} \sigma_k^2(s; \theta, D).
\]

Here, subscript “K” stands for “kriging” and \( \mathcal{P} \) is typically either \( D \) or \( S \). If the mean function is constant, designs that minimize \( \phi_K(\cdot; \theta) \) tend to be “regular,” i.e., their sites are rather uniformly dispersed over the study region. If, instead, the mean function is planar, the sites again have good spatial coverage but tend to be denser near the periphery of the study region. In either case, the design is qualitatively much different than an optimal design for estimating covariance parameters. A design that minimizes \( \phi_K(\cdot; \theta) \) for the same scenario described in the previous subsection is shown in the second panel of Figure 9.1.

Although designs that minimize \( \phi_K(\cdot) \) are of interest in their own right, they do not necessarily perform well for the prediction problem of greatest practical importance, which is to predict unobserved values of \( Z \) when the spatial covariance parameters are unknown. The standard predictor in this situation, known as the empirical BLUP (E-BLUP), is given by an expression identical to the BLUP when \( \theta \) is known but with the covariance function evaluated at an estimate \( \hat{\theta} \), rather than at the hitherto-assumed-known \( \theta \). That is, the E-BLUP at \( s_0 \) is given by \( \hat{Z}(s_0; \hat{\theta}, D) \). The E-BLUP is an unbiased predictor under fairly general conditions, but an exact expression for its prediction error variance, \( \sigma_{E, K}^2(s_0; D) \), is unknown (except in very special cases) and simulation-based approaches to approximate it.
at more than a few sites are computationally prohibitive. Harville and Jeske (1992) proposed to approximate it by

\[ \sigma_{EK}^2(s_0; \theta, D) \approx \sigma_{ME}^2(s_0; \theta, D) + \text{tr}[A(s_0; \theta, D)\mathbf{I}^{-1}(\theta, D)] \]

where “EK” denotes “empirical kriging” and \( A(s_0; \theta, D) = \text{var}_\theta(\partial \hat{Z}(s_0; \theta, D) / \partial \theta) \). Accordingly, for optimal design for prediction with estimated covariance parameters, Zimmerman (2006) proposed to minimize the criterion

\[ \phi_{EK}(D; \theta) = \max_{s \in \mathcal{P}} \sigma_{EK}^2(s; \theta, D), \]

which combines a measure of design quality for prediction under a model having known covariance parameters (the first term in Harville and Jeske’s approximation) with a measure of quality for covariance parameter estimation (the second term). Not surprisingly, then, designs that are locally optimal for empirical kriging represent a compromise between designs that are optimal for each of its component objectives. That is, they often have relatively good spatial coverage while also having a few clusters. A design that minimizes \( \phi_{EK}(\cdot; \theta) \) for the aforementioned scenario is displayed in the third panel of Figure 9.1. How good the spatial coverage is in this particular design is debatable.

The prediction-based design approach based on \( \phi_{EK}(\cdot) \) is focused on good empirical point prediction. Zhu and Stein (2006) carried this approach further by constructing a criterion that gives some attention to the utility of the design for interval prediction in addition to point prediction. The spatial configuration of sites in designs that are good with respect to this extended criterion do not appear to be markedly different from designs that are good for empirical point prediction, however.

A more natural way to account for covariance parameter uncertainty when constructing efficient designs for spatial prediction is a Bayesian approach, as expounded by Diggle and Lophaven (2006). They put a prior distribution on the unknown parameters and then minimized an estimate of the spatially averaged prediction variance obtained by Monte Carlo sampling from the posterior distribution for \( \theta \) and hence from the predictive distribution of the (assumed) Gaussian random field. The designs obtained in this fashion for the examples they considered are qualitatively very similar to those obtained by the empirical prediction-based approaches of Zimmerman (2006) and Zhu and Stein (2006).

### 9.2.2.3 Mean estimation-based criteria

Suppose that greatest interest lies not in covariance parameter estimation or prediction but in estimating the model’s mean parameters, \( \beta \), as well as possible. Although a well-developed classical theory exists for exact optimization of an experimental design for the estimation of mean (regression function) parameters (e.g., Federov, 1972; Silvey, 1980), among the assumptions on which it is based are independence of the observations and the possibility of sampling repeatedly at any combination of the regressors. These assumptions are not met in the spatial sampling design problem, so attempts to apply the theory to this problem have faced significant challenges and are still evolving. We will describe instead an approximate approach to optimal design for mean parameter estimation that was studied by Müller (2005) for use when the covariance parameters are known, and a natural extension of it to the case of unknown covariance parameters.

If \( \theta \) is known, the best (in the sense of minimizing the estimator’s variance) linear unbiased estimator (BLUE) of \( \beta \) is \( \hat{\beta} = (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{z} \). A standard measure of a design’s quality for estimating \( \beta \), known as the D-optimality criterion, is the determinant of the information matrix or equivalently (assuming a Gaussian process) the determinant of the inverse of the covariance matrix of \( \hat{\beta} \), i.e., \( \phi_{ME}(D; \theta) = | \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X} | \), where “ME”
Environmental Sampling Design

denotes “mean estimation.” Provided that the model is second-order stationary and the mean function is constant, designs that maximize this criterion tend to consist of regularly spaced sites that cover the study region, which makes sense since such a configuration maximally reduces the dependence among pairs of observations. On the other hand, if the mean function is a planar trend, i.e., \( E[Z(s)] = s'\beta \), then the optimal design tends to have sites on or near the boundary of the study region. These differences are likely due to the higher leverage that boundary sites have when the mean function is planar rather than constant, with the result that estimation is improved by sampling more points near the boundary in spite of the strong correlation among observations taken there. The rightmost panel of Figure 9.1 displays the design that maximizes \( \phi_{ME}(\cdot; \theta) \) for the aforementioned scenario.

If, more realistically, \( \theta \) is not known and is estimated by \( \hat{\theta} \), then a natural estimator of \( \beta \) is the empirical best linear unbiased estimator \( \hat{\beta} = (X'{\Sigma}(\hat{\theta})^{-1}X)^{-1}X'{\Sigma}(\hat{\theta})^{-1}z \). An approximation for the covariance matrix of \( \hat{\beta} \) akin to that given for \( \sigma^2_{ER}(s_0; \theta, D) \) may be given, or one may again use a Bayesian approach to incorporate uncertainty about \( \theta \). In the former approach, the determinant of the inverse of the approximate covariance matrix of \( \hat{\beta} \) may serve as a design criterion. Designs that maximize this criterion usually do not differ greatly from those that maximize the D-optimality criterion for the case of known \( \theta \) and planar mean function.

### 9.2.2.4 Multi-objective and entropy-based criteria

Rather than optimizing a criterion measuring the quality of a design with respect to a single, narrowly focused objective, we may optimize a criterion that combines several objectives. In fact, as noted previously, the EK criterion of Zimmerman (2006) combines criteria for covariance parameter estimation and prediction, and the criterion of Zhu and Stein (2006) combines criteria for point and interval prediction. Two general strategies for combining criteria yield what are called compound designs (Müller and Stehlík, 2010) and Pareto-optimal designs (Lu, Anderson-Cook, and Robinson, 2011; Müller et al., 2015). A compound design criterion is a weighted linear combination of two or more component design criteria, while Pareto-optimal designs (also called the Pareto frontier) are those designs for which one of the component design criteria cannot be improved without worsening the design with respect to another component criterion. Naturally, both approaches tend to yield compromise optimal designs having some features of the designs optimal with respect to each of the component criteria.

In some situations, it may be difficult to elicit precise design objectives from the users of the environmental monitoring data, and it is also possible that the objectives may change over time. In fact, future uses of the data, and therefore future design objectives, may not necessarily be foreseen at the outset of a long-term monitoring program. For these situations, perhaps the best thing to do is try to minimize the uncertainty of the actual responses, rather than of model parameters or predictions based upon them. In a series of papers published over the last 30+ years, J. Zidek and various co-authors proposed to measure this uncertainty by the entropy of the joint predictive distribution of the responses at “ungauged” sites (i.e., unobserved \( Z(s) \)) given those at “gauged” sites. For further details on this approach, we refer the reader to summaries of this literature in Chapter 11 of Le and Zidek (2006) or Chapter 13 of this volume, plus Fuentes et al. (2007). The approach tends to add sites to an existing design that correspond to responses that are highly unpredictable, either because of their lack of correlation with responses at gauged sites or because of their greater intrinsic variability.
Probability-based spatial sampling designs, for 25 sites in a square study region. Left panel, simple random sample; middle panel, systematic sample; right panel, stratified random sample using square strata resulting from a $5 \times 5$ grid partition of the study region.

9.2.3 Probability-based spatial design

In probability-based sampling design, no model is assumed for the spatial attribute process $Z(\cdot)$; in fact, for fixed $s$, $Z(s)$, though it is unknown (unless $s$ is included in the sample), is regarded as nonrandom. Instead, randomness is introduced into the system by the investigator through a random process for selecting the sites in the design from $S$. The locations of sampled sites are determined entirely by this random process; no objective function is optimized, hence no numerical optimization algorithm is needed. The random process is such that the probability of a site being selected for the design, called the inclusion probability if $D$ is finite and the inclusion density function if $D = S$, is known for every site. Inclusion densities need not be constant across sites, and different joint inclusion densities correspond to different designs. The inclusion densities provide a mathematical foundation for an inferential paradigm known as design-based inference, which is much different than model-based inference. The objects of interest in design-based inference are usually the mean or cumulative distribution function of $Z$. Predicting unobserved values of $Z(s)$ or estimating the parameters of a model for $Z(\cdot)$ are not purposes for which probability-based sampling was developed (though one can, of course, still perform those activities for such samples).

9.2.3.1 Simple random sampling

In simple random sampling, sites are selected independently and without replacement from a uniform inclusion distribution on $D$. One such design is displayed in the left panel of Figure 9.2. The advantages of this type of sampling include that it is seen as “fair” or unbiased; it requires no knowledge about $Z$ or how it is distributed in space; and it yields a relatively simple theoretical basis for inference. The main disadvantages are that it could, by chance, result in a design in which some sample sites are quite close to one another (clusters), yielding redundant information if $Z$ is positively spatially autocorrelated, and that some subregions of appreciable size may not be sampled at all. That the locations of these clusters and empty regions would likely change from one simple random sample to another implies that estimators of population quantities of interest, though they may be unbiased, could vary a lot between samples. As a result, simple random sampling is almost never used in environmental monitoring studies.
9.2.3.2 Systematic random sampling

In systematic random sampling, a single initial site is chosen according to simple random sampling and the remaining \( n - 1 \) sites are located nonrandomly, according to some regular pattern (to the extent possible within \( D \)). The pattern could be a square, equilateral triangular, or hexagonal grid, for example. The middle panel of Figure 9.2 shows a systematic square grid design for which the grid is oriented with the axes of the square study region. It would also be possible to choose the angle of orientation randomly, according to a uniform distribution on \( (0^\circ, 90^\circ) \).

The characteristics of a systematic design differ in important ways from those of the simple random design described previously and the stratified random design to be described next. First and foremost, neither large unsampled subregions nor clusters of sites can occur, so the design has a property known as spatial balance. Spatial balance may be desirable for several reasons. For example, in air or groundwater pollution studies it may be politically unacceptable to not sample at all from large areas where significant numbers of citizens reside. Also, from a model-based perspective, spatial balance ensures that no two sites are so close together that they provide essentially redundant information. On the other hand, as noted previously, some clusters in the design may be desirable for the purpose of estimating the covariance parameters of a model. Furthermore, from a pure probability-based perspective, a disadvantage of a systematic design is that it carries with it a risk of producing poor estimates if there is a periodicity in \( Z(\cdot) \) that coincides with the orientation and spacing of the grid.

Among the various regular grids, which is best? Although site determination is generally easiest for the square grid, the triangular and hexagonal grids facilitate estimation of the covariance function in three directions, which is useful for checking for isotropy. Furthermore, the triangular grid appears to be slightly more efficient than the others for estimating the overall mean of \( Z(\cdot) \) (Matérn, 1960) as well as for purposes of covariance parameter estimation and spatial prediction (Olea, 1984; Yfantis et al., 1987).

9.2.3.3 Stratified random sampling

A sampling scheme that overcomes the weaknesses of simple random sampling without incurring problems due to periodicity is stratified random sampling. In stratified random sampling, \( S \) is partitioned into strata (subregions) and simple random samples, typically of size one or two, are taken independently within each stratum; see the right panel of Figure 9.2 for an example. Although this can still result in some small clusters, if the strata are chosen appropriately it will make it impossible for large subregions to be unsampled. Consequently, a stratified random design can be more efficient (i.e., quantities of interest are estimated with smaller variance) than a simple random design.

How should the strata be chosen? Efficiency gains are largest when the variability within strata is as small as possible, relative to the variability between strata. Thus, for example, for sampling a pollutant around a point source for which there is no favored diffusion direction, annular regions centered at the point source would be a sensible choice. For spatial sampling design more generally, it makes sense to choose strata that are geographically compact since \( Z(\cdot) \) is usually positively spatially correlated, hence more homogeneous, within such strata.

In the absence of any additional information, approximately square, triangular, or hexagonal partitionings are reasonable possibilities. A particular stratified random design that has received considerable attention is the randomized-tessellation stratified design (Overton and Stehman, 1993). The strata for this design are hexagons formed by the tessellation of a triangular grid. The regularity of the hexagonal strata confers good spatial balance upon the design, while the randomization within strata greatly diminishes any phase-correspondence with a periodic surface over \( S \). Consequently, the randomized-tessellation stratified design
is more efficient than a systematic design of the same size for estimating the mean of a population; moreover, it yields a better estimate of the variance of the mean estimator (Overton and Stehman, 1993). If $S$ is so irregularly shaped that it is not possible to form strata that are approximate hexagons, a $k$-means clustering algorithm may be applied to a fine discretization of the study region to construct geographically compact strata (Walvoort et al., 2010).

In some sampling situations, information may be available on an auxiliary variable (e.g., elevation) that is correlated with $Z$. In those situations it would often be desirable to choose strata to minimize within-strata variation of the auxiliary variable, subject to some geographical compactness constraints. A variance quadtree algorithm that accomplishes this for square strata of varying sizes was proposed by Minasny et al. (2007).

### 9.2.3.4 Variable probability sampling

More complex sampling designs are possible, including some with unequal marginal inclusion probabilities resulting in what is called variable probability sampling. Such a design might be appropriate for sampling lakes for levels of mercury in fish, for example, if one wants the probability that a lake is selected to be proportional to its size because larger lakes tend to hold more fish. The key to making unbiased inferences from a variable probability sampling design is the Horvitz-Thompson Theorem (Horvitz and Thompson, 1952) and its analogue for sampling in a continuum (Cordy, 1993). By this theorem, an unbiased estimator — known as the Horvitz-Thompson estimator — of the population total is given by the weighted sum of all the sampled values, where the weights are the reciprocals of the inclusion probabilities. Furthermore, the variance of the Horvitz-Thompson estimator is given by an expression involving the marginal and pairwise inclusion probabilities and can be estimated unbiasedly by a function of the data and those same probabilities; for details see Stevens (1997).

Stevens (1997) also extended the randomized-tessellation stratified design to allow the inclusion probability density function to vary across strata while retaining good spatial balance. The design, which he called the multiple-density, nested, random-tessellation stratified (MD-NRTS) design, achieves variable site inclusion densities by using a series of nested grids. The choices of sampling intensity in this design are limited, however. The generalized random-tessellation stratified (GRTS) design (Stevens and Olsen, 2004) overcomes this limitation by using a hierarchical randomization method to map points in two-dimensional space into one-dimensional space, while preserving proximity relationships between points to a considerable degree, and then samples systematically along the single dimension. MD-NRTS and GRTS designs have been used with considerable success in numerous monitoring studies carried out within the United States Environmental Protection Agency’s Environmental Monitoring and Assessment Program (EMAP).

### 9.2.4 Space-filling designs

Space-filling designs are, as their name suggests, designs for which no convex subregion of the study region having appreciable area (relative to the sampling intensity) is devoid of sampling points. The concept is essentially the same as that of spatial balance described previously for probability-based designs. In contrast to a probability-based design, however, the sample locations of a space-filling design are determined not by randomization but by optimizing a mathematical criterion. Furthermore, optimization of a space-filling design criterion does not require any prior knowledge of or reference to a model governing the observations, and is usually less computationally intensive than optimization of a model-based criterion. Space-filling designs have been developed primarily for use in computer...
FIGURE 9.3
Space-filling designs, for 7 sites in a square study region. Left panel, maximin design; middle panel, minimax design; right panel, maximin (and minimax) Latin hypercube design.

experiments, but they are also relevant to environmental monitoring studies. They may be particularly useful when the investigator cannot define a suitable model-based criterion or when there are multiple competing design objectives. Five types of space-filling designs have received the most attention: maximin distance designs, minimax distance designs, Latin hypercube designs, spatial coverage designs, and regular grids.

An \( n \)-point maximin distance design (Johnson et al., 1990) is an \( n \)-point design for which the shortest Euclidean distance between sampled sites is maximized, i.e., a design for which

\[
\phi_{Mm}(D) = \min_{s_i, s_j \in D, i \neq j} \| s_i - s_j \|
\]

is maximized over all possible \( n \)-point designs \( D \). Not surprisingly, such a design necessarily includes points on the boundary of the study region. Maximin distance designs are solutions to the problem of packing circles (in two dimensions) or spheres (in three dimensions) inside a region, about which much is known when the region is square, rectangular, or circular; see, e.g., http://www.packomania.com/. An example of a 7-point maximin distance design in the unit square is displayed in the left panel of Figure 9.3.

An \( n \)-point minimax distance design (Johnson et al., 1990), in contrast to a maximin distance design, is a design for which the maximum distance from any point in the study region to the closest point in the design is minimized, i.e., a design for which

\[
\phi_{mM}(D) = \max_{s \in \mathcal{S}} \min_{s_i \in D} \| s - s_i \|
\]

is minimized over all possible \( n \)-point designs \( D \). It may be more appropriate to call such a design a “minimaximin” distance design in light of the form of \( \phi_{mM}(\cdot) \), but the shorter name has prevailed. For fixed \( n \), minimax distance designs often have fewer points on the boundary than maximin distance designs. Minimax designs are related to so-called covering problems in geometry. An example of a 7-point minimax distance design in the unit square is displayed in the middle panel of Figure 9.3.

An arguably undesirable feature of maximin and minimax distance designs is that they are not necessarily space-filling when projected to their lower-dimensional subspaces. This can be visualized in the maximin and minimax designs of Figure 9.3, where the projections to either the bottom or the left side of the square would result in both coincident points and relatively large gaps between points. An \( n \)-point Latin hypercube (Lh) design (McKay et al., 1979; Pistone and Vicario, 2010) addresses this shortcoming by having the property that all of its one-dimensional projections are \( n \)-point maximin designs; this latter property is equivalent to each one-dimensional projection comprising the sequence \{0, 1/(n−1), 2/(n−1),..., 1\} (when \( \mathcal{S} \) is a unit square or cube). For any \( n \) there are many Lh designs, some of
which may be very poor with respect to global (d-dimensional) space-filling. It is common, therefore, to seek an Lh design that is globally space-filling within that class. The right panel of Figure 9.3 shows a 7-point Lh design; in fact, this particular design is both globally maximin and globally minimax within the class of 7-point Lh designs.

Royle and Nychka (1998) defined an n-point spatial coverage design as a design that minimizes

\[ \phi_{C,r,q}(D) = \left[ \sum_{s \in S} \left\{ \left( \sum_{s_i \in D} \|s - s_i\|^r \right)^{q/r} \right\} \right]^{1/q} \]

over all possible n-point designs D, where r and q are negative and positive integers, respectively. Note that \( \phi_{C,r,q}(D) \) is a function of the same quantities involved in the minimax criterion \( \phi_{mM}(D) \); more specifically, \( \phi_{C,r,q}(D) \) is an \( L_q \)-average of generalized distances between candidate points and the design. Since \( r < 0 \), the generalized distance \( \left( \sum_{s_i \in D} \|s_i - s\|^t \right)^{1/r} \) converges to 0 as s converges to any site in D. The design that minimizes \( \phi_{C,r,q}(\cdot) \) depends on \( r \) and \( q \), but appears to be insensitive to these values if \( r \) is very negative. In the limit as \( r \to -\infty \) and \( q \to \infty \), \( \phi_{C,r,q}(\cdot) \) converges to the minimax distance criterion. Royle and Nychka (1998) featured spatial coverage designs with \( r = -5 \) and \( q = 1 \).

The last category of space-filling designs, regular grids, has already been discussed in the context of probability-based design. The only difference between regular grids of the probability-based variety and those of the space-filling variety is that the locations of the gridpoints of the latter are not determined randomly but are selected either to maximize the minimum distance between them (a maximin grid design) or to minimize the maximum distance between a point in \( S \) and the nearest gridpoint (a minimax grid design). If \( S \) is rectangular, some points of a maximin grid design will necessarily lie on the boundary of \( S \), whereas points of a minimax grid design will lie entirely within the \( S \).

In practice, irregularity of \( S \), constraints on the candidate sampling sites and/or the sheer size of the design problem will preclude an investigator from finding an optimal maximin, minimax, or Lh design within the catalog of known ones, in which case the design must be found using a numerical optimization algorithm. Many of the same algorithms used to optimize model-based design criteria may be used successfully to optimize space-filling criteria (Pronzato and Müller, 2012). Exchange algorithms seem to be especially popular in this regard. Royle and Nychka (1998) presented an exchange algorithm for the construction of their spatial coverage designs and demonstrated empirically that it could obtain near-optimal designs efficiently for reasonably large problems.

How do space-filling designs compare to model-based designs? A comparison of Figure 9.3 with Figure 9.1 makes it plain that the site configuration of a globally space-filling design is similar to that of a design that optimizes a model-based prediction-error variance criterion like \( \phi_K(\cdot) \), but very different from that of a design that optimizes some other model-based criteria. Consequently, space-filling designs perform very well for spatial prediction with known covariance parameters; in fact, Zimmerman and Li (2013) and Li and Zimmerman (2015) showed, for cases with \( n = 4 \) or 5 and \( D = 4 \times 4 \) or \( 5 \times 5 \) grid (as it is in Figure, that the \( \phi_K \)-optimal design is a maximin or minimax Lh design, and Johnson et al. (1990) established a theoretical connection between maximin and minimax distance designs and designs that optimize certain other prediction-oriented criteria. However, for objectives other than spatial prediction, especially for estimation of covariance parameters, space-filling designs can perform very poorly.

For a more detailed review of space-filling designs, including some not described here, see Pronzato and Müller (2012). For applications of space-filling approaches to the design of real environmental monitoring networks, see Nychka and Saltzman (1998) and Holland et al. (1999).
9.2.5 Design for multivariate data and stream networks

Our presentation to this point has assumed implicitly that the environmental monitoring program is focused on a single variable (\(Z\)). Often, however, it is of interest to monitor two or more variables. The variables are usually correlated with each other, both non-spatially (at the same location) and spatially (at different, especially nearby sites), and the best linear unbiased multivariate prediction procedure known as co-kriging exploits both types of correlation to improve prediction quality over that which is possible by kriging each variable separately. Contributions to the literature on probability-based and space-filling sampling design generally have not been concerned with how many variables are of interest, but have assumed, either implicitly or explicitly, that if more than one variable is of interest then every such variable is observed at each site in the design. This multivariate spatial sampling design feature, known as collocation, is often practically and economically advantageous, relative to observing some of the variables at some sites and other variables at other sites. Consequently, most treatments of multivariate sampling design from a model-based perspective (e.g., Le and Zidek, 1994; Bueso et al., 1999; Yeh et al., 2006; Vašát et al., 2010) have also restricted attention to collocated designs. Within the model-based design paradigm, however, collocation is not necessarily statistically optimal.

Li and Zimmerman (2015) present an extensive investigation of multivariate model-based optimal sampling design, allowing for collocation but not requiring it. They found that, generally, the within-variable characteristics of optimal designs with respect to multivariate criteria are similar to those of designs that are optimal with respect to the analogous univariate criteria. For example, designs optimal for covariance and cross-covariance parameter estimation consist of a few clusters or strands, while designs optimal for co-kriging have excellent spatial coverage. Figures 9.4 and 9.5, reprinted from Li and Zimmerman (2015), display examples of such designs, as determined by a simulated annealing algorithm, for a scenario in which there are two variables of interest, \(D\) is a \(25 \times 25\) square grid with unit spacing, and \(n = 15\) for each variable. The observations are assumed to arise from a second-order stationary, isotropic, separable (across variables) bivariate Gaussian process with unknown constant means and a bivariate Matérn covariance function with equal correlation decay parameters and both smoothness parameters equal to 0.5. Nine cases, corresponding to all combinations of three correlation decay parameters and three cross-correlation parameters, are displayed in the figure for each design criterion. The design criteria for which the displayed designs are optimal are, in Figure 9.4, the determinant of the information matrix associated with the REML estimator of covariance and cross-covariance parameters (which is maximized), and in Figure 9.5, the maximum determinant of the \(2 \times 2\) matrix of prediction errors over a fine discretization of the unit square (which is minimized). What is at least as interesting as the spatial configuration of sites in these optimal designs is the extent of collocation. The optimal design for covariance and cross-covariance estimation tends to be completely collocated, while the optimal design for co-kriging (with either known or unknown covariance and cross-covariance parameters) can have much less collocation. The extent of collocation in the optimal co-kriging design tends to decrease as either the spatial correlation gets stronger or the cross-correlation gets weaker. On the other hand, in most cases in which collocated designs are not optimal for co-kriging, it turns out that they are reasonably efficient. These results, plus the aforementioned practical and economic considerations, provide support for using a completely collocated design in multivariate spatial sampling.

Another application area for the principles and methods of environmental sampling design described herein is stream networks. The same broad categories of designs (model-based, probability-based, and space-filling) apply, but there are some important differences as a result of the unique topology of stream networks (Dobbie et al., 2008) and the distinct
FIGURE 9.4
Optimal designs for covariance and cross-covariance estimation for the example described in Section 9.2.5. Closed circles correspond to one variable, and open squares to the other.

characteristics of models for spatial dependence on streams, namely, the tail-up and tail-down models of Ver Hoef et al. (2006) and Ver Hoef and Peterson (2010); see also Chapter 18 of this volume. Stream networks have a hierarchical branching structure and a flow direction and volume, so for variables such as point-source pollutants it may be sensible to regard observations taken at flow-connected sites to be correlated in such a way that the correlation decays with hydrologic distance (i.e., distance within the stream network) rather than with Euclidean distance, while regarding observations taken at flow-unconnected sites to be uncorrelated; this is what a tail-up model does. Alternatively, for variables like counts of fish (which can move both upstream and downstream) it can be argued that one should allow for the possibility of correlation between observations at all sites, be they flow-connected or flow-unconnected; this corresponds to a tail-down model. Furthermore, it stands to reason that an observation at a site downstream of the confluence of two tributaries of unequal flow volume would be more highly correlated with observations on the tributary having greater flow volume.

These distinctives of stream network variables may affect the relevance and utility of various sampling designs. For example, for stratified random sampling, Liebetrau (1979) proposed taking stream segments (sections of streams between confluences) as the sampling units and stratifying not on geography but on stream order (number of upstream tributaries), drainage area, or mean annual flow. A space-filling approach was proposed by Dixon et al. (1999), with the wrinkle that the maximum subcatchment area (defined
Environmental Sampling Design

as the drainage area of a sampling site minus the drainage area of any other sampling sites further upstream), rather than maximum intersite distance, was minimized. Som et al. (2014) and Falk, McGree, and Pettitt (2014) investigated model-based sampling design for stream networks from frequentist and (pseudo-)Bayesian perspectives, respectively. Som et al. (2014) found in particular that: for estimating the covariance parameters of tail-up models, the best designs have clusters on flow-unconnected segments; for estimating the covariance parameters of tail-down models, the best designs tend to have triadic clusters at stream confluences (one site just downstream of the confluence and one on each tributary just above the confluence); and for all other inference objectives, the best designs usually have a site on the outlet segment and many headwater segments, with the remaining sites evenly dispersed throughout the remainder of the network. Cost considerations may preclude using a design with too many headwater segments, however, as those segments are usually the least accessible. The findings of Falk, McGree, and Pettitt (2014) were similar.

9.2.6 Space-time designs

Environmental monitoring studies typically occur in space and time, and often have as one of their aims the detection of change or trend in $Z$ over time. Consequently, their sampling design may have a temporal component in addition to a spatial component. That said,
for some studies temporal design need not be considered. For example, if the monitoring instruments measure more or less continuously in time (e.g., water temperature loggers in streams) or aggregate over time (air filters), or if the nature of Z or the use to which it will be put demands that it be sampled at specific times (e.g., annual April 1 snow water equivalent measurements in the western U.S. mountain snowpack used to predict summer water supplies), then temporal design is not relevant. Temporal design is potentially useful only when measurements represent values of Z on a portion of the time domain T to which one desires to make inferences.

Designs for which the temporal component is given consideration may be classified as either static or dynamic. In a static design, the sites do not change over time (equivalently, the times are common across sites) and the temporal component of design may be considered independently of the spatial component. A static design may be appropriate when measuring stations are expensive to set up and cannot be moved easily and the cost of sampling and measurement at those stations is so low that it makes little sense to not sample at all of them whenever one is sampled. Analogous to the purely spatial design approaches reviewed previously in this chapter, model-based, probability-based, and “time-filling” approaches may be taken to choose the nT sampling times of a static design. Of these, model-based designs seem to be much less common. Perhaps the most common static temporal design consists simply of regular time points 1, 2, ..., nT (in appropriate time scale units), due partly to its time-filling properties and partly to the relative convenience of a constant sampling interval.

If a design is not static, it is dynamic. A class of dynamic designs for environmental monitoring that has received much attention is the class of panel designs (Urquhart et al., 1993), also called revisit designs (McDonald, 2003). Here, a “panel” refers to a set of sites that are sampled contemporaneously. Two important types of panel designs are rotating panel designs and serially alternating designs. For concreteness suppose that the basic sampling interval is a year. A rotating panel design prescribes that each panel will be sampled in each of several consecutive years and then removed permanently from future consideration; as each panel is removed, another is started. The top portion of Table 9.1a shows a rotating panel design in which three panels are sampled in each of four years. In contrast, a serially alternating design prescribes that each panel is sampled in every rth year, where r is an integer. The top portion of Table 9.1b displays a serially alternating design in which there are four panels, each of which is sampled every fourth year, for a total sampling period of 8 years.

Both types of panel designs just described may be augmented with a panel that is sampled every year. The designs in Table 9.1, with the bottom portions included, are an augmented rotating panel design and an augmented serially alternating design, respectively. Augmentation improves the designs for certain purposes; for example, without augmentation, the estimation of a change in Z from one year to the next is nonestimable in the serially alternating design of Table 9.1b. Urquhart et al. (1993) recommended allocating 10-20% of the sites to the augmented panel for best performance of either panel design. Urquhart and Kincaid (1999) showed that an augmented serially alternating design has greater power to detect trend than an augmented rotating panel design with the same total number of sites, but that the latter estimates current “status” (the mean of Z over space at a particular time) more precisely than the former.

Rotating and serially alternating panel designs have obvious advantages over static designs, but their structure requires that every site be sampled in multiple years. A fully dynamic design allows each panel to be sampled only once. Wikle and Royle (1999) considered fully dynamic model-based design for monitoring spatiotemporal processes. For a relatively simple Gaussian separable model with first-order Markovian temporal dependence, they found that the degree to which the average prediction error variance at time
Environmental Sampling Design

(a)

<table>
<thead>
<tr>
<th>Panel</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Augmented</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Panel</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Augmented</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

TABLE 9.1
(a) An augmented rotating panel design, and (b) an augmented serially alternating panel design. An asterisk indicates that the panel is sampled in that year.

$t$ is reduced by a dynamic design, from what it is for a static design, is largest when the temporal dependence is strong. Furthermore, in that circumstance the spatial locations of the optimal dynamic design for spatiotemporal prediction at time $t$ are typically quite far removed from those at time $t-1$. This is intuitively reasonable, since $Z(s, t-1)$ is strongly informative of $Z(s^*, t)$ when $s^*$ is at or near $s$ and the temporal dependence is strong. Not surprisingly, when the spatiotemporal process has more complicated dynamics or is nonseparable or non-Gaussian, the dynamic prediction-optimal design can be more complicated, even nonintuitive, but is still much more efficient than the optimal static design (Wikle and Royle, 1999, 2005).

A more highly specialized space-time design is the Before-After-Control-Impact (BACI) design (Green, 1979; Stewart-Oaten, 1986). The objective of a BACI design is to assess the impact on the environment of some human activity or disturbance. For example, it may be desirable to determine the effect that effluent from a newly built municipal sewage treatment plant, discharged into a river, has on aquatic organisms downstream. In a standard BACI design, samples are taken at two classes of sites, one designated as impact sites where the effects of the activity or disturbance, if any, would occur, and the other as control sites. At each site, observations are taken at two sets of times, one before the activity begins and the other after. Whether the activity has impacted the environment is assessed by acting as though the BACI design is a classical randomized experiment and carrying out an F-test for Time by Treatment interaction in a linear model that has fixed effects for two times (before and after) and two treatments (control and impact) and their interaction. However, since the treatments are not actually randomly assigned to sites, inferences must be model-based
rather than design-based. This and several other issues have led to the development of many variations on the basic BACI design. For an overview of these variations, see Smith (2002).

### 9.2.7 Discussion

Thus far we have described three main approaches to design for environmental monitoring — model-based, probability-based, and space-filling — and we have indicated which approaches are applicable and likely to be effective for various objectives. A fundamental requirement of the model-based approach is, of course, that a model be assumed for the spatial process. If there are several competing models and uncertainty about which is best, one possibility would be to select, at the first stage of sampling at least, a design that optimally discriminates among them. A large literature exists on the topic of optimal design for model discrimination in a classical regression context, but virtually no attention has been given to this problem in a spatial context. Accordingly, if the model is uncertain or unknown, presently we recommend using a probability-based, space-filling, or “hybrid” design. Hybrid designs are probability-based designs that have good spatial coverage over the region of interest, yet have some sites close together. An example of a class of hybrid designs is the “lattice plus close pairs” design of Diggle and Lophaven (2006), which has recently been studied further by Chipeta et al. (2016).

Although spatial sampling design has matured considerably within the past 20 years, research continues along several fronts. For example, as geostatistics for functional data has developed, so has spatial sampling design for such data (Bohorquez, Giraldo, and Mateu, 2016). For the future, there is surely a need for the development of designs that can discriminate among competing spatial models, and for more work on adaptive (multi-stage), multivariate, and space-time design.

### 9.3 Sampling for Estimation of Abundance

When estimating the abundance of individuals in a biological population, the issue of detectability of objects of interest (usually animals or animal groups) typically arises. That is, we cannot usually assume that all objects present on a sampling unit are detected. If we can, then simple plot sampling methods can be adopted, in which some form of probability-based design (Section 9.2.3) is adopted, and design-based methods yield the required abundance estimates. The abundance estimator for any given design can be formulated as a special case of the unbiased Horvitz-Thompson estimator (Horvitz and Thompson, 1952).

When detection of an object on a plot is uncertain, we must estimate the probability of detection, so that the inclusion probability of the Horvitz-Thompson estimator must be estimated. Provided we adopt a consistent estimator of the detection probability, the resulting “Horvitz-Thompson-like” estimator (Borchers and Burnham, 2004) is asymptotically unbiased.

We first consider distance sampling, which may be considered an extension of plot sampling in which not all objects on a sampled plot are counted. In the case of point transect sampling, the plots are circles of radius $w$ centered on the point, and for line transect sampling, the plots are strips of width $2w$, extending out a distance $w$ either side of the transect. An observer travels to each point, or along each line, recording distance from the point or line of each object detected. If bearing (points) or distance along the line (lines) is also recorded for each detected object, then the data may be regarded as a thinned spatial point process. Conventional distance sampling methods are a hybrid of model-based methods to
Environmental Sampling Design

199

estimate probability of detection and design-based methods to estimate abundance in the wider survey region, given estimates of detection probabilities, but fully model-based methods can also be adopted. We only address design issues here; see Buckland et al. (2015) for analysis methods.

Survey design has received less attention for mark-recapture methods, but this is starting to change, with the rise of spatial capture-recapture methods (Efford, 2004; Borchers and Efford, 2008; Royle et al., 2014). In spatial capture-recapture, sensors are located according to some design through the study region. With the need to capture a number of animals at more than one location, the spacing of sensors becomes crucial, especially for large study areas, when sensors may have to be distributed in clusters, with large gaps between clusters. Note the use of the term ‘sensor’ here. These may be the more traditional traps of conventional capture-recapture, but they may represent some other means of detecting (and re-detecting) animals, for example camera traps, acoustic sensors (for which the ‘objects’ sampled might be individual animals, but might be individual calls from animals), or observers.

9.3.1 Distance sampling

Distance sampling comprises a suite of methods based on a design of lines or points. Line transect sampling is the most widely-used technique, while point transect sampling is sometimes preferred, especially for songbird surveys. There are several variants on these two schemes. The sampling units are the plots centered on the lines (strips of half-width $w$) or points (circles of radius $w$). A complication is that $w$ is often not known in advance, but is estimated from the data. Further, in multi-species surveys, $w$ may vary by species. Thus a strategy of pre-defining all possible plots in the study region, and of selecting a sample of these plots, is often not an option.

9.3.1.1 Standard probability-based designs

Point designs

Any of the designs of Figure 9.2 may be used. Systematic designs are generally preferred, as they give abundance estimates with higher precision when object density through the survey region exhibits spatial trends (Fewster et al., 2009). Sometimes, strata are defined, and a systematic sample of points selected in each stratum.

With either simple random sampling or stratified random sampling, because the sampling units are circles of radius $w$, it is possibly for units to overlap when the design is based on points. (Systematic sampling avoids this minor complication.) If $w$ is known, the study region can be divided into squares of side $2w$, and a simple random or stratified random sample of squares selected. If $w$ is estimated from data, the squares can be made larger, so that the length of one side is certain to exceed $2w$. The sample points are then located at the centers of the sampled squares.

If travel between points is costly or time-consuming, a cluster of points might be positioned at each sampling location. The analyst can then either pool data across each cluster, or adopt cluster sampling methods to estimate abundance.

In some surveys, a line design is used, and points are placed systematically along the lines. See Strindberg et al. (2004) for details of this and other variants.

Line designs

Line designs are more problematic than point designs, as the plots are typically much longer than they are wide. If it is practical to have long lines traversing the width of the study area, a simple design is a systematic sample of parallel lines spanning the full length of the
study area (Figure 9.6, top left panel). For surveys in which lines can be traversed quickly but surveyed slowly, such as in dung surveys (Marques et al., 2001), short, systematically-spaced segments are often covered along each line (Figure 9.6, top right). If the distance between segments is the same as the distance between successive parallel lines, this gives a systematic sample of line segments through the study region.

Typically, the number of lines in a line transect survey, or within a single stratum of a stratified design, is modest, so that a given realization of a simple random design can give uneven cover. A systematic random design addresses this issue (Fewster et al., 2009). Another option is shown in the bottom left panel of Figure 9.6; here, the survey region is partitioned by a grid of rectangles, each of width $2w$, where $w$ is the distance from the line beyond which detections are not recorded. A simple random sample of rectangles can then be selected, and a line placed down the center of each sampled rectangle. A similar method is to have fewer, wider rectangles, and a line is randomly located within each rectangle. As noted in Section 9.2.3, this is a form of stratified sampling with one unit sampled per stratum.

Especially for shipboard surveys of large regions, but also for aerial surveys when successive lines would ordinarily be widely-separated, zig-zag or saw-tooth designs are commonly used. These ensure that search effort is more-or-less continuous, without long gaps while the vessel travels to the next line. Thus for a given level of resources, additional surveying can be done. Such designs are similar to a systematic random design, but coverage probability varies if the width of the study region varies. The bottom right panel of Figure 9.6 illustrates how such designs can be implemented so that the variability in coverage probability is smaller than for a design in which the sampler angle is constant (Strindberg et al., 2004). A major axis is defined along the length of the survey region, and a systematic grid of parallel lines is placed at right-angles. Waypoints are defined where these lines intersect the region boundary, and the transects are formed by straight lines linking waypoints, alternating from one boundary to the other.

Automated survey design

Automated survey design can be very useful when used in conjunction with a Geographic Information System. Survey regions are often complex shapes, and unconstrained random designs can be inefficient, for example when a ship must travel long distances to travel from one line to another. An automated design algorithm can be used to quantify coverage probability through the survey region for constrained designs, zig-zag designs, or other designs that do not given uniform coverage. This also allows different designs to be compared for efficiency, allowing the user to evaluate options before committing resources.

Edge effects

For small survey regions, edge effects can be problematic. A simple solution is to adopt “plus-sampling”: the samplers (circles or strips) are located through an enlarged region that includes a “bufferzone” extending a distance $w$ beyond the boundary of the survey region. Observers then cover points or line segments located in the bufferzone, recording any detected objects that are within the survey region boundaries. If it is not possible or cost-effective to do this, then “minus-sampling” is used, resulting in reduced coverage along the boundary. When using an automated survey design, simulation can be used to quantify coverage probability through the region, so that bias is not introduced to abundance estimates.

Model-based versus design-based methods

When design-based analysis methods are used to extrapolate from the sampled plots to
Environmental Sampling Design

FIGURE 9.6
Four options for line transect designs. Top left: systematic sample of parallel lines. Top right: systematic grid of line segments (thick lines), placed along lines traversing the study region (dashed lines). Bottom left: a simple random sample of rectangular sampling units; the line is placed down the center of each selected unit. Bottom right: a zig-zag or saw-tooth sampler, determined by joining systematically-spaced waypoints alternating between top and bottom boundary of the study region.

the survey region, careful consideration needs to be given to the survey design, to ensure that abundance and precision are reliably estimated. When model-based methods are used to model object density through the survey region (Buckland et al., 2016), the survey design is less influential, giving greater flexibility for designing the survey. However, the design should seek to ensure that samplers are spread throughout the survey region, and that they span the range of values for any covariate used to model animal density.

9.3.1.2 Adaptive distance sampling designs

Distance sampling is especially effective for surveying objects that are sparsely distributed through a large region. However, if those objects tend to occur in a relatively small number of patches, standard distance sampling may yield both a low number of detections (because much effort is expended away from these patches) and high variance.

Adaptive cluster sampling (Thompson, 2002) was designed for surveying objects whose spatial distribution is patchy, providing a mechanism for conducting additional survey effort in localities where objects have been found, while giving design-unbiased estimation. The
method was adapted for distance sampling by Pollard (2002) and Pollard and Buckland (2004).

An initial sample of units is selected at random. If the number of observations in a sampled unit satisfies some trigger condition, then units in a pre-defined neighborhood of the triggering unit are also sampled. If any of the adaptive units in the neighborhood meet the condition, then the neighborhood of each of these units is also sampled. The process repeats until no newly-sampled units meet the condition. The combination of an initial unit and its associated adaptive units is termed a cluster. Within the cluster, any units which do not meet the trigger condition are termed edge units, whilst any units which meet the condition form a network.

Consider first point transect sampling. In Figure 9.7, we show a definition of the neighborhood around a sample point based on square units. Designs can be based on other units, for example triangles or hexagons (Pollard and Buckland, 2004). Figure 9.8 shows a realization of a design based on square units, in which an initial systematic sample of squares is selected, and squares in the neighborhood are surveyed if at least one object is detected in a sampled square. Further detections trigger additional effort, so that initially separate units can merge into a single cluster.

**FIGURE 9.7**
An adaptive design based on square units. The shaded circle is one of the original sample of plots, and the neighborhood is defined here as the four plots immediately adjacent to the shaded plot. The point in the shaded plot is surveyed, and if extra effort is triggered, the point at the center of each of the four plots in the neighborhood is surveyed.

For line transect surveys, we can define units as in the bottom left panel of Figure 9.6. Thus the units are rectangles, and for sampled units, the line passes through the center of the rectangle. A realization of a design of this type is shown in Figure 9.9. Here, the initial line transect is taken as the centerline of a strip of units placed end-to-end, and the neighborhood of a unit is defined as the two units to either side of that unit, above and below the strip. In this example, a single detection triggers effort in the neighborhood, and if at least one further detection is made in a neighborhood unit, the next unit away from the initial strip is also surveyed.
A major limitation of design-unbiased adaptive distance sampling methods is that the amount of survey effort is not known in advance. For marine surveys carried out by ship or aircraft, the platform is typically available for a pre-determined length of time. If the survey finishes early, resources are wasted, while if it is not possible to complete the survey in the time available, some of the survey region will remain unsurveyed, compromising abundance estimation. For this reason, Pollard et al. (2002) developed a form of adaptive distance sampling in which the additional effort triggered is a function of how much total survey effort is still available, thus ensuring that the survey finishes on schedule. While no longer design-unbiased, bias was found to be small. Pollard et al. (2002) tested this fixed-effort adaptive distance sampling method on harbour porpoise in the Gulf of Maine and Bay of Fundy region. Adaptive and conventional methods gave very similar estimates of porpoise density, with slightly improved precision for the adaptive method. Perhaps the most significant gain from the adaptive method however was a substantial increase in the number of detections (551, compared with 313 for the conventional survey).

9.3.1.3 Designed distance sampling experiments

In distance sampling, the objective is usually to estimate abundance of the population of interest. Sometimes, we might instead wish to assess the effects (if any) of a management regime or an impact on population densities. For example, does spraying a forest block with pesticide to eradicate an insect pest have a detrimental effect on songbird densities?
FIGURE 9.9
Part of a design comprising an initial systematic sample of lines. Each line is divided into segments, and the rectangles formed by extending each segment out to a distance $w$ either side of the line form the units. A detection within a unit triggers effort in the neighborhood, which is formed by the units immediately above and below it.

Or does a proposed conservation measure in farmland have the desired effect on species of conservation concern?

Typically in a designed experiment using distance sampling, there is only one treatment, and we wish to compare that treatment with a control. There might be no time element (we simply compare densities on treatment plots with those on control plots at the time of the surveys), or we may have repeat visits to plots, and are interested in change over time. We cannot assume that repeat counts at the same plot are independent.

Design considerations are similar to those for the design of agricultural experiments, while analysis methods need to account for imperfect detection of objects on plots, and for the analysis of counts (Buckland et al., 2015). In the case of environmental impact assessments, a BACI design (before/after control/impact; McDonald et al., 2000; Buckland et al., 2009) can be useful when it is possible to have adequate replication. Thus plots are established on both control and treatment sites, and monitored before and after the treatment is applied.

Matched-pairs designs are useful for comparing a single treatment with a control. Where the treatment can readily be applied to a single plot, the design can utilize paired plots, as in a large-scale study to assess the effect of field conservation buffers on bird densities (Evans et al., 2013; Oedekoven et al., 2014). Where the treatment is applied at the level of
site, with replicate plots at each site, then sites may be paired, as in a study assessing the
effect on bird densities of controlled burns in ponderosa pine forest (Buckland et al., 2009).

9.3.2 Capture-recapture

9.3.2.1 Standard capture-recapture

There has been relatively little work done on survey design in non-spatial capture-recapture.
When the goal of the study is to estimate the size of a population, the requirements of the
approach are quite demanding. For example, no animal in the population should have zero
probability of capture; in practice, for reliable estimation, we seek a design that ensures no
probabilities are close to zero. Thus traps need to be spread throughout the range of the
population of interest. Realistically, the approach is usually only useful for populations that
are contained within a small study area.

When the population extends beyond the study area in which traps are set, animal
movement onto the study area results in over-estimation of abundance, unless the effective
area trapped can be estimated reliably. The bias reduces as the size of the study area
increases; study area size should be appreciably larger than average home range size. For a
camera trap survey of ocelots, Maffei and Noss (2008) concluded that the study area should
be at least four times the size of the average home range.

More commonly, capture-recapture methods are used to estimate survival rates. In this
context, survey design is seldom considered in detail. It is assumed that marked animals
are representative of animals in the population, or at least of those that have the same
values of any modeled covariates. There is no longer any requirement that all animals in
the population have non-zero probability of capture, and study area size is only important
if death of a marked animal is indistinguishable from emigration from the study area.

9.3.2.2 Spatial capture-recapture

To estimate abundance using spatial capture-recapture methods, it is no longer necessary
for all animals in the population to have non-zero probability of capture, nor is it necessary
for the study area to be large relative to the size of average home range. Of course, the
abundance being estimated is defined by the boundaries of the study area, rather than by
the biological population of interest.

As noted by Royle et al. (2014), solutions to the problem of survey design in spatial
capture-recapture are largely ad hoc. They recommend simulation to explore different trap
configurations. Successful design relies on knowledge of the spatial behavior of the species,
to ensure that simulations identify an efficient design.

Royle et al. (2014) note that model-based design (Section 9.2.2) is a standard approach
in classical sampling, but has not seen widespread use in ecology. They believe that the
approach has great potential for spatial capture-recapture, and give a preliminary formul-
ation. They note that normal linear models, as described in Section 9.2.2, are unsuitable
for spatial capture-recapture, for which a Poisson or binomial model is more appropriate.
They thus develop methods adopting the formulation of generalized linear models, allowing
a link function \( g(\cdot) \). Thus they take

\[
g(E(y)) = \alpha_0 + \alpha_1 d(x, s)^2 = M' \alpha
\]

where \( y \) is the vector of length \( J \) of encounter frequencies in the \( J \) traps, \( d(x, s) \) is the
distance between trap location \( x \) and individual activity center \( s \), \( \alpha \) is a vector of parameters
of length two, and \( M \) is the \( J \times 2 \) design matrix. They then proceed to develop an optimal
design criterion for spatial capture-recapture.
9.3.3 Discussion

The rapid rise in the use of spatial capture-recapture methods is likely to result in renewed interest in the design of capture-recapture studies. Most such studies are carried out within small study areas, and the issue of optimal design when the methods are applied to large study areas is a difficult one that requires effective solutions, if such studies are to become commonplace. To ensure adequate recaptures at distinct locations in a large study area, sensors must be placed according to a clustered design, and the optimal trade-off between number of clusters and number of sensors within each cluster is unclear. In distance sampling, large-scale designed experiments are starting to be used to assess the impact of a treatment, development or management action, and again, optimal design strategies are needed to ensure efficient use of resources.

Bibliography


