MARKOV CHAIN DESIGNS FOR ONE-PER-STRATUM SPATIAL SAMPLING

F. Jay Breidt, Iowa State University
221 Snedecor Hall, Ames, IA 50011

Keywords: Balanced systematic sampling, Global positioning system, Soil surveys, Spatial sampling design, Systematic sampling.

Abstract: Markov chain designs are developed for the sampling of a continuous two-dimensional spatial domain. These designs include as special cases systematic sampling, balanced systematic sampling, and stratified random sampling with one sampling unit per stratum. Designs are compared on the basis of their anticipated variances under superpopulation models which include both large-scale variation and small-scale variation. An example of the implementation of Markov chain sampling for a multiphase soil mapping project in Crawford County, Iowa is given.

1 Introduction

The design for the Crawford County soil mapping project is a three-phase sample. In the first phase, 1600 points are selected and surface horizon data are collected. In the second phase, 400 points are subsampled from the 1600 Phase 1 points, and data on several variables are collected for all horizons down to some specified depth. In the third phase, 200 of the Phase 2 points are selected. Laboratory analyses are conducted on about six horizons at each of the Phase 3 points, for a total of about 1200 laboratory samples. Data elements collected during the three phases include general landscape features at the point (aspect, gradient, etc.), and horizon-specific physical characteristics of the soil (color, effervescence, % clay/sand, depth to mottles/carbonates/redoximorphic conditions/free water, etc.).

Points are located precisely (to within a few meters) using a global positioning system, but features at the point cannot be determined without a field visit. Since data cannot be collected on roads, ditches or railroads, and since data can only be collected after authorization near buried cables and pipelines, it is desirable to avoid aligning or equally spacing the points. Delays from drawing additional points or obtaining authorization are undesirable since the best field season is the relatively short time between spring thaw and crop emergence.

Selection of the points at the first phase is via a controlled version of one-per-stratum sampling. The idea is to draw a point sample which is well-dispersed spatially, like a systematic (or grid-based) sample would be, but is protected against systematic sources of error (such as roads, underground cables, etc.) through additional randomness. Systematic designs can be extremely inefficient for populations with systematic features like trends and periodicities (e.g., Madow and Madow 1944, Cochran 1977, Bellhouse 1988).

Spatial control is provided by making the X and Y coordinates of the sample points evolve according to Markov chains, and so the design is referred to as a Markov chain (MC) design. See Breidt (1995) for discussion of MC designs in the one-dimensional, finite population case. The Markov chains used in the Crawford County soil mapping project are two-dimensional and have a continuous state space.

Quenouille (1949) makes a distinction between aligned and unaligned spatial sampling designs. Given the potential problems with aligned points in our application, we consider only unaligned designs.

2 Markov chain designs

Let (x_0, y_0) denote the longitude and latitude of the southwest corner of Crawford County, which will be represented as a rectangular spatial domain, $D \subset \mathbb{R}^2$. A sampling rate is determined and the county is accordingly split into rectangular strata with sides of a degrees of longitude and b degrees of latitude. Number the rows of strata $i = 1, \ldots, I$ and the columns $j = 1, \ldots, J$.

Denote by $\Phi(\cdot)$ the cumulative distribution function of the standard normal. Points selected according to the MC design are then

$$\{(X_{ij}, Y_{ij}) : i = 1, \ldots, I;\ j = 1, \ldots, J\},$$

where

$$X_{ij} = x_0 + a(j-1) + X^*_{ij},$$

$$Y_{ij} = y_0 + b(i-1) + Y^*_{ij},$$

$$X^*_{ij} = a\Phi\left(V^{(1)}_{ij}\right), \quad Y^*_{ij} = b\Phi\left(V^{(2)}_{ij}\right),$$

$$V^{(1)}_{ij} = \begin{cases} W^{(1)}_{ij}, & j = 1, \\ \phi_1 V^{(1)}_{i,j-1} + W^{(1)}_{ij}\sqrt{1-\phi_1^2}, & j = 2, \ldots, J, \end{cases}$$

and

$$V_{ij}^{(2)} = \begin{cases} W_{ij}^{(2)}, & i = 1, \\ \phi_2 V_{i-1,j}^{(2)} + W_{ij}^{(2)}\sqrt{1-\phi_2^2}, & i = 2, \ldots, I. \end{cases}$$

Here, $\{W_{ij}^{(1)}\}$ is independent and identically distributed (iid) N(0,1) and $\{W_{ij}^{(2)}\}$ is iid N(0,1), independent of $\{W_{ij}^{(1)}\}$. The $\{V_{ij}^{(k)}\}$ are first-order autoregressive processes with standard normal marginal distributions.

Because of Gaussianity, the processes $\left\{\Phi\left(V_{ij}^{(1)}\right)\right\}$ and $\left\{\Phi\left(V_{ij}^{(2)}\right)\right\}$ are reversible, meaning that the probabilistic structure remains the same whether the sample is drawn east to west or west to east, north to south or south to north. Interestingly, for $\phi_k \neq 0$, reversibility holds in this design if and only if $\{W_{ij}^{(k)}\}$ is Gaussian (Weiss 1975).

The parameters ϕ_k dictate the type and degree of control: for $\phi_1 = 1$, the sample in a given row is systematic after a random start (SY); for $\phi_1 = 0$, the row sample is stratified simple random sampling with one point per stratum (ST); and for $\phi_1 = -1$, the row sample is balanced systematic sampling (BA) (Murthy 1967). In spite of this control, all points are equally likely; that is, the unconditional probability density function on any of the $I \times J$ strata is uniform.

Denote realizations of X_{ij} and Y_{ij} by

$$x_{ij} = x_0 + a(j-1) + x_{ij}^* \text{ and } y_{ij} = y_0 + b(i-1) + y_{ij}^*.$$

The conditional probability density function of X_{ij} given $X_{i,j-1} = x_{i,j-1}$ is

$$f(x_{ij}|x_{i,j-1}) = \frac{a}{(1-\phi_1^2)^{1/2}} \exp\left\{\frac{-\phi_1}{2(1-\phi_1^2)}\zeta_{\phi_1}(x_{ij}^*, x_{i,j-1}^*)\right\} \times I_{[0,a]}(x_{ij}^*) I_{[0,a]}(x_{i,j-1}^*)$$

where

$$\zeta_{\phi_1}(x_{ij}^*, x_{i,j-1}^*) = \phi_1 \Phi^{-1}(x_{ij}^*)^2 + \phi_1 \Phi^{-1}(x_{i,j-1}^*)^2 - 2\Phi^{-1}(x_{ij}^*)\Phi^{-1}(x_{i,j-1}^*),$$

and the conditional probability density function of Y_{ij} given $Y_{i-1,j} = y_{i-1,j}$ is

$$f(y_{ij}|y_{i-1,j}) = \frac{b}{(1-\phi_2^2)^{1/2}} \exp\left\{\frac{-\phi_2}{2(1-\phi_2^2)}\zeta_{\phi_2}(y_{ij}^*, y_{i-1,j}^*)\right\} \times I_{[0,b]}(y_{ij}^*) I_{[0,b]}(y_{i-1,j}^*).$$

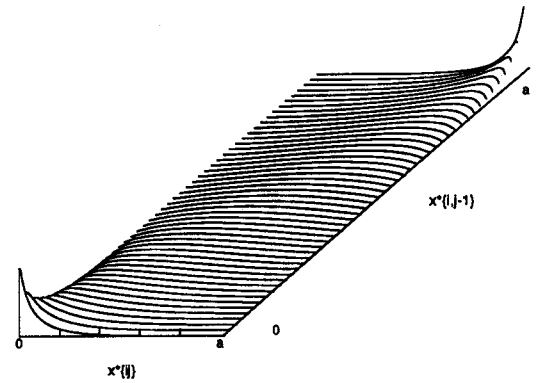

Figure 1: *Conditional densities for* $X_{ij}^* = X_{ij} - x_0 - a(j-1)$ *given* $X_{i,j-1} = x_0 + a(j-2) + x_{i,j-1}^*$ *(*$\phi_1 = 0.75$*).*

Figure 1 shows conditional densities of X_{ij}^* given different possible values of $X_{i,j-1}^*$, where $\phi_1 = 0.75$. Note the ridge of high density; with high probability, the sampled x-coordinate in stratum (i,j) is near the same relative position as the sampled x-coordinate in stratum $(i,j-1)$. If $\phi_1 = 1$, the sampled x-coordinates would be in exactly the same relative position from stratum to stratum across the row, and each conditional density would have all of its mass on the ridge. If $\phi_1 = 0$, the sampled x-coordinates would be independent from stratum to stratum and all the conditional densities would be flat (uniform on $[0,a]$).

Figure 2 shows a possible Markov chain sample in five rows and five columns using $\phi_1 = \phi_2 = 0.75$. Lines represent stratum boundaries. The sampled locations in stratum (i,j) and its four nearest neighbors are labeled.

The sampled location (X_{ij}, Y_{ij}) is independent of all $(X_{i'j'}, Y_{i'j'})$ to the northwest $(i < i',\ j > j')$, to the northeast $(i < i',\ j < j')$, to the southeast $(i > i', j < j')$, and to the southwest $(i > i', j > j')$. Given the sampled locations for the four nearest neighbors, the sampled location (X_{ij}, Y_{ij}) is also conditionally independent of $(X_{i,j'}, Y_{i,j'})$ for $j' \neq j-1, j, j+1$ and conditionally independent of $(X_{i',j}, Y_{i',j})$ for $i' \neq i-1, i, i+1$.

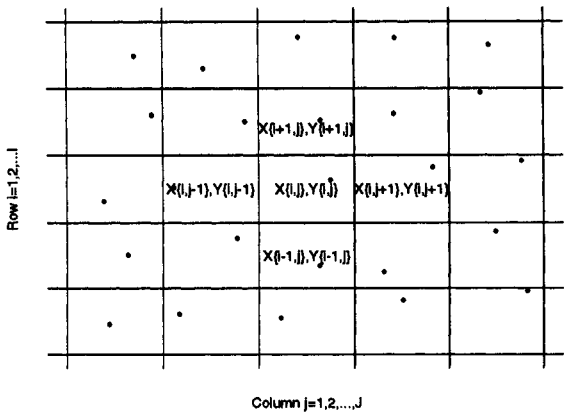

Figure 2: *Example map of points selected via one-per-stratum Markov chain design with* $\phi_1 = \phi_2 = 0.75$. *Lines represent stratum boundaries. Sampled locations in stratum* (i, j) *and its four nearest neighbors are labeled.*

3 Estimation under MC

Suppose that the population parameter of interest is

$$t_z = \int_{x_0}^{x_0+aJ} \int_{y_0}^{y_0+bI} z(x,y)\, dx\, dy.$$

A design-unbiased estimator of t_z under any MC design is

$$\hat{t}_z = ab \sum_{i=1}^{I} \sum_{j=1}^{J} z(X_{ij}, Y_{ij}).$$

The design variance of $\hat{t}_z$, $\mathrm{V}_p\left(\hat{t}_z\right)$, depends on all the values of the study variable $z(x,y)$ in the domain D as well as the covariance structure of the X_{ij}'s and Y_{ij}'s, and so is not easily used for comparing designs.

4 A superpopulation model

Following Cochran (1946), assume that the values of the study variable $z(x,y)$ are realizations of a stochastic process $Z(x,y)$ following some model, ξ. Designs can then be compared on the basis of anticipated variance, $\mathrm{E}_\xi\left[\mathrm{V}_p\left(\hat{t}_z\right)\right]$. Note that the total variance, over both design and model, of $\hat{t}_z$ is

$$\begin{aligned} \mathrm{V}_{p\xi}\left(\hat{t}_z\right) &= \mathrm{V}_\xi\left(\mathrm{E}_p\left[\hat{t}_z\right]\right) + \mathrm{E}_\xi\left[\mathrm{V}_p\left(\hat{t}_z\right)\right] \\ &= \mathrm{V}_p\left(\mathrm{E}_\xi\left[\hat{t}_z\right]\right) + \mathrm{E}_p\left[\mathrm{V}_\xi\left(\hat{t}_z\right)\right], \end{aligned}$$

so that

$$\begin{aligned} &\mathrm{E}_\xi\left[\mathrm{V}_p\left(\hat{t}_z\right)\right] = \\ &\quad \mathrm{V}_p\left(\mathrm{E}_\xi\left[\hat{t}_z\right]\right) + \mathrm{E}_p\left[\mathrm{V}_\xi\left(\hat{t}_z\right)\right] - \mathrm{V}_\xi\left(\mathrm{E}_p\left[\hat{t}_z\right]\right). \end{aligned}$$

Since $\hat{t}_z$ is design-unbiased, $\mathrm{V}_\xi\left(\mathrm{E}_p\left[\hat{t}_z\right]\right)$ is constant across MC designs and will be ignored in what follows.

A common geostatistical model for spatial data (e.g., Cressie 1991, §2.3) is

$$\xi: \quad Z(x,y) = \mu(x,y) + \epsilon(x,y) + \nu(x,y),$$

where $\mu(x,y)$ is the non-stochastic mean structure or *large-scale variation* and $\{\epsilon(x,y) : (x,y) \in D\}$ is a zero-mean, second-order stationary stochastic process with autocovariance function (ACVF)

$$\mathrm{Cov}_\xi\{\epsilon(x,y), \epsilon(x+dx, y+dy)\} = \gamma(dx, dy).$$

The ACVF describes the *small-scale variation* in the data.

A particular version of the above large-scale variation model is the linear trend model,

$$\mu(x,y) = \alpha x + \beta y,$$

where, without loss of generality for variance computations, the intercept is taken to be zero. This model has been considered by Bellhouse (1981) for a two-dimensional lattice process.

Result 1 *The contributions of the linear trend* $\mu(x,y) = \alpha x + \beta y$ *to the total variance are*

$$E_p\left[V_\xi\left(ab\sum_{i=1}^{I}\sum_{j=1}^{J}\mu(X_{ij}, Y_{ij})\right)\right] = 0$$

and

$$\begin{aligned} V_p\left(E_\xi\left[ab\sum_{i=1}^{I}\sum_{j=1}^{J}\mu(X_{ij}, Y_{ij})\right]\right) = \\ (ab)^2\alpha^2 I\, V_p\left(\sum_{j=1}^{J} X_{ij}^*\right) \\ +(ab)^2\beta^2 J\, V_p\left(\sum_{i=1}^{I} Y_{ij}^*\right), \end{aligned}$$

which simplifies to

$$\begin{cases} 0, & I, J \text{ even}, \\ (ab)^2\frac{\alpha^2 a^2 I}{12}, & I \text{ even}, J \text{ odd}, \\ (ab)^2\frac{\beta^2 b^2 J}{12}, & I \text{ odd}, J \text{ even}, \\ (ab)^2\frac{\alpha^2 a^2 I + \beta^2 b^2 J}{12}, & I, J \text{ odd}, \end{cases}$$

under BA;

$$(ab)^2\frac{\alpha^2 a^2 IJ + \beta^2 b^2 IJ}{12}$$

under ST; and

$$(ab)^2 \frac{\alpha^2 a^2 IJ^2 + \beta^2 b^2 I^2 J}{12}$$

under SY. □

Result 2 *The contributions of the spatially-autocorrelated error $\epsilon(x, y)$ to the total variance are*

$$E_p\left[V_\xi\left(ab\sum_{i=1}^{I}\sum_{j=1}^{J}\epsilon(X_{ij}, Y_{ij})\right)\right] = (ab)^2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J} E_p\left[\gamma\left(|X_{ij} - X_{i'j'}|, |Y_{ij} - Y_{i'j'}|\right)\right]$$

and

$$V_p\left(E_\xi\left[ab\sum_{i=1}^{I}\sum_{j=1}^{J}\epsilon(X_{ij}, Y_{ij})\right]\right) = 0. \quad \square$$

One possible choice for the ACVF of ϵ is the exponential ACVF

$$\gamma(dx, dy) = \sigma^2 \exp\left\{-\delta\left(dx^2 + dy^2\right)^{1/2}\right\},$$

where $\sigma > 0$ and $\delta > 0$. This ACVF has been considered by Matérn (1947), Zubrzycki (1958) and Hájek (1961), among others. See Das (1950) and Cressie (1991, §2.3.1) for other possible ACVF choices.

5 Variance comparisons

From the above results, the total variance can be evaluated by computing the design variances of $\sum_{j=1}^{J} X_{ij}^*$ and $\sum_{i=1}^{I} Y_{ij}^*$ and the design expectation of $\gamma\left(|X_{ij} - X_{i'j'}|, |Y_{ij} - Y_{i'j'}|\right)$. In general, analytical evaluation of these quantities is difficult, but Monte Carlo evaluation is straightforward. Note that drawing repeated MC samples from D is far simpler than drawing repeated realizations of the stochastic process $\{Z(x, y)\}$. See Ripley (1981, §2.5) for some discussion of simulating spatial processes.

Figure 3 shows the effect of the Markov chain sampling parameter ϕ and the spatial dependence parameter δ on the total variance, for an autocorrelated superpopulation with no trend and with $\sigma = 1$, $a = b = 10$ and $I = J = 10$. The contours show the ratio of the total variance under the MC design with $\phi_1 = \phi_2 = \phi$ to the total variance under unaligned

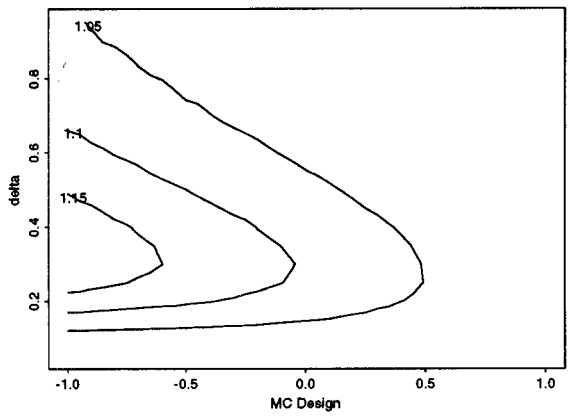

Figure 3: *Ratio of total variance under MC design to total variance under SY. Model has $\alpha = \beta = 0$, $\sigma = 1$. Each design uses $a = b = 10$ and $I = J = 10$. Design expectation of model variance was computed over 1000 independent replications of the sample for each MC design ($\phi_1 = \phi_2 = -1.00, -0.95, \ldots, 0.95, 1.00$) with each value of the exponential ACVF parameter ($\delta = 0.05, 0.10, \ldots, 0.90, 0.95$).*

systematic sampling. Design expectations and variances were computed over 1000 independent replications of the sample for each design.

For a broad range of δ values, MC designs with high positive values of ϕ are the most efficient designs, having total variances within 5% of the total variance under SY. Other MC designs, including ST and BA, are less efficient. As δ increases, the spatial dependence disappears, and all designs have the same efficiency.

As δ decreases, all designs again have the same efficiency. This phenomenon was pointed out by Hájek (1961), who used it as a counterexample to a conjecture in Zubrzycki (1958) that SY was more efficient than ST under certain conditions. The result is somewhat surprising, since in one dimension a non-negative, nonincreasing and convex autocovariance function implies that SY will be the most efficient equal probability design (Cochran 1946, Hájek 1959, Bellhouse 1988). This kind of optimality result can be extended to plane sampling only under special conditions (e.g., Quenouille 1949, Dalenius, Hájek and Zubrzycki 1960, Bellhouse 1977, Iachan 1985).

Figure 4 shows the effect of the spatial dependence parameter δ on the total variance for a population with $\alpha = \beta = 0.025$, and $\sigma = 1$ under a variety of MC designs with $a = b = 10$ and $I = J = 10$.

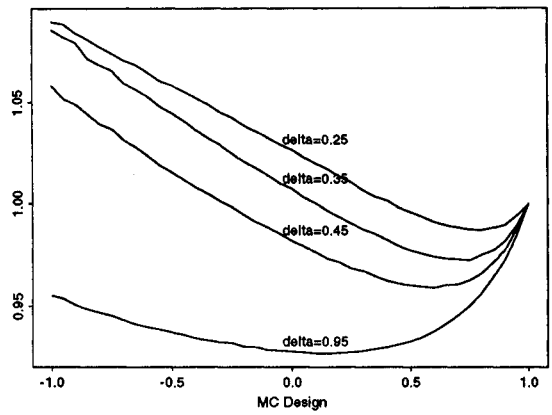

Figure 4: *Effect of the exponential ACVF parameter δ on the ratio of total variance under MC design to total variance under SY. Model has $\alpha = \beta = 0.025$, $\sigma = 1$. Each design uses $a = b = 10$ and $I = J = 10$. Design expectations and variances were computed over 1000 independent replications of the sample for each MC design ($\phi_1 = \phi_2 = -1.00, -0.95, \ldots, 0.95, 1.00$).*

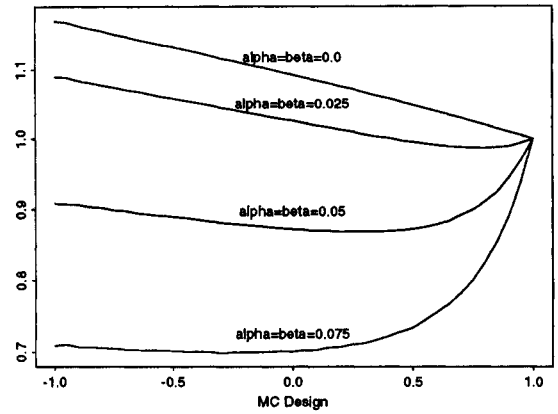

Figure 5: *Effect of the trend parameters α and β on the ratio of total variance under MC design to total variance under SY. Model has $\delta = 0.25$, $\sigma = 1$. Each design uses $a = b = 10$ and $I = J = 10$. Design expectations and variances were computed over 1000 independent replications of the sample for each MC design ($\phi_1 = \phi_2 = -1.00, -0.95, \ldots, 0.95, 1.00$).*

Design expectations and variances were computed over 1000 independent replications of the sample for each design. As δ decreases, the spatial dependence strengthens, and SY becomes relatively more efficient. As δ increases, the spatial dependence disappears, the trend dominates, and the optimal design shifts away from SY toward BA.

Figure 5 shows the effect of the trend parameters α and β on the total variance for a population with $\delta = 0.25$ and $\sigma = 1$ under a variety of MC designs with $a = b = 10$ and $I = J = 10$. Design expectations and variances were computed over 1000 independent replications of the sample for each design. As α and β decrease, the spatial trend weakens, and SY becomes relatively more efficient. As α and β increase, the trend dominates, and the optimal design shifts away from SY toward BA.

In all cases considered in Figures 3–5, MC designs with high positive values of ϕ which are strictly less than one are never far from optimal. For a single study variable of interest, z, with a given trend function and ACVF, an optimal MC design could be selected. In applications, however, many study variables are of interest, and their trends and ACVF's are unknown. This is the case in the Crawford County soil mapping project. In these circumstances, a reasonable procedure is to choose large positive values of ϕ which are strictly less than one, e.g., $\phi_1 = \phi_2 = 0.75$, the values used for Crawford County.

Acknowledgements

This research was partly supported by Cooperative Agreement 68–3A75–5–72 with the National Resources Conservation Service.

References

Bellhouse, D.R. (1977). Some optimal designs for sampling in two dimensions. *Biometrika* **64**, 605–611.

Bellhouse, D.R. (1981). Spatial sampling in the presence of a trend. *J. Statistical Planning and Inference* **5**, 365–375.

Bellhouse, D.R. (1988). Systematic sampling. In: P.R. Krishnaiah and C.R. Rao, eds., *Handbook of Statistics*, Vol. 6. North-Holland, Amsterdam, pp. 125–145.

Breidt, F. J. (1995). Markov chain designs for one-per-stratum sampling. *Survey Methodology* **21**, 63–70.

Cochran, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Statist.* **17**, 164–177.

Cochran, W.G. (1977). *Sampling Techniques,* 3rd ed. Wiley, New York.

Cressie, N.A.C. (1991). *Statistics for Spatial Data.* Wiley, New York.

Dalenius, T., Hájek, J. and Zubrzycki, S. (1960). On plane sampling and related geometrical problems. *Proc. 4th Berkeley Symp.* **1**, 125–150.

Das, A.C. (1950). Two dimensional systematic sampling and the associated stratified and random sampling. *Sankhyā* **10**, 95–108.

Hájek, J. (1959). Optimum strategy and other problems in probability sampling. *Časopis Pro Pěstování Matematiky* **84**, 387–423.

Hájek, J. (1961). Concerning relative accuracy of stratified and systematic sampling in a plane. *Colloquium Mathematicum* **8**, 133–134.

Iachan, R. (1985). Plane sampling. *Statist. and Prob. Letters* **3**, 151–159.

Madow, W.G. and Madow, L.H. (1944). On the theory of systematic sampling, I. *Ann. Math. Statist.* **15**, 1–24.

Matérn, B. (1947). Methods of estimating the accuracy of line and sample plot surveys. *Medd. fr. Statens Skogsforsknings Institut* **36**, 1–138.

Murthy, M.N. (1967). *Sampling Theory and Methods.* Statistical Publishing Society, Calcutta.

Quenouille, M.H. (1949). Problems in plane sampling. *Ann. Math. Statist.* **10**, 355–375.

Ripley, B.D. (1981). *Spatial Statistics*, Wiley, New York.

Weiss, G. (1975). Time-reversibility of linear stochastic processes. *J. Appl. Prob.* **12**, 831–836.

Zubrzycki, S. (1958). Remarks on random, stratified and systematic sampling in a plane. *Colloquium Mathematicum* **6**, 251–264.