THIELE CENTRE
FOR APPLIED MATHEMATICS IN NATURAL SCIENCE

Systematic sampling with errors in sample locations

Johanna Ziegel, Adrian Baddeley,
Karl-Anton Dorph-Petersen and Eva B. Vedel Jensen

Research Report

No. 10 | July 2008

Systematic sampling with errors in sample locations

Johanna Ziegel* Adrian Baddeley† Karl-Anton Dorph-Petersen‡
Eva B. Vedel Jensen§

Abstract

Systematic sampling of points in continuous space is widely used in microscopy and spatial surveys. Classical theory provides asymptotic expressions for the variance of estimators based on systematic sampling as the grid spacing decreases. However, the classical theory assumes the sample grid is exactly periodic; real physical sampling procedures may introduce errors in the placement of the sample points. This paper studies the effect of errors in sample positioning on the variance of estimators. First we sketch a general approach to variance analysis using point process methods. We then analyse three different models for the error process, calculate exact small-sample variances, and derive asymptotic variances. Errors in the placement of sample points can lead to substantial inflation of the variance, dampening of 'Zitterbewegung' effects, and a slower order of convergence. This suggests that current practice in some areas of microscopy may be based on over-optimistic predictions of estimator accuracy.

1 Introduction

Systematic sampling in continuous space is a useful technique in stereology, in ecological survey, and in other spatial sciences, see Baddeley & Jensen (2005) and references therein. In one dimension, a systematic sample is a grid of equally-spaced sample points, with fixed spacing t, randomly shifted with respect to the origin. It may be constructed by setting $x_k = U + kt$ for all integers k, where U is uniformly distributed on $[0, t)$. Systematic sampling can be used to estimate the integral

$$\Theta = \int_{\mathbb{R}} f(x)\,\mathrm{d}x$$

*ETH Zurich, Department of Mathematics, Rämistrasse 101, 8092 Zurich, Switzerland; johanna.ziegel@math.ethz.ch

†School of Mathematics & Statistics, University of Western Australia, WA 6009 Perth, Western Australia;adrian@maths.uwa.edu.au

‡Centre for Psychiatric Research, Aarhus University Hospital, 8240 Risskov, Denmark; karl-anton@dorph-petersen.dk

§Thiele Center, Department of Mathematical Sciences, University of Aarhus, 8000 Aarhus C, Denmark; eva@imf.au.dk

of any integrable function f, using the unbiased estimator

$$\widehat{\Theta} = t \sum_k f(x_k).$$

Similarly in two or three dimensions, a systematic sample is a randomly shifted regular grid of points with fixed geometry; the integral of any integrable function f can be estimated by summing the function values at the sample points and multiplying by the area or volume of one tile in the grid. Such estimators were already known in the nineteenth and the early twentieth century (Delesse (1847, 1848); Crofton (1885); Rosiwal (1898); Steinhaus (1929, 1954); Thomson (1930); Glagolev (1933)). Important early theoretical work on the performance of random grids and their relation to systematic sampling can be found in Moran (1966, 1968), see also Jones (1948).

A simple geometric example of systematic sampling in one dimension concerns the estimation of the volume of a bounded object in $\mathbb{R}^3$. Here, we may let $f(x)$ be the area of the intersection of the object with a horizontal plane at height $x \in \mathbb{R}$. The resulting sampling design is the 'egg-slicer design'. The corresponding estimator is sometimes called 'the Cavalieri estimator', see (Baddeley & Jensen, 2005, p. 155), due to Cavalieri's principle, stating that two solid objects which have equal cross-sectional areas on all horizontal planes must have equal volumes. There are important applications of the Cavalieri estimator throughout biological science.

Systematic sampling, as formulated above, has since the mid-1980's experienced a renaissance in stereology. The main practical purpose of stereology is to estimate quantitative parameters of a spatial object from microscopical images of sections through the object. The aim of stereology is not to reconstruct the three-dimensional geometry of the object. Modern stereology is, however, not confined to two-dimensional sections. The same principles apply to a two-dimensional projection, a three-dimensional volume image, a cylindrical core sample, etc. Stereological methods are nowadays powerful tools in many fields of science. A very recent account of the mathematical and statistical foundations of stereology and the closely related field of stochastic geometry can be found in Weil & Schneider (2008).

Estimation of the precision of $\widehat{\Theta}$ based on systematic sampling is a question of great current interest, see the recent volume of Journal of Microscopy, Mattfeldt (ed.) (2006), devoted to this topic. There is extensive literature on the representation and approximation of the variance of $\widehat{\Theta}$, see (Baddeley & Jensen, 2005, Chapter 13) and references therein. Matheron (1965, 1970) proposed to study this variance by means of the transitive theory, which provides a variance representation based on the Euler-MacLaurin formula; see also Cruz-Orive (1989). The variance can be expressed as the sum of the *extension term*, which gives the overall trend of the variance, the *'Zitterbewegung'*, which oscillates around zero, and higher order terms. The extension term is used to estimate the variance of $\widehat{\Theta}$. Matheron worked with the fundamental fact that the extension term depends on the behaviour of the *geometric covariogram*

$$g(z) := \int f(z+x)f(x)dx, \quad x \in \mathbb{R},$$

of f near the origin. In Kiêu (1997) and Kiêu et al. (1999), a general form of the Euler-MacLaurin formula was derived, which reveals the connection between the

variance of $\widehat{\Theta}$ and the jumps of the function f and its derivatives. See also Arnau & Cruz-Orive (1998).

Two main findings of the classical theory are that, as the sample spacing decreases, the variance of $\widehat{\Theta}$ decreases at a faster rate than we might have expected ('superefficiency') and that the variance does not decrease monotonically but fluctuates between high and low values ('Zitterbewegung' or oscillation) because of resonance effects (Baddeley & Jensen, 2005, Chapter 13).

However, the classical theory assumes that the grid points are exactly periodic. In real sampling procedures, which may involve physically placing the sample points or physically cutting a material, the positions of the sample points may be subject to error. It appears to be unknown what effect these errors might have on the variance of $\widehat{\Theta}$.

The key idea of the present paper is to describe the noisy sampling points by means of a point process Φ. This approach has earlier been used with success in Pache et al. (1993) and Baddeley et al. (2006). The estimator to be considered takes the following form

$$\widehat{\Theta} = \tau \sum_{x \in \Phi} f(x),$$

where $\tau > 0$ is a suitable normalization constant. The estimator $\widehat{\Theta}$ will be denoted a *generalized Cavalieri estimator*.

We will mainly study the case where the function f is defined on the line. There are a number of important examples of this sampling situation in stereology, the most prominent ones are volume estimation from measurement of section areas and number estimation from disector counts (Baddeley & Jensen, 2005, pp. 155 and 258).

We study three models for errors in sample locations. They are inspired by recent stereological studies of brain structure, see Dorph-Petersen (1999); Dorph-Petersen et al. (2005, 2007); Sweet et al. (2005). The models are formulated here so that they have general probabilistic interest. In the first model, called *perturbed systematic sampling*, it is assumed that the sampling points are perturbed by independent and identically distributed errors D_k, $k \in \mathbb{Z}$. This model was briefly discussed in Baddeley et al. (2006). Under the second model, called *systematic sampling with cumulative error*, the increments between successive sampling points are independent and identically distributed. The last model, *systematic sampling with independent p-thinning*, applies if observations are lost independently of each other with probability p. Each of the two first models may be combined with p-thinning. The models of perturbed systematic sampling and systematic sampling with independent p-thinning have earlier been discussed in another spatial sampling context in Lund & Rudemo (2000) under the names of displacement and thinning, respectively.

One of the key results of this paper is that the effect of error in sample locations on the variance of the estimator $\widehat{\Theta}$ may be substantial. We assess the asymptotic variance of $\widehat{\Theta}$ as $t \to 0$ in the case of systematic sampling with errors. For perturbed systematic sampling the asymptotic variance can be determined, using the transitive theory and its further development in Kiêu (1997) and Kiêu et al. (1999). The order of magnitude of the asymptotic variance depends on the smoothness of f. For systematic sampling with cumulative error, the asymptotic behavior of the variance

is dominated by the term $tCg(0)$, where g is the covariogram of f and C is a model constant. We use renewal theory to show this result. It is remarkable that for both perturbation mechanisms the *'Zitterbewegung'*-effect is asymptotically negligible as $t \to 0$. We also derive the asymptotic variance under p-thinning combined with either perturbed systematic sampling or systematic sampling with cumulative error. In both cases, the variance behaves like $tCg(0)$, where C is a constant depending on whether perturbed systematic sampling or systematic sampling with cumulative error applies.

Section 2 contains preliminaries about point processes. In Section 3, we show under mild regularity conditions that $\widehat{\Theta}$ is unbiased and derive an expression for the variance of $\widehat{\Theta}$ in terms of the covariogram g of f and the second order reduced factorial moment measure of Φ. In Section 4, the three types of models for noisy sampling points are described in more detail. The density of the second order reduced factorial moment measure of Φ is derived in each of the three cases and the resulting expression for $\mathrm{var}(\widehat{\Theta})$ is given. Section 5 contains the study of the asymptotic variance as $t \to 0$, while Section 6 gives an example of the effect of errors in sampling locations. Section 7 discusses the obtained results.

2 Preliminaries

This section introduces the basics of point process theory needed in the sequel. For a detailed exposition, see Daley & Vere-Jones (1988) and Stoyan et al. (1995). Let $\mathcal{B}^d$ denote the Borel σ-algebra on $\mathbb{R}^d$. All point processes considered are assumed to be simple.

Definition 2.1 (Moment measures)**.** Let Φ be a point process on $\mathbb{R}^d$. For $A_1, \ldots, A_k \in \mathcal{B}^d$ we define

$$M_k(A_1 \times \cdots \times A_k) := \mathbb{E}\left(\Phi(A_1) \cdots \Phi(A_k)\right)$$

and

$$M_{[k]}(A_1 \times \cdots \times A_k) := \mathbb{E}\left(\sum\nolimits^{\neq}_{x_1,\ldots,x_k \in \Phi} \mathbf{1}_{A_1}(x_1) \cdots \mathbf{1}_{A_k}(x_k)\right),$$

where $\Phi(A_i)$ is the number of points in A_i, and the symbol $\sum^{\neq}_{x_1,\ldots,x_k \in \Phi}$ indicates summation over all k-tuples in Φ^k such that the components are pairwise different. If M_k and $M_{[k]}$ are finite on bounded sets, they extend to uniquely defined symmetric measures on the product σ-algebra $\otimes_{i=1}^k \mathcal{B}^d = \mathcal{B}^d \otimes \cdots \otimes \mathcal{B}^d$. In this case, M_k is called the *k-th order moment measure* and $M_{[k]}$ the *k-th order factorial moment measure* of Φ.

The first order moment measure M_1 is also called the *intensity* measure of the process. Note that $M_1 = M_{[1]}$. If M_k or $M_{[k]}$ have densities with respect to Lebesgue measure on $(\mathbb{R}^d)^k$ we denote them by m_k, $m_{[k]}$, respectively.

Definition 2.2 (Stationarity)**.** A point process Φ on $\mathbb{R}^d$ is *strictly stationary* if, for all $x \in \mathbb{R}^d$ and all $k \in \mathbb{N}$, $A_1, \ldots, A_k \in \mathcal{B}^d$, $n_1, \ldots, n_k \in \mathbb{N}$,

$$\mathrm{pr}(\Phi(A_i) \le n_i, i = 1, \ldots, k) = \mathrm{pr}(\Phi(T_x A_i) \le n_i, i = 1, \ldots, k),$$

where the shift operator T_x is defined as $T_x A := x + A$.

Definition 2.3 (Moment Stationarity)**.** A point process Φ is *k-th order stationary* if its k-th order moment measure exists, and for each $j = 1, \ldots, k$, bounded sets $A_1, \ldots, A_j \in \mathcal{B}^d$, and $x \in \mathbb{R}^d$,

$$M_j(T_x A_1 \times \cdots \times T_x A_j) = M_j(A_1 \times \cdots \times A_j).$$

It can be shown, see (Daley & Vere-Jones, 1988, p. 355), that in the case of simple point processes the conditions on M_j for $j < k$ are redundant. If a process is first order stationary then the first order moment measure is a finite positive multiple $m \equiv m_1(x)$ of the Lebesgue measure $\mathcal{L}^d$ on $\mathbb{R}^d$. The proportionality constant m is called the *intensity* of the process.

For a k-th order stationary point process, the k-th order moment measure M_k can be factorized as follows. For any measurable function f on $(\mathbb{R}^d)^k$ with bounded support we have

$$\begin{aligned}\int_{(\mathbb{R}^d)^k} f(x_1, \ldots, x_k) M_k(dx_1 \times \cdots \times dx_k) \\ = \int_{\mathbb{R}^d} \int_{(\mathbb{R}^d)^{k-1}} f(x, x + y_1, \ldots, x + y_{k-1}) \tilde{M}_k(dy_1 \times \cdots \times dy_{k-1})\, dx, \quad (1)\end{aligned}$$

where $\tilde{M}_k$ is a reduced measure on $(\mathbb{R}^d)^{k-1}$, called the k-th order *reduced* moment measure, see (Daley & Vere-Jones, 1988, Corollary 10.4.IV). The k-th order *reduced* factorial moment measure $\tilde{M}_{[k]}$ is defined in the analogous way. Note that the measure disintegration in equation (1) can be performed for any boundedly finite Borel measure on $(\mathbb{R}^d)^k$, which is invariant under diagonal shifts, see (Daley & Vere-Jones, 1988, Lemma 10.4.III). In analogy with the notation above we use $\tilde{m}_k$ and $\tilde{m}_{[k]}$, respectively, to denote the density of the reduced measure if it exists.

If the density $m_{[2]}$ of the second order factorial moment measure $M_{[2]}$ exists, then $m_{[2]}(x, y)dxdy$ may be interpreted as the probability that two neighborhoods of x and y of volume dx and dy, respectively, each contain a point from the point process. The function $\rho(x, y) = m_{[2]}(x, y)/(m_1(x)m_1(y))$ is usually called the pair correlation function. The process is second order stationary if it is first order stationary and $\rho(x, y)$ depends only on $x-y$ and is locally integrable. This was used as the definition of second order stationarity in Baddeley et al. (2006). Finally note that a strictly stationary point process for which the k-th order moment measure exists is k-th order stationary.

For a function $f : \mathbb{R}^d \to \mathbb{R}$ we define $\check{f} : \mathbb{R}^d \to \mathbb{R}$ by $\check{f}(x) = f(-x)$. The convolution of two functions $f, g : \mathbb{R}^d \to \mathbb{R}$ is denoted by $f * g$. Furthermore we define the *k-fold convolution* of f by $f^{k*} = f^{(k-1)*} * f$, $f^{1*} = f$, for $k \geq 2$. A function f belongs to the space of *locally integrable* functions L^1_{loc}, if for all compact sets K the function $\mathbf{1}_K f$ is Lebesgue integrable. The space of integrable (essentially bounded) functions is L^1 (L^∞) with norm $\|\cdot\|_1$ ($\|\cdot\|_\infty$). The space of p-times continuously differentiable functions is denoted by C^p. We write C^p_0 if they are also required to have compact support.

3 First and second order properties

The following two theorems, which are generalizations of results in Baddeley et al. (2006), allow us to study the first and second order properties of estimators based on systematic sampling with errors. Let f be a measurement function, i.e. an integrable function with bounded support on $\mathbb{R}$. Define $\Theta := \int_{\mathbb{R}} f(x)dx$.

Theorem 3.1. *Suppose that Φ is a first order stationary point process with intensity $m_1(x) = m$, where $m > 0$. Then the generalized Cavalieri estimator $\widehat{\Theta} = \tau \sum_{x\in\Phi} f(x)$ with $\tau = 1/m$ is an unbiased estimator of Θ.*

Proof. First order stationarity yields that $M_1 = m\mathcal{L}$. By Campbell's Theorem (Daley & Vere-Jones, 1988, p. 188) we have

$$\mathrm{E}\big(\widehat{\Theta}\big) = \tau\,\mathrm{E}\Big(\sum_{x\in\Phi} f(x)\Big) = \tau \int_{\mathbb{R}} f(x)M_1(dx) = \tau m \int_{\mathbb{R}} f(x)dx = \Theta,$$

by the choice of τ. □

Theorem 3.2. *Suppose that Φ is a second order stationary point process with intensity $m_1(x) = m > 0$ and second order reduced factorial moment measure $\tilde{M}_{[2]}$. Choosing $\tau = 1/m$ as above the variance of $\widehat{\Theta}$ is given by*

$$\mathrm{var}\big(\widehat{\Theta}\big) = \frac{g(0)}{m} + \frac{1}{m^2}\int_{\mathbb{R}} g(z)\tilde{M}_{[2]}(dz) - \int_{\mathbb{R}} g(z)dz,$$

where $g(z) = \int_{\mathbb{R}} f(z+x)f(x)dx$ is the geometric covariogram of f.

Proof. Using the factorization of the second order factorial moment measure as given by (1), Campbell's Theorem and Fubini we obtain

$$\begin{aligned}
\mathrm{var}\big(\widehat{\Theta}\big) &= \mathrm{E}\big(\widehat{\Theta}^2\big) - \mathrm{E}\big(\widehat{\Theta}\big)^2 \\
&= \frac{1}{m^2}\,\mathrm{E}\Big(\sum_{x,y\in\Phi,\, x\neq y} f(x)f(y) + \sum_{x\in\Phi} f^2(x)\Big) - \Theta^2 \\
&= \frac{1}{m^2}\left(\int_{\mathbb{R}^2} f(x)f(y)M_{[2]}(dx,dy) + \int_{\mathbb{R}} f^2(x)m\,dx\right) - \Theta^2 \\
&= \frac{1}{m^2}\int_{\mathbb{R}}\int_{\mathbb{R}} f(x)f(x+y)\tilde{M}_{[2]}(dy)\,dx + \frac{g(0)}{m} - \Theta^2 \\
&= \frac{g(0)}{m} + \frac{1}{m^2}\int_{\mathbb{R}} g(y)\tilde{M}_{[2]}(dy) - \Theta^2.
\end{aligned}$$

Again by Fubini we have $\int_{\mathbb{R}} g(z)dz = \int_{\mathbb{R}}\int_{\mathbb{R}} f(y)f(y+z)dzdy = \Theta^2$. □

It is possible to extend the above results to higher dimensions, see the Appendix.

4 Models for Φ

In this section we discuss three different models for systematic sampling with errors on $\mathbb{R}$.

4.1 Perturbed systematic sampling

The first model we address is called *perturbed systematic sampling*. This model was already considered in Baddeley et al. (2006). We assume that the intended equally spaced sampling points $x_k = U + kt$ are perturbed by random errors $(D_k)_{k\in\mathbb{Z}}$, so that the actual locations are $y_k = x_k + D_k$. The random variable U is uniformly distributed on $[0, t)$, where $t > 0$ is the intended spacing of the sampling points. The sequence $(D_k)_{k\in\mathbb{Z}}$ is independent and identically distributed with common density function h, which has bounded support. In order to have a realistic model, it would normally be assumed that h is supported in $[-t/2, t/2]$, which ensures that the sample points y_k are strictly increasing, with probability 1. However, this fact is not used in the sequel unless explicitly stated. In relation to cutting of tissue in stereological studies, perturbed systematic sampling will, for example, be a reasonable model for devices consisting of an array of cutting blades (Gundersen et al., 1988, Figure 7). Slight drift of the blades while cutting will perturb the actual cut around the fixed position of each blade.

Lemma 4.1. *Let Φ be a point process that follows the perturbed systematic sampling model with error density h, which has bounded support. Then, the process $\Phi = (y_k)_{k\in\mathbb{Z}}$ is second order stationary, the intensity measure M_1 is equal to $\frac{1}{t}\mathcal{L}$ and the second order reduced factorial moment measure $\tilde{M}_{[2]}$ has density*

$$\tilde{m}_{[2]}(y) = \frac{1}{t} \sum_{n\in\mathbb{Z},\, n\neq 0} h * \check{h}(-y + nt).$$

Remark. Note that the convolution $h * \check{h}$ is the density of $D_k - D_l$ for $k \neq l$.

Proof. The density of $U + D_k$ is

$$f_{U+D_k}(y) = \frac{1}{t} \int_{\mathbb{R}} h(x)\mathbf{1}_{[y-t,y]}(x)\, dx.$$

Let $a < b < \infty$, we then obtain

$$\begin{aligned}
M_1([a, b)) &= \mathrm{E}\left(\Phi([a, b))\right) \\
&= \sum_{k\in\mathbb{Z}} \mathrm{E}\left(\mathbf{1}_{[a,b)}(kt + D_k + U)\right) \\
&= \sum_{k\in\mathbb{Z}} \int_{\mathbb{R}} \mathbf{1}_{[a-kt,b-kt)}(y) \frac{1}{t} \int_{\mathbb{R}} h(x)\mathbf{1}_{[y-t,y]}(x)\, dx\, dy \\
&= \frac{1}{t} \sum_{k\in\mathbb{Z}} \int_{\mathbb{R}} \int_{\mathbb{R}} h(x)\mathbf{1}_{[z-(k+1)t,z-kt]}(x)\mathbf{1}_{[a,b)}(z)\, dx\, dz \\
&= \frac{1}{t} \int_a^b \int_{\mathbb{R}} h(x)\, dx\, dz = \frac{1}{t}(b - a) = \frac{1}{t}\mathcal{L}([a, b)).
\end{aligned}$$

The joint density of $(U + D_k, U + D_l)$ is

$$f_{U+D_k,U+D_l}(y_1, y_2) = \frac{1}{t} \int_{\mathbb{R}} \mathbf{1}_{[0,t)}(u)h(y_1 - u)h(y_2 - u)du.$$

For $B_1, B_2 \in \mathcal{B}(\mathbb{R})$, we have

$$\begin{aligned} M_{[2]}(B_1 \times B_2) &= \mathrm{E}\Big(\sum_{x,y\in\Phi,\, x\neq y} \mathbf{1}_{B_1}(x)\mathbf{1}_{B_2}(y)\Big) \\ &= \sum_{k,l\in\mathbb{Z},\, k\neq l} \mathrm{E}\left(\mathbf{1}_{B_1-kt}(U+D_k)\mathbf{1}_{B_2-lt}(U+D_l)\right) \\ &= \sum_{k,l\in\mathbb{Z},\, k\neq l} \frac{1}{t}\int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbf{1}_{B_1}(y_1+kt)\mathbf{1}_{B_2}(y_2+lt) \\ &\qquad \mathbf{1}_{[0,t)}(u)h(y_1-u)h(y_2-u)\,du\,dy_1\,dy_2 \\ &= \sum_{n\in\mathbb{Z},\, n\neq 0} \frac{1}{t}\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbf{1}_{B_1}(z_1)\mathbf{1}_{B_2}(z_2) \\ &\qquad \times \int_{\mathbb{R}} \sum_{k\in\mathbb{Z}} \mathbf{1}_{(z_1-(k+1)t, z_1-kt]}(v) \\ &\qquad \times h(v)h(z_2-z_1-nt+v)\,dv\,dz_1\,dz_2, \end{aligned} \tag{2}$$

using the substitutions $z_1 = y_1+kt$, $z_2 = y_2+lt$, $v = z_1-kt-u$ and setting $n = l-k$. Recall that $\check{h}(x) := h(-x)$. We can further simplify equation (2) as follows

$$\begin{aligned} M_{[2]}(B_1 \times B_2) &= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbf{1}_{B_1}(z_1)\mathbf{1}_{B_2}(z_2)\frac{1}{t}\sum_{n\in\mathbb{Z},\, n\neq 0} \check{h} * h(z_2-z_1-nt)\,dz_1\,dz_2 \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbf{1}_{B_1}(z_1)\mathbf{1}_{B_2}(z_1+y)\frac{1}{t}\sum_{n\in\mathbb{Z},\, n\neq 0} \check{h} * h(y-nt)\,dy\,dz_1, \end{aligned}$$

so we obtain that $\tilde{M}_{[2]}$ has density $\tilde{m}_{[2]}(y) = \frac{1}{t}\sum_{n\in\mathbb{Z},\, n\neq 0} \check{h} * h(y-nt)$. It is not difficult to check that $\tilde{m}_{[2]}$ is always in L^1_{loc}, hence the second order reduced factorial moment measure of Φ exists for any density function h with bounded support. □

Applying Theorem 3.2 we obtain the following representation of the variance for the generalized Cavalieri estimator under perturbed systematic sampling

$$\begin{aligned} \mathrm{var}\big(\widehat{\Theta}\big) &= tg(0) + t\sum_{n\in\mathbb{Z},\, n\neq 0}\int_{\mathbb{R}} g(z)\check{h} * h(z-nt)\,dz - \int_{\mathbb{R}} g(z)dz \\ &= tg(0) + t\sum_{n\in\mathbb{Z},\, n\neq 0} g * \check{h} * h(nt) - \int_{\mathbb{R}} g(z)dz. \end{aligned} \tag{3}$$

4.2 Systematic sampling with cumulative error

The second model we consider is called *systematic sampling with cumulative error*. We assume that the actual locations $(y_k)_{k\in\mathbb{Z}}$ of the sampling points are such that the increments $w_k = y_k - y_{k-1}$, $k \in \mathbb{Z}$, are independent and identically distributed with density $h : \mathbb{R}_+ \to \mathbb{R}_+$ and finite expectation $t > 0$. We choose the 'starting distribution' $\tilde{H}$ for y_1 as follows

$$\tilde{H}(x) = \frac{1}{t}\int_0^x (1-H(y))\,dy,$$

where H is the distribution function of h. Applying (Daley & Vere-Jones, 1988, Theorem 12.3.II) we then have that $\Phi = (y_k)_{k\in\mathbb{Z}}$ is a strictly stationary point process with finite intensity $m = \frac{1}{t}$. Systematic sampling with cumulative error is appropriate if the sampling procedure works like a meat slicer, where each successive section is cut by advancing the material towards a stop plate a fixed distance from the slicing blades. If the 'block advance' is slightly variable, e.g. due to elasticity of the material leading to a variable degree of compression, then we get cumulative errors.

Lemma 4.2. *Let Φ be a point process that follows the systematic sampling with cumulative error model with error density h with mean $t > 0$. Then, $\Phi = (y_k)_{k\in\mathbb{Z}}$ is second order stationary, the intensity measure M_1 is equal to $\frac{1}{t}\mathcal{L}$ and the second order reduced factorial moment measure $\tilde{M}_{[2]}$ has density*

$$\tilde{m}_{[2]}(x) = \frac{1}{t}\sum_{k=1}^{\infty}(h^{k*}(x) + \check{h}^{k*}(x)),$$

where h^{k*} denotes the k-fold convolution of h. The density $\tilde{m}_{[2]}$ is locally integrable.

Proof. Denote by $\overset{\circ}{M}_1$ the first moment measure of the Palm distribution $\mathcal{P}_0$ of Φ. We have

$$\overset{\circ}{M}_1(A) - \delta_0(A) = \mathrm{E}_{\mathcal{P}_0}\left(\Phi(A\backslash\{0\})\right), \quad A \in \mathcal{B}. \tag{4}$$

Define $y'_0 := 0$, $y'_k := \sum_{i=1}^{k} w_i$ and $y'_{-k} := \sum_{i=0}^{k-1} -w_{-i}$ for $k \geq 1$. We can calculate the right-hand side of equation (4) in the following way

$$\begin{aligned}\mathrm{E}_{\mathcal{P}_0}\left(\Phi(A\backslash\{0\})\right) &= \mathrm{E}\Big(\sum_{k=1}^{\infty}\mathbf{1}\{y'_k \in A\} + \sum_{k=1}^{\infty}\mathbf{1}\{y'_{-k} \in A\}\Big)\\ &= \int_A \sum_{i=1}^{\infty}(h^{i*}(x) + \check{h}^{i*}(x))\,dx.\end{aligned}$$

The term $\sum_{k=1}^{\infty} h^{k*}(x)$ is the renewal density of a renewal process with holding times that are independent and identically distributed with density h. Standard renewal theory yields that $\sum_{k=1}^{\infty} h^{k*}(x)$ is locally integrable, so in particular the series converges for almost all $x \in \mathbb{R}$; see for example (Daley & Vere-Jones, 1988, Chapter 4). The same argument holds for $\sum_{k=1}^{\infty} \check{h}^{k*}(x)$, where we have to consider a renewal process with reversed time. Therefore $\tilde{m}_{[2]}(x) \in L^1_{loc}$, which implies the existence of the first moment measure $\overset{\circ}{M}_1$ of the Palm distribution. By (Daley & Vere-Jones, 1988, Proposition 12.2.V) this implies the existence of the second order factorial moment measure $M_{[2]}$ and

$$\tilde{M}_{[2]}(A) = m(\overset{\circ}{M}_1(A) - \delta_0(A)).$$

Plugging in $m = \frac{1}{t}$ yields the claimed formula. □

Using Lemma 4.2, Theorem 3.2 yields the following variance for the generalized Cavalieri estimator under systematic sampling with cumulative error

$$\mathrm{var}\big(\widehat{\Theta}\big) = tg(0) + t\int_{\mathbb{R}}\sum_{k=1}^{\infty} g(z)(h^{k*}(z) + \check{h}^{k*}(z))dz - \int_{\mathbb{R}} g(z)dz. \tag{5}$$

4.3 Systematic sampling with independent p-thinning

The third model we want to address is *systematic sampling with independent p-thinning*. Suppose we have sampling points at locations $\Psi = (y_k)_{k\in\mathbb{Z}}$, which form a second order stationary point process, the so-called center process, with intensity $m = \frac{1}{t}$ and second order reduced factorial density $\tilde{m}^c_{[2]}$. Let $p > 0$ be the probability that the value of f cannot be determined at location y_k. Let $(U_k)_{k\in\mathbb{Z}}$ be a sequence of independent and identically distributed uniform random variables on $[0, 1]$ and independent of $(y_k)_{k\in\mathbb{Z}}$. The resulting point process is

$$\Phi = \{y_k \,:\, U_k > p\}\,.$$

Using (Daley & Vere-Jones, 1988, Proposition 8.2.IV), we derive the following expressions for the intensity measure M_1 and the second order factorial measure $M_{[2]}$ of Φ

$$\begin{aligned}
M_1(A) &= \int_{\mathbb{R}} (1-p)\mathbf{1}_A(x) m\, dx = (1-p)m\mathcal{L}(A), \quad A \in \mathcal{B}(\mathbb{R}),\\
M_{[2]}(A\times B) &= \int_{\mathbb{R}}\int_{\mathbb{R}} (1-p)^2\mathbf{1}_A(x)\mathbf{1}_B(y) M^c_{[2]}(dx, dy)\\
&\quad + \int_{\mathbb{R}} ((1-p)\mathbf{1}_A(x)\mathbf{1}_B(x) - (1-p)\mathbf{1}_{A\cap B}(x))m\, dx\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} (1-p)^2\mathbf{1}_A(x)\mathbf{1}_B(x+y)\tilde{m}^c_{[2]}(y)\, dy\, dx, \quad A, B \in \mathcal{B}(\mathbb{R}).
\end{aligned}$$

Hence, the intensity of Φ is $(1-p)m$ and the second order reduced factorial moment measure of Φ has density $(1-p)^2\tilde{m}^c_{[2]}(y)$. Therefore we obtain the following representation of the variance for the generalized Cavalieri estimator under systematic sampling with independent p-thinning

$$\operatorname{var}\big(\widehat{\Theta}\big) = \frac{g(0)}{(1-p)m} + \frac{1}{m^2}\int_{\mathbb{R}} g(z)\tilde{m}^c_{[2]}(z)\, dz - \int_{\mathbb{R}} g(z)dz. \tag{6}$$

Note that only the first term on the right-hand side is different from the formula for the variance of the generalized Cavalieri estimator based on the process Ψ, see Theorem 3.2.

5 Limiting behavior

5.1 Perturbed systematic sampling

In this section we study the asymptotic behavior of the variance for perturbed systematic sampling. The measurement function will always be denoted by f, the density of the error distribution by h. For perturbed systematic sampling we can

rewrite equation (3) for the variance as follows:

$$\begin{aligned}\operatorname{var}\big(\widehat{\Theta}\big) &= tg(0) + t \sum_{n\in\mathbb{Z},\, n\neq 0} g * \check{h} * h(nt) - \int_{\mathbb{R}} g(z)dz \\ &= t(g(0) - g * \check{h} * h(0)) + t\sum_{n\in\mathbb{Z}} g * \check{h} * h(nt) - \int_{\mathbb{R}} g * \check{h} * h(z)dz \\ &\quad + \int_{\mathbb{R}} g * \check{h} * h(z)dz - \int_{\mathbb{R}} g(z)dz.\end{aligned}$$

Using Fubini's theorem it is easy to see that the last two terms in the above equation cancel each other. Recall that the geometric covariogram g is defined as $g(z) = \int_{\mathbb{R}} f(x)f(x+z)dx$, where f is the measurement function. Using this definition, it is easy to check that

$$g * \check{h} * h(z) = \int_{\mathbb{R}} \left(f * h(x+z) f * h(x)\right) dx.$$

Define $F := f * h$. We now consider F as the measurement function. Its covariogram is $G(z) = g * \check{h} * h(z)$. Let $\hat{W} := t\sum_{j\in\mathbb{Z}} F(U + jt)$, where U is uniformly distributed on $[0, t)$. We then obtain

$$\operatorname{var}(\hat{W}) = t\sum_{j\in\mathbb{Z}} G(jt) - \int_{\mathbb{R}} G(z)dz, \tag{7}$$

see (Baddeley & Jensen, 2005, (13.18)). As we want to study the asymptotic behavior of the variance of $\widehat{\Theta}$ as $t \to 0$, we need to specify how the error density h depends on t. Throughout this section we assume that

$$h_t(x) = \frac{1}{t} h_0\left(\frac{x}{t}\right), \quad x \in \mathbb{R},$$

$t > 0$, where h_0 is a probability density function belonging to the class $\mathcal{C}_K$ of Lebesgue measurable functions with compact support and a finite number of jumps of finite size.

In the proof of the asymptotic variance of $\widehat{\Theta}$ we will use the definitions and properties of piecewise smooth functions as given in Kiêu (1997). In particular, the following definition will be important.

Definition 5.1. For a function $q : \mathbb{R} \to \mathbb{R}$ let

$$s_q(x) := \lim_{y\to x^+} q(y) - \lim_{y\to x^-} q(y), \quad x \in \mathbb{R},$$

where we assume that the limits are defined everywhere. Let $D_q := \operatorname{supp}(s_q)$. The function q is said to be (m, p)-*piecewise smooth*, $m, p \in \mathbb{N}_0$, if

(i) $q^{(l)} \in \mathcal{C}_K$ for all $0 \le l \le m + p$

(ii) $D_{q^{(l)}} = \emptyset$ for $0 \le l < m$.

Thus, an (m,p)-piecewise smooth function has compact support. Furthermore, all its derivatives of order less than m are continuous while derivatives of order m up to $m+p$ have a finite number of jumps of finite size.

Proposition 5.1. *Let f be an $(m,1)$-piecewise smooth measurement function. Then its covariogram g is $(2m+1,1)$-piecewise smooth and the variance of the generalized Cavalieri estimator has the following expansion as $t \to 0$*

$$\begin{aligned}\operatorname{var}\big(\widehat{\Theta}\big) &= t(g(0) - g * \check{h}_t * h_t(0)) \\ &\quad - t^{2m+2} s_{g^{(2m+1)}}(0) \int_{\mathbb{R}} h_0 * \check{h}_0(x) P_{2m+2}(x)dx + o(t^{2m+2}),\end{aligned} \tag{8}$$

where $P_i(\cdot)$ denotes the i-th Bernoulli polynomial as defined in (Knopp, 1996, Paragraph 297). Let $c_k := \int_{\mathbb{R}} |x|^k h_0 * \check{h}_0(x)dx$. For $m = 0$, the first term in (8) is asymptotically equal to the following expression

$$t\big(g(0) - g * \check{h}_t * h_t(0)\big) \sim -t^2 \frac{c_1}{2} s_{g'}(0). \tag{9}$$

If $\operatorname{supp}(h_0) \subseteq [-1/2, 1/2]$ and $m = 0$, equation (8) simplifies to

$$\operatorname{var}\big(\widehat{\Theta}\big) \sim -t^2 \left(\frac{c_2}{2} + \frac{1}{12}\right) s_{g'}(0). \tag{10}$$

For $m \geq 1$, we have

$$t\big(g(0) - g * \check{h}_t * h_t(0)\big) = -t^3 \frac{c_2}{2} g^{(2)}(0) + o(t^3). \tag{11}$$

Remark. Note that by (Kiêu, 1997, Corollary 5.8) $s_{g^{(2m+1)}}(0) \neq 0$. From the definition of g it is clear that for $m \geq 1$ we have $g^{(2)}(0) \neq 0$.

Remark. In Kiêu (1997), the Bernoulli polynomial P_i is denoted by $P_{i,1}$. They are defined as follows. For $x \in [0,1]$ we first define inductively $\tilde{P}_0(x) = 1$, $\tilde{P}_1(x) = x - \frac{1}{2}$ and for $i \geq 2$

$$\begin{aligned}\tilde{P}'_{i+1} &= \tilde{P}_i \\ \tilde{P}_i(0) &= \tilde{P}_i(1).\end{aligned}$$

Then let $P_i(x) = \tilde{P}_i(x - [x])$. So P_i is bounded, 1-periodic and $P'_{i+1} = P_i$ for $i \geq 0$, in particular $P_2(x) = \frac{1}{2}\big((x-[x])(x-[x]-1) + \frac{1}{6}\big)$.

Proof of Proposition 5.1. Suppose that the measurement function f is (m,p)-piecewise smooth with $p \geq 1$. Let $F_t := f * h_t$. Then, it follows from (Kiêu, 1997, Proposition 5.6 and Corollary 5.8) that the covariogram $G_t = g * h_t * \check{h}_t$ of F_t is $(2m+2)$-times continuously differentiable and

$$(g * h_t * \check{h}_t)^{(2m+2)} = g^{(2m+2)} * h_t * \check{h}_t + s_{g^{(2m+1)}} * h_t * \check{h}_t, \tag{12}$$

where $s * q(x) := \sum_a s(a) q(x-a)$ for a function s with finite support and a function q whose support has non-zero Lebesgue measure. In (Kiêu, 1997, Proposition 4.2)

a refined Euler-MacLaurin formula for $(m,1)$-piecewise smooth functions is given. The proof relies on a partial integration formula for piecewise smooth functions. We would like to apply the formula to the right-hand side of (7) with $G = G_t$. This is not directly possible as the error term approximations are only valid, if G does not depend on t, but following (Kiêu, 1997, proof of Proposition 4.2), we obtain, using (12),

$$\begin{aligned}\operatorname{var}(\hat{W}_t) = &-t^{2m+2}\int_{\mathbb{R}} g^{(2m+2)} * h_t * \check{h}_t(x) P_{2m+2}\left(\frac{x}{t}\right) dx \\ &- t^{2m+2} \sum_{a \in D_{g^{(2m+1)}}} s_{g^{(2m+1)}}(a) \int_{\mathbb{R}} h_t * \check{h}_t(x-a) P_{2m+2}\left(\frac{x}{t}\right) dx.\end{aligned} \tag{13}$$

Note that by (Kiêu, 1997, Corollary 5.8), we always have $s_{g^{(2m+1)}}(0) \neq 0$ and, as g is an even function, we obtain $s_{g^{(2m+1)}}(0) = 2g^{(2m+1)}(0^+)$, where $g^{(2m+1)}(0^+) := \lim_{x \to 0^+} g^{(2m+1)}(x)$.

The second term on the right hand side of the above equation can be decomposed as follows

$$\begin{aligned}&t^{2m+2} \sum_{a \in D_{g^{(2m+1)}}} s_{g^{(2m+1)}}(a) \int_{\mathbb{R}} h_t * \check{h}_t(x-a) P_{2m+2}\left(\frac{x}{t}\right) dx \\ &= t^{2m+2} s_{g^{(2m+1)}}(0) \int_{\mathbb{R}} h_0 * \check{h}_0(x) P_{2m+2}(x) dx \\ &\quad + t^{2m+2} \sum_{\substack{a \in D_{g^{(2m+1)}} \\ a \neq 0}} s_{g^{(2m+1)}}(a) \int_{\mathbb{R}} h_0 * \check{h}_0\left(x - \frac{a}{t}\right) P_{2m+2}(x) dx.\end{aligned} \tag{14}$$

For all $a \neq 0$ and $x \in \mathbb{R}$ the term $h_0 * \check{h}_0\left(x - \frac{a}{t}\right) P_{2m+2}(x)$ converges to zero as $t \to 0$. As $h_0 * \check{h}_0$ is compactly supported and bounded, Lebesgue's dominated convergence theorem shows that $\int_{\mathbb{R}} h_0 * \check{h}_0\left(x - \frac{a}{t}\right) P_{2m+2}(x) dx$ converges to zero. Therefore the second term of the right hand side of (14) converges to zero of order $o(t^{2m+2})$.

The asymptotic behavior of the first term on the right hand side of (13) can be determined by the following reasoning. If $g^{(2m+2)}$ is $(0,1)$-piecewise smooth we can again apply (Kiêu, 1997, Proposition 5.6) to obtain $(g^{(2m+2)} * h_t * \check{h}_t)^{(1)} = g^{(2m+3)} * h_t * \check{h}_t + s_{g^{(2m+2)}} * h_t * \check{h}_t$. This derivative is again continuous, so partial integration of the integral in the first term on the right hand side of (13) yields

$$\begin{aligned}&\int_{\mathbb{R}} g^{(2m+2)} * h_t * \check{h}_t(x) P_{2m+2}\left(\frac{x}{t}\right) dx \\ &\quad = -t \int_{\mathbb{R}} g^{(2m+3)} * h_t * \check{h}_t(x) P_{2m+3}\left(\frac{x}{t}\right) dx \\ &\qquad - t \sum_{a \in D_{g^{(2m+2)}}} s_{g^{(2m+2)}}(a) \int_{\mathbb{R}} h_t * \check{h}_t(x-a) P_{2m+3}\left(\frac{x}{t}\right) dx.\end{aligned}$$

We here use that $\int P_{2m+3}\left(\frac{x}{t}\right) dx = \frac{1}{t} P_{2m+2}\left(\frac{x}{t}\right)$.

It is not difficult to see that both terms on the right hand side of the above equation converge of order at least $O(t)$, hence the first term on the right hand side

of (13) converges to zero of order $o(t^{2m+2})$. If instead of assuming that $g^{(2m+2)}$ is $(0,1)$-piecewise smooth, we only require that $g^{(2m+2)} \in \mathcal{C}_K$, then $g^{(2m+2)}$ is Riemann integrable, so for each $\varepsilon > 0$ there exists a step function $\tilde{g} \in \mathcal{C}_K$, such that $\tilde{g} \leq g^{(2m+2)}$ and

$$0 \leq \int_{\mathbb{R}} g^{(2m+2)}(x)dx - \int_{\mathbb{R}} \tilde{g}(x)dx \leq \frac{\varepsilon}{2\|P_{2m+2}\|_\infty}.$$

It is not difficult to check that this implies

$$0 \leq \int_{\mathbb{R}} g^{(2m+2)} * h_t * \check{h}_t(x)dx - \int_{\mathbb{R}} \tilde{g} * h_t * \check{h}_t(x)dx \leq \frac{\varepsilon}{2\|P_{2m+2}\|_\infty}$$

and

$$\left| \int_{\mathbb{R}} g^{(2m+2)} * h_t * \check{h}_t(x) P_{2m+2}\left(\frac{x}{t}\right) dx - \int_{\mathbb{R}} \tilde{g} * h_t * \check{h}_t(x) P_{2m+2}\left(\frac{x}{t}\right) dx \right| \leq \frac{\varepsilon}{2}.$$

As $\tilde{g}$ is $(0,1)$-piecewise smooth, we can apply the same argument as above in order to show that $\tilde{I} := \int_{\mathbb{R}} \tilde{g} * h_t * \check{h}_t(x) P_{2m+2}\left(\frac{x}{t}\right) dx = O(t)$. In particular $|\tilde{I}| \leq \frac{\varepsilon}{2}$, for t small enough. This implies

$$\left| \int_{\mathbb{R}} g^{(2m+2)} * h_t * \check{h}_t(x) P_{2m+2}\left(\frac{x}{t}\right) dx \right| \leq \varepsilon$$

for t small enough, hence this integral tends to zero as $t \to 0$. Therefore the first term on the right hand side of (13) converges to zero of order $o(t^{2m+2})$.

It only remains to show the representation of the first term in (8) as given in (9) and (11). It is easy to see, that

$$g(0) - g * \check{h}_t * h_t(0) = \int_{\mathbb{R}} (g(0) - g(tx)) \check{h}_0 * h_0(x)\, dx.$$

Fix $x \in \mathbb{R}$. For $t > 0$ small enough we can use Taylor expansion to obtain

$$g(xt) - g(0) = \sum_{k=1}^{m} \frac{1}{(2k)!} g^{(2k)}(0) x^{2k} t^{2k} + \frac{1}{(2m+1)!} g^{(2m+1)}(\xi) x^{2m+1} t^{2m+1}$$

as all uneven continuous derivatives of g are odd functions so they are zero at zero; ξ is between 0 and xt. If $m = 0$ and $x > 0$, then

$$\frac{1}{t^2} t(g(xt) - g(0)) = g'(\xi)x \to g'(0^+)x \quad \text{as } t \to 0^+.$$

Using Lebesgue's dominated convergence theorem and $s_{g'}(0) = 2g'(0^+)$, see (Kiêu, 1997, p. 56), one can deduce (9). Equation (11) also follows by dominated convergence, using the boundedness of $g^{(2m+1)}$. Finally, one obtains equation (10) combining (8) and (9) using the definition of the second Bernoulli polynomial. □

5.2 Systematic sampling with cumulative error

In this section we study the asymptotic behavior of $\operatorname{var}\left(\widehat{\Theta}\right)$ under the model of systematic sampling with cumulative error. We assume that the error density for a certain spacing $t > 0$ is given by $h_t(x) = \frac{1}{t} h_0\left(\frac{x}{t}\right)$, where h_0 is a probability density on the positive halfline with expected value 1. Define $u_t^+ := \sum_{k=1}^{\infty} h_t^{k*}$, $u_t^- := \sum_{k=1}^{\infty} \check{h}_t^{k*}$. Note that u_t^+ is supported on the positive halfline, while u_t^- is supported on the negative halfline. Furthermore $u_t^{\pm}(x) = \frac{1}{t} u_0^{\pm}\left(\frac{x}{t}\right)$. Rewriting equation (5) with this notation yields

$$\operatorname{var}\left(\widehat{\Theta}\right) = tg(0) + \int_0^{\infty} g(z) u_0^+\left(\frac{z}{t}\right) dz + \int_{-\infty}^{0} g(z) u_0^-\left(\frac{z}{t}\right) dz - \int_{\mathbb{R}} g(z) dz. \qquad (15)$$

The function u_0^+ is the renewal density of a renewal process with holding times that are independent identically distributed with density h_0. The following theorem, see (Alsmeyer, 1991, Satz 3.3.1, Satz 13.2.2), reveals the asymptotic behavior of $u_0^{\pm}$.

Theorem 5.2. *Let $h : \mathbb{R}_+ \to \mathbb{R}_+$ be a probability density with expectation $\mu > 0$ and let $u := \sum_{k=1}^{\infty} h^{k*}$. Suppose that $h \in L^{\infty}$ and $\lim_{s\to\infty} h(s) = 0$. Then*

(a) $u \in L^{\infty}$ and $u - h$ is continuous and bounded.

(b) $\lim_{s\to\infty} u(s) = \mu^{-1}$, where $\infty^{-1} := 0$.

(c) If h is absolutely continuous and there is an integer $m \geq 2$, such that

$$\int_{\mathbb{R}} |x^{m-1} h'(x)| dx < \infty$$

and the m-th moment of h exists, then

$$u(s) - \mu^{-1} = o(s^{1-m}), \quad \text{as } s \to \infty.$$

Remark. Absolutely continuous functions are almost everywhere differentiable. Every Lipschitz-function is absolutely continuous. For further reference on absolute continuity, see Rudin (1986).

We assume now that $h_0 \in L^{\infty}$ and that $\lim_{s\to\infty} h_0(s) = 0$. By part (b) of the above theorem we obtain for each $z \in [0, \infty)$

$$\lim_{t\to 0} g(z) u_0^+\left(\frac{z}{t}\right) = g(z)$$

and analogously for $z \in (-\infty, 0]$ and u_0^-. Furthermore $|g(z) u_0^+\left(\frac{z}{t}\right)| \leq \|u_0^+\|_{\infty} |g(z)| \in L^1$, using part (a) of the above theorem, and again analogously for u_0^-. Lebesgue's dominated convergence theorem now implies

$$\lim_{t\to 0} \left(\int_0^{\infty} g(z) u_0^+\left(\frac{z}{t}\right) dz + \int_{-\infty}^{0} g(z) u_0^-\left(\frac{z}{t}\right) dz \right) = \int_{\mathbb{R}} g(z) dz$$

and hence

$$\lim_{t\to 0} \operatorname{var}\left(\widehat{\Theta}\right) = 0.$$

The actual order of convergence is determined in the proposition below. Note that if h_0 does not have expected value 1, then $\operatorname{var}\left(\widehat{\Theta}\right)$ does not converge to zero for $t \to 0$.

Proposition 5.3. *Assume that h_0 satisfies the conditions in part (c) of Theorem 5.2 for some $m \geq 3$ and that the covariogram g is continuous at 0 and bounded. Then the variance of the generalized Cavalieri estimator under the model of systematic sampling with cumulative error has the following limiting behavior*

$$\operatorname{var}\left(\widehat{\Theta}\right) = tg(0)\nu^2 + o(t)$$

as $t \to 0$, where $\nu^2 < \infty$ is the variance of a random variable with probability density h_0.

Proof. The assumptions on h_0 yield that $(u_0^+ - 1)$ is integrable. Using substitution we obtain

$$\begin{aligned}\int_0^\infty g(z)\left(u_0^+\left(\frac{z}{t}\right) - 1\right)dz =& t\int_0^\infty (g(tz) - g(0))(u_0^+(z) - 1)dz \\ &+ tg(0)\int_0^\infty (u_0^+(z) - 1)dz.\end{aligned}$$

The first term on the right hand side of the above equation converges of order $o(t)$ as $t \to 0$. This can be seen by using dominated convergence and the continuity of g at 0. As g is symmetric and $u_0^-(z) = u_0^+(-z)$ we obtain

$$\operatorname{var}\left(\widehat{\Theta}\right) = tg(0)\left(2\int_0^\infty (u_0^+(z) - 1)dz + 1\right) + o(t), \tag{16}$$

using equation (15). Let U be the renewal measure of the renewal process with holding times that are independent identically distributed with density h_0. Then u_0^+ is a density for $U - \delta_0$. The function $(u_0^+(z) - 1)\mathbf{1}_{[0,K]}(z)$ converges in L^1 to $u_0^+(z) - 1$ as $K \to \infty$, therefore

$$\begin{aligned}\int_0^\infty (u_0^+(z) - 1)dz &= \lim_{K\to\infty}\int (u_0^+(z) - 1)\mathbf{1}_{[0,K]}(z)dz \\ &= \lim_{K\to\infty}((U - \delta_0)([0, K]) - K) \\ &= \frac{\nu^2 - 1}{2},\end{aligned}$$

by (Alsmeyer, 1991, Theorem 3.4.1). Combining this with equation (16) yields the claim. □

5.3 Systematic sampling with independent p-thinning

The model of systematic sampling with independent p-thinning is a two-stage model. We have to specify the underlying center process Ψ and the thinning probability $p > 0$. We consider the two main cases of Ψ.

Perturbed systematic sampling with independent p-thinning: Suppose the center process Ψ follows the model of perturbed systematic sampling with error density h_t as given in Section 5.1.

Proposition 5.4. *Let f be an $(m,1)$-piecewise smooth measurement function. Then its covariogram g is $(2m+1,1)$-piecewise smooth and the variance of the generalized Cavalieri estimator under perturbed systematic sampling combined with independent p-thinning with thinning probability $p>0$ has the following asymptotic behavior as $t \to 0$*

$$\operatorname{var}\bigl(\widehat{\Theta}\bigr) = t\frac{p}{1-p}g(0) + o(t).$$

Proof. This follows by combining equation (6) in Section 4.3 with Proposition 5.1. □

Systematic sampling with cumulative error and independent p-thinning: Let the center process Ψ follow the model of systematic sampling with cumulative error with increment density h_t as defined in Section 5.2.

Proposition 5.5. *Assume that the conditions on h_0 of part (c) of Theorem 5.2 are fulfilled for some $m \geq 3$ and that the covariogram g is continuous at 0 and bounded. Then the variance of the generalized Cavalieri estimator under systematic sampling with cumulative error combined with independent p-thinning with thinning probability $p>0$ has the expansion*

$$\operatorname{var}\bigl(\widehat{\Theta}\bigr) = tg(0)\left(\nu^2 + \frac{p}{1-p}\right) + o(t)$$

as $t \to 0$, where $\nu^2 < \infty$ is the variance of a random variable with probability density h_0.

Proof. This follows by combining equation (6) in Section 4.3 with Proposition 5.3. □

6 An example

As an example, we have investigated the effect of errors in sample locations of section planes on the precision of the estimator of the volume of the unit ball. In this case, the measurement function f and the geometric covariogram g can be calculated as follows

$$\begin{aligned} f(x) &= \pi(1-x^2)\mathbf{1}_{[-1,1]}(x) \\ g(x) &= \pi^2\left(\frac{16}{15} - \frac{4}{3}x^2 + \frac{2}{3}|x|^3 - \frac{1}{30}|x|^5\right)\mathbf{1}_{[-2,2]}(x). \end{aligned}$$

In Figure 1, the variance of the generalized Cavalieri estimator under the model of perturbed systematic sampling is displayed and compared to the variance of the estimator under 'exact' systematic sampling. The density h_0 is a truncated normal density with mean zero and truncation points $\pm 1/2$. In Figure 1, the variance of the resulting estimators are plotted against the expected number of sections n. Note that $t = 2/n$ as we are cutting a unit ball. The variances used in Figure 1 for the truncated normal density are $\sigma^2 = 0,\ 0.05^2,\ 0.10^2$ for the lower, middle and upper plots, respectively. Here, $\sigma^2 = 0$ corresponds to exact systematic sampling.

The order of magnitude of the positive σ^2s has been chosen in accordance with what has been found in recent morphological studies where the model of perturbed systematic sampling is appropriate, see Dorph-Petersen et al. (2005, 2007). Methods of statistical analysis of this type of data will be provided in a forthcoming paper written for users by Dorph-Petersen, Baddeley, Ziegel and Jensen.

The measurement function f of the unit ball is $(1, \infty)$-piecewise smooth. Applying Proposition 5.1 we obtain that

$$\operatorname{var}\big(\widehat{\Theta}\big) = -\frac{c_2}{2} g^{(2)}(0) t^3 + o(t^3)$$

under the model of perturbed systematic sampling. This asymptote can also be seen as a line in Figure 1 for each of the two cases of positive σ^2.

We also computed the variance of the generalized Cavalieri estimator under the model of systematic sampling with cumulative error and compared it to the variance of the estimator under 'exact' systematic sampling. The increment density h_0 is a truncated normal density with mean 1, truncation points 0 and 2 and variance σ^2. Again, $\sigma^2 = 0$ corresponds to exact systematic sampling. For the calculation we approximated the k-th fold convolution of the truncated normal density h_0 by a truncated normal density with mean k, truncation points 0 and $2k$ and variance $\sqrt{k}\sigma^2$. The variances used in Figure 2 are $\sigma^2 = 0$, 0.05^2, 0.10^2 for the lower, middle and upper curve, respectively.

It is of note that, as shown in Figure 2, cumulative error may have a substantial effect on variance. For example, if 100 sections are used, exact sampling gives a very small coefficient of variation ($CV = \sqrt{\text{Variance}}/(4\pi/3)$) which is about 0.002%. But for systematic sampling with cumulative error even with the smaller standard deviation of $\sigma = 0.05$, the CV is about 0.55%. The effect for perturbed systematic sampling, on the other hand, appears to be less significant, cf. Figure 1.

7 Discussion

The reason why random sampling experiments have become so important in biological applications of stereological methods is that most biological structures are highly organized and spatially inhomogeneous so that sampling inference cannot be drawn from a single arbitrarily positioned sample Weibel (1978). Randomization of sampling points is needed if the material cannot be regarded as homogeneous. A first mention of a design based approach in stereology can be found in the farsighted paper Thompson (1932), see also the accompanying paper Thompson et al. (1932) and Royall (1970). An alternative to randomization of sampling points would be to develop a stochastic model for the biological structure under study. This is, however, not needed for estimating parameters Θ expressible as integrals. There is a strong analogy between randomizing the position of the grid for estimating Θ and designing a sample survey for estimating the population total of a finite population.

In the present paper, we have proposed two quite different models to deal with errors in systematic sampling. The choice of the model will be specific to the application. Furthermore, the suggested models may be modified to take special features of the sampling procedure into account, such as loss of observations. Our example

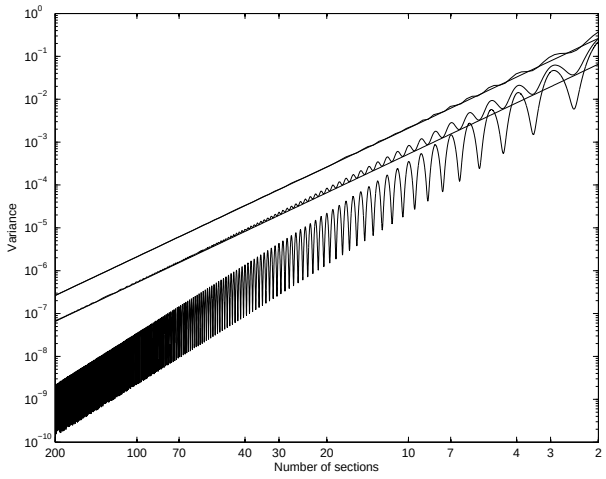

Figure 1: Variance of the Cavalieri estimator of volume of a unit ball as a function of the expected number of sections is shown on a log-log scale. The lower curve is based on exact systematic sampling. The upper and middle curves were calculated using the model of perturbed systematic sampling with a truncated normal error distribution h_0 with mean zero, truncation points $\pm 1/2$ and standard deviation $\sigma = 0.05$ (middle curve) and $\sigma = 0.10$ (upper curve), respectively. The straight lines represent the main terms of the asymptotic expansion of the variances.

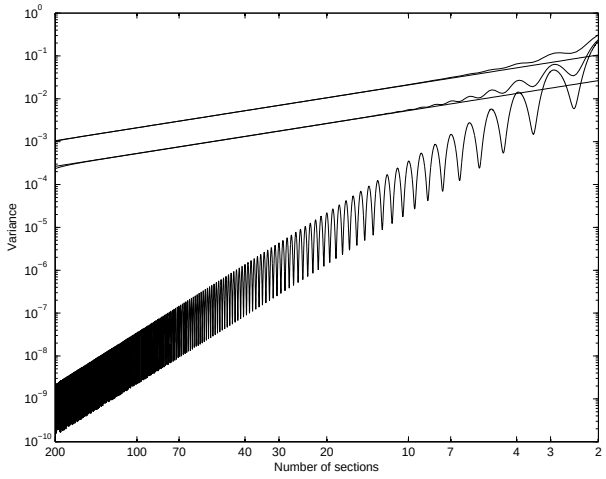

Figure 2: Variance of the Cavalieri estimators of volume of a unit ball as a function of the expected number of sections is shown on a log-log scale. The lower curve is based on exact systematic sampling. The upper and middle curves were calculated using the model of systematic sampling with cumulative error with a truncated normal increment distribution h_0 with mean 1, truncation points 0 and 2 and standard deviation $\sigma = 0.05$ (middle) and $\sigma = 0.10$ (upper), respectively. The straight lines represent the main terms of the asymptotic expansion of the variances.

in Section 6 shows that errors in the placement of sampling points may lead to a substantial inflation of the estimator variance.

There are a number of ways in which the methods presented here may be extended. Measurement functions f with first order derivative being non-continuous with infinite jumps are not covered by the asymptotic theory developed in the present paper. In the case of the classical Cavalieri estimator, the asymptotic variance has been derived for such measurement functions in García-Fiñana & Cruz-Orive (2000, 2004); García-Fiñana (2006). The variance exhibits a fractional trend. The trend is often of order T^{2p+2}, typically with $0 < p < 1$. For the perturbed systematic sampling model a next step will be to use this theory to extend Proposition 5.1 to very general measurement functions. Another obvious extension concerns the effect on the variance of errors in placement of sampling points in the case where sampling in two or three dimensions is performed. Appendix A of this paper represents a first step in this direction.

For applications in microscopy and spatial surveys, it is of great importance to be able to estimate the variance of the generalized Cavalieri estimator from data. One obvious possibility is to try to estimate the leading terms of the asymptotic expansions in Section 5 from data. This task is part of our future research plans.

Acknowledgements

Johanna Ziegel would like to thank Prof. Paul Embrechts for his mathematical, financial and personal support and Dr. Johanna Neslehova for fruitful discussions. This work was supported by a grant from the Danish Natural Science Research Council.

Appendix

Systematic sampling in higher dimensions

Assume we want to estimate the volume of a bounded Borel set $B \subseteq \mathbb{R}^d$ with the unbiased estimator $\hat{V}$, defined by

$$\hat{V} = t^d \sum_{z \in \mathbb{Z}^3} \mathbf{1}_B(t(U+z)),$$

where $t > 0$ and U is a uniform random variable in $[0,1]^3$. We can generalize this estimator in the following way. Let Φ be a first order stationary point process in $\mathbb{R}^d$ with intensity measure $M_1 = m\mathcal{L}$, $m > 0$ and $f : \mathbb{R}^d \to \mathbb{R}$ a measurement function with compact support. In the case of volume estimation we simply have $f = \mathbf{1}_B$.

Proposition 7.1. *The estimator*

$$\tilde{V} := \frac{1}{m} \sum_{x \in \Phi} f(x)$$

is an unbiased estimator of the integral $W := \int f(x)dx$.

Proof.

$$\mathrm{E}\big(\tilde{V}\big) = \frac{1}{m}\,\mathrm{E}\Big(\sum_{x\in\Phi} f(x)\Big) = \frac{1}{m}\int f(x)M_1(dx) = \int f(x)dx.$$

□

Proposition 7.2. *Let C_2 denote the second cumulant measure (covariance measure) of Φ. Suppose it exists. Then*

$$\mathrm{var}\big(\tilde{V}\big) = \frac{1}{m^2}\int f(x)f(y)C_2(dx\times dy).$$

Proof. Let M_2 denote the second moment measure of Φ, then C_2 is defined as $C_2(A\times B) = M_2(B\times A) - M_1(A)M_2(B)$ for Borel sets A, B.

$$\begin{aligned}
\mathrm{var}\big(\tilde{V}\big) &= \mathrm{E}\big(\tilde{V}^2\big) - \mathrm{E}\big(\tilde{V}\big)^2 \\
&= \frac{1}{m^2}\,\mathrm{E}\left(\Big(\sum_{x\in\Phi} f(x)\Big)^2\right) - W^2 \\
&= \frac{1}{m^2}\,\mathrm{E}\Big(\sum_{x,y\in\Phi} f(x)f(y)\Big) - W^2 \\
&= \frac{1}{m^2}\left(\int f(x)f(y)M_2(dx\times dy) - \left(\int f(x)M_1(dx)\right)^2\right) \\
&= \frac{1}{m^2}\int f(x)f(y)C_2(dx\times dy).
\end{aligned}$$

□

References

Alsmeyer, G. (1991). *Erneuerungstheorie.* B. G. Teubner, Stuttgart.

Arnau, X. G. & Cruz-Orive, L. M. (1998). Variance prediction under systematic sampling with geometric probes. *Adv. Appl. Prob. (SGSA)* 30 889–903.

Baddeley, A., Dorph-Petersen, K. A. & Jensen, E. B. V. (2006). A note on the stereological implications of irregular spacing of sections. *J. Microsc.* 222 177–181.

Baddeley, A. & Jensen, E. B. V. (2005). *Stereology for Statisticians.* Chapman & Hall/CRC, Boca Raton.

Crofton, M. W. (1885). Probability. In *Encyclopaedia Britannica.* 9th ed.

Cruz-Orive, L. M. (1989). On the precision of systematic sampling: a review of Matheron's transitive methods. *J. Microsc.* 153 315–333.

Daley, D. J. & Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes.* Springer, New York.

Delesse, A. (1847). Procédé mécanique pour déterminer la composition des roches. *Comptes Rendues de l'Académie des Sciences (Paris)* 25 544–545.

Delesse, A. (1848). Procédé mécanique pour déterminer la composition des roches. *Annales des Mines* 13 379–388.

Dorph-Petersen, K.-A. (1999). Stereological estimation using vertical sections in a complex tissue. *J. Microsc.* 195 79–86.

Dorph-Petersen, K.-A., Pierri, J. N., Perel, J. M., Sun, Z., Sampson, A. R. & Lewis, D. A. (2005). The influence of chronic exposure to antipsychotic medications on brain size before and after tissue fixation: A comparison of haloperidol and olanzapine in macaque monkeys. *Neuropsychopharmacology* 30 1649–1661.

Dorph-Petersen, K.-A., Pierri, J. N., Wu, Q., Sampson, A. R. & Lewis, D. A. (2007). Primary visual cortex volume and total neuron number are reduced in schizophrenia. *J. Comp. Neurol.* 501 290–301.

García-Fiñana, M. (2006). Confidence intervals in Cavalieri sampling. *J. Microsc.* 222 146–157.

García-Fiñana, M. & Cruz-Orive, L. M. (2000). Fractional trend of the variance in Cavalieri sampling. *Image Anal. Stereol.* 19 71–79.

García-Fiñana, M. & Cruz-Orive, L. M. (2004). Improved variance prediction for systematic sampling on $\mathbb{R}$. *Statistics* 38(3) 243–272.

Glagolev, A. A. (1933). On geometrical methods of quantitative mineralogic analysis of rocks. *Trans. Inst. Econ. Min.* 59 1–47.

Gundersen, H. J. G., Bagger, P., Bendtsen, T. F., Evans, S., Korbo, L., Marcussen, N., Møller, A., Nielsen, K., Nyengaard, J. R., Pakkenberg, B., Sørensen, F. B., Vesterby, A. & West, M. J. (1988). The new stereological tools: Disector, fractionator, nucleator and point sampled intercepts and their use in pathological research and diagnosis. *APMIS* 96 857–881.

Jones, A. E. (1948). Systematic sampling of continuous parameter populations. *Biometrika* 35 283–296.

Kiêu, K. (1997). Three lectures on systematic geometric sampling. *Memoirs 13, Department of Theoretical Statistics, University of Aarhus* .

Kiêu, K., Souchet, S. & Istas, J. (1999). Precision of systematic sampling and transitive methods. *J. Stat. Plan. Infer.* 77 263–279.

Knopp, K. (1996). *Theorie und Anwendung der unendlichen Reihen*. Springer, Berlin.

Lund, J. & Rudemo, M. (2000). Models for point processes observed with noise. *Biometrika* 87 235–249.

MATHERON, G. (1965). *Les Variables Régionalisées et Leur Estimation.* Masson et Cie, Paris.

MATHERON, G. (1970). The theory of regionalized variables and its applications. *Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau* 5.

MATTFELDT (ED.), T. (2006). Special volume on variance estimation in stereology. *J. Microsc.* 222 143–255.

MORAN, P. A. P. (1966). Measuring the length of a curve. *Biometrika* 53 359–364.

MORAN, P. A. P. (1968). Statistical theory of a high-speed photoelectric planimeter. *Biometrika* 55 419–422.

PACHE, J.-C., ROBERTS, N., VOCK, P., ZIMMERMANN, A. & CRUZ-ORIVE, L. M. (1993). Vertical LM sectioning and CT scanning designs for stereology: application to human lung. *J. Microsc.* 170 9–24.

ROSIWAL, A. (1898). Über geometrische Gesteinsanalysen. Ein einfacher Weg zur ziffermässigen Feststellung des Quantitätsverhältnisses der Mineralbestandteile gemengter Steine. *Verhandlungen der Kaiserlich-Königlichen Geologischen Reichsanstalt Wien* 143–175.

ROYALL, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* 57 377–387.

RUDIN, W. (1986). *Real and Complex Analysis.* McGraw-Hill, Singapore.

STEINHAUS, H. (1929). Sur la portée pratique et théorique de quelques théorèmes sur la mesure des ensembles de droites. In *Comptes Rendues 1er Congr. Mathématiciens des Pays Slaves.* Warszawa, 348–354.

STEINHAUS, H. (1954). Length, shape and area. *Colloq. Math.* 3 1–13.

STOYAN, D., KENDALL, W. S. & MECKE, J. (1995). *Stochastic Geometry and its Applications.* John Wiley and Sons, Chichester.

SWEET, R. A., DORPH-PETERSEN, K.-A. & LEWIS, D. A. (2005). Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. *J. Comp. Neurol.* 491 270–289.

THOMPSON, W. R. (1932). The geometric properties of microscopic configurations. I. General aspects of projectometry. *Biometrika* 24 21–26.

THOMPSON, W. R., HUSSEY, R., MATTEIS, J. T., MEREDITH, W. C., WILSON, G. C. & TRACY, F. E. (1932). The geometric properties of microscopic configurations. II. Incidence and volume of islands of Langerhans in the pancreas of a monkey. *Biometrika* 24 27–38.

THOMSON, E. (1930). Quantitative microscopic analysis. *J. Geol.* 38 193–222.

Weibel, E. R. (1978). The non-statistical nature of biological structure and its implications on sampling for stereology. In R. E. Miles & J. Serra, eds., *Geometrical Probability and Biological Structures: Buffon's 200th Anniversary*, Lecture Notes in Biomathematics, No 23. Springer Verlag, Berlin-Heidelberg-New York.

Weil, W. & Schneider, R. (2008). *Stochastic and Integral Geometry*. Springer, Heidelberg.