CENTRE FOR **STOCHASTIC GEOMETRY**
AND ADVANCED **BIOIMAGING**

2D non-uniform systematic sampling

Ina Trolle Andersen and Johanna F. Ziegel

No. 14, December 2014

2D non-uniform systematic sampling

Ina Trolle Andersen[1,2] and Johanna F. Ziegel[3]

[1]Department of Mathematics, Aarhus University, ita@math.au.dk
[2]Stereological Research Laboratory, Department of Clinical Medicine, Aarhus University
[3]Institute of Mathematical Statistics and Actuarial Science, University of Bern, johanna.ziegel@stat.unibe.ch

Abstract

In this paper, we introduce a natural generalization of non-uniform systematic sampling to two dimensions. Under optimal auxiliary information about the function of interest, this design yields an estimator with zero variance. In a simulation study imitating sampling situations in microscopy, the 2D non-uniform systematic designs show similar efficiency as proportional-to-size sampling with replacement. An exception is area estimation where the 2D non-uniform systematic designs are superiour in a number of cases considered.

Keywords: Cavalieri estimator, efficiency, Horvitz-Thompson estimator, microscopy, systematic sampling, 2D sampling

1 Introduction

In this paper, we introduce a new 2D non-uniform systematic sampling design that respects the spatial information available and study its efficiency. In designs like the proportionator (Gardi et al., 2008a,b), which is used in microscopy, all spatial information is lost, prior to sampling. Therefore, a natural idea to improve upon such designs, is to include spatial information in the sampling procedure. We propose to use transformations of 2D uniform systematic sampling into non-uniform sampling, while still maintaining some of the spatially balanced arrangement. Unlike Grafström and Tillé (2012) and Stevens Jr and Olsen (2004), which also try to balance non-uniform samples spatially, our proposal is a genuine 2D sampling procedure.

The suggested 2D non-uniform systematic sampling design is a generalization of Dorph-Petersen et al. (2000), where non-uniform systematic sampling was introduced in 1D. In that paper it was concluded from simulations, that non-uniform sampling was more efficient than traditional uniform sampling, known as the classical 2D Cavalieri estimator, in an example of area estimation from lengths of linear intercepts. In the present paper we propose a 2D non-uniform systematic sampling design, which under optimal auxiliary information about the function of interest, yields zero variance of the estimator. Furthermore, a simulation study resembling

sampling in microscopy, has been performed to investigate the applicability of the method.

The paper is organised as follows. First we recall results from Dorph-Petersen et al. (2000) for the 1D sampling procedure. Then, we derive the generalization to higher dimensions and propose various transformations of 2D uniform systematic sampling. Subsequently, methods and results of the simulation study are presented, followed by a discussion. Technical proofs are deferred to two appendices.

2 Generalized systematic sampling in 1D

2.1 Theoretical considerations

Let f be a bounded non-negative function with bounded support, assumed to be the interval $[0, 1]$ without loss of generality. The objective is to estimate the integral

$$Q = \int_0^1 f(x)\,\mathrm{d}x, \tag{2.1}$$

using values of f at a set of random sampling points. In uniform systematic sampling (Cruz-Orive, 1993; Gundersen et al., 1999), a random systematic set of n points is selected, $Y_i = (U + i)/n$, where $U \sim \text{unif}[0, 1)$ and $i = 0, 1 \ldots, n - 1$, from which f is estimated by a simple step function. Thus, Q can be estimated unbiasedly by

$$\hat{Q}_n = \frac{1}{n}\sum_{i=0}^{n-1} f(Y_i). \tag{2.2}$$

Instead of equidistant sampling points, it may be more efficient to choose points closer to each other in some areas, while keeping a spatially spread sample. This can be obtained by considering an increasing bijective function $G : [0, 1] \to [0, 1]$, which transforms the sampling points into new points $X_i = G^{-1}(Y_i), i = 0, \ldots, n-1$. Using that

$$Q = \int_0^1 f(x)\,\mathrm{d}x = \int_0^1 f(G^{-1}(x))\frac{1}{G'(G^{-1}(x))}\,\mathrm{d}x, \tag{2.3}$$

we obtain a new unbiased estimator

$$\hat{Q}_n = \frac{1}{n}\sum_{i=0}^{n-1} \frac{f(X_i)}{G'(X_i)} = \frac{1}{n}\sum_{i=0}^{n-1} \frac{f(G^{-1}(Y_i))}{G'(G^{-1}(Y_i))},$$

which corresponds to the uniform systematic sampling estimator of the function

$$\tilde{f}(y) = f(G^{-1}(y))/G'(G^{-1}(y)).$$

2.2 Choice of sampling points

Finding efficient sampling points, hence G and thereby $\tilde{f}$, is considered in Dorph-Petersen et al. (2000), where three choices are investigated. An obvious possibility

is to choose G such that $\tilde{f}$ is constant. Then $\hat{Q}_n$ is constant and always yields the true value Q. Under the assumption that $f > 0$, the function G can be defined by

$$G(x) = \frac{\int_0^x f(t)\,\mathrm{d}t}{\int_0^1 f(t)\,\mathrm{d}t}. \tag{2.4}$$

In Dorph-Petersen et al. (2000), the relation between this choice and sampling proportional-to-size in the discrete setting, is established. Another suggestion uses that the transformation corresponds to uniform systematic sampling for $\tilde{f}$, hence asymptotic results in the uniform case can be used. Transitive methods yield that the asymptotic variance depends on the smoothness of $\tilde{f}$, hence it is preferable to have $\tilde{f}$ smoother than f. The third suggestion is based on the idea of sampling the most, where f varies the most.

The three possibilities considered in the paper each depend on the function f, thus prior knowledge of f or properties of f are needed. In practice f is unknown, therefore a function f_0 similar to f is used instead. For instance using this in (2.4), the estimator $\hat{Q}_n$ becomes

$$\hat{Q}_n = \int_0^1 f_0(t)\,\mathrm{d}t \cdot \frac{1}{n}\sum_{i=0}^{n-1} \frac{f(X_i)}{f_0(X_i)}, \quad X_i = G^{-1}((U+i)\tfrac{1}{n}). \tag{2.5}$$

3 Generalized systematic sampling in 2D

3.1 Theoretical considerations

The idea is now to generalize this to higher dimensions. For simplicity, we restrict our considerations to two dimensions as the theoretical considerations are directly transferable to any dimension. Let f be a bounded non-negative function with bounded support $[0,1]^2$. The objective is to estimate the integral

$$Q = \int_{[0,1]^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y, \tag{3.1}$$

using values of f at a set of random sampling points in $[0,1]^2$. In uniform systematic sampling, a random systematic set of $n \times m$ points is selected,

$$(Y_{1i}, Y_{2j}) = ((U_1+i)\tfrac{1}{n}, (U_2+j)\tfrac{1}{m}),$$

where U_1 and U_2 are independent and $U_1, U_2 \sim \text{unif}[0,1)$, $i = 0, 1 \ldots, n-1$ and $j = 0, 1 \ldots, m-1$. From these points f is estimated by a simple step function, hence Q can be estimated unbiasedly by

$$\hat{Q}_{nm} = \frac{1}{nm}\sum_{i=0}^{n-1}\sum_{j=0}^{m-1} f(Y_{1i}, Y_{2j}).$$

Similar to the 1D case, it might be better to choose points closer to each other in some areas. This can be obtained by considering a diffeomorphism $G : [0,1]^2 \to [0,1]^2$,

which transforms the sampling points into new points $(X_{1i}, X_{2j}) = G^{-1}(Y_{1i}, Y_{2j})$, $i = 0, \dots, n-1$, $j = 0, \dots, m-1$. Using that

$$\begin{aligned} Q &= \int_{[0,1]^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{[0,1]^2} f(G^{-1}(x,y)) \frac{1}{|\det(G'(G^{-1}(x,y)))|}\,\mathrm{d}x\,\mathrm{d}y, \end{aligned}$$

where G' denotes the Jacobi matrix of G, we obtain a new unbiased estimator

$$\begin{aligned} \hat{Q}_{nm} &= \frac{1}{nm} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \frac{f(X_{1i}, X_{2j})}{|\det(G'(X_{1i}, X_{2j}))|} \\ &= \frac{1}{nm} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \frac{f(G^{-1}(Y_{1i}, Y_{2j}))}{|\det(G'(G^{-1}(Y_{1i}, Y_{2j})))|} \end{aligned}$$

which, again, corresponds to the uniform systematic sampling estimator of the function

$$\tilde{f}(x,y) = f(G^{-1}(x,y))/|\det(G'(G^{-1}(x,y)))|.$$

In the Appendix, we proof that $\hat{Q}_{nm}$ is in fact unbiased.

3.2 Choice of sampling points

Analogously to 1D we would like to choose the sampling points in an efficient manner. In particular, we can determine G such that $\tilde{f}$ is constant, which implies that the variance of $\hat{Q}_{nm}$ becomes zero. Assuming $f(x,y) > 0$, $\tilde{f}$ is constant if we choose any G with

$$|\det(G'(u,v))| = cf(u,v), \tag{3.2}$$

for some constant $c > 0$. There are many possible choices of diffeomorphisms with property (3.2), one example is $G = (G_1, G_2)$, where

$$G_1(x,y) = \int_0^x g(u)\,\mathrm{d}u,\; G_2(x,y) = \frac{1}{g(x)\Delta} \int_0^y f(x,v)\,\mathrm{d}v, \tag{3.3}$$

and

$$g(x) = \frac{1}{\Delta} \int_0^1 f(x,v)\,\mathrm{d}v,\; \Delta = \int_{[0,1]^2} f(u,v)\,\mathrm{d}u\,\mathrm{d}v. \tag{3.4}$$

The Jacobi matrix is given by

$$\begin{pmatrix} \frac{\partial G_1}{\partial x} & \frac{\partial G_1}{\partial y} \\ \frac{\partial G_2}{\partial x} & \frac{\partial G_2}{\partial y} \end{pmatrix} = \begin{pmatrix} g(x) & 0 \\ \cdots & \frac{f(x,y)}{g(x)\Delta} \end{pmatrix},$$

hence $|\det(G'(u,v))| = f(u,v)/\Delta$. Replacing f with a known function f_0, which is similar to f, will be used in practice. This could be colour values obtained by image

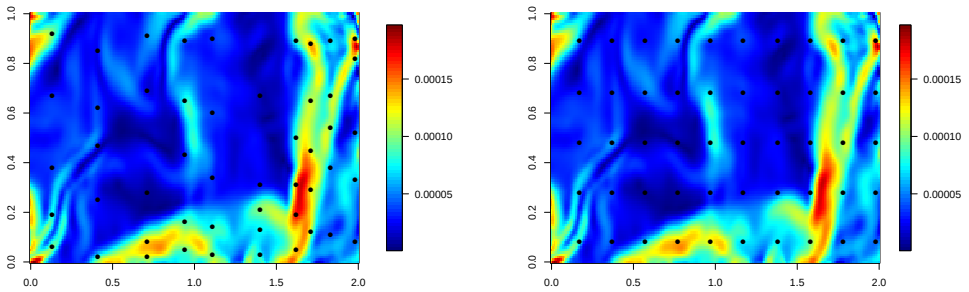

Figure 1: Left: 2D uniform systematic sampling. Right: 2D non-uniform systematic sampling, where the points have been transformed proportional to the colour values using (3.3) and (3.4). The colours are obtained from the *bei*-data in the R-package Spatstat, see Baddeley and Turner (2005).

analysis. Like in 1D this choice of G is related to sampling proportional-to-size in the discrete set-up. In the Appendix, this relation is established. Replacing f with f_0 in (3.3) and (3.4), the estimator becomes

$$\hat{Q}_{nm} = \int_{[0,1]^2} f_0(u,v)\,\mathrm{d}u\,\mathrm{d}v \frac{1}{nm}\sum_{i=0}^{n-1}\sum_{j=0}^{m-1}\frac{f(X_{1i},X_{2j})}{f_0(X_{1i},X_{2j})}, \tag{3.5}$$
$$(X_{1i},X_{2j}) = G^{-1}((U_1+i)\tfrac{1}{n},(U_2+j)\tfrac{1}{m}).$$

Figure 1 illustrates a 2D uniform systematic sample, which is transformed into a non-uniform systematic sample, proportional to the colour values. Note that the non-uniform sample is still spatially spread. However, the statistical behaviour of the estimator $\hat{Q}_{nm}$ depends on which coordinate in (3.3) and (3.4) is chosen first. If we interpret Φ as the result of a gravitational force acting on the systematic grid of points, G has to be given as $G(x,y) = \nabla\Phi(x,y)$ for some scalar potential Φ. Hence to fulfil property (3.2), we seek a solution of the Monge-Amper equation

$$\frac{\partial^2\Phi}{\partial x^2}\frac{\partial^2\Phi}{\partial y^2} - \frac{\partial^2\Phi}{\partial x\partial y} = cf(x,y)$$

see e.g. Kołodziej (1998), subject to some suitable boundary and convexity conditions. Unfortunately this equation is known to be very difficult to solve even numerically.

Instead, keeping the ideas behind G in (3.3) and (3.4), it is possible to decrease the strong dependence on the order of the coordinates by working with the composition $G = \Psi \circ \Phi$ of two transformations, where $\Phi : [0,1]^2 \to [0,1]^2$ is given by

$$\Phi_1(x,y) = \int_0^x g(u)\,\mathrm{d}u,$$
$$\Phi_2(x,y) = \frac{1}{g(x)\Delta_g}\int_0^y f(x,v)^\alpha\,\mathrm{d}v,$$

where $\alpha \in (0,1)$,

$$g(x) = \frac{1}{\Delta_g} \int_0^1 f(x,v)^\alpha \, \mathrm{d}v, \quad \Delta_g = \int_{[0,1]^2} f(u,v)^\alpha \, \mathrm{d}u \, \mathrm{d}v.$$

and $\Psi : [0,1]^2 \to [0,1]^2$ is given by

$$\Psi_1(x,y) = \frac{1}{h(y)\Delta_h} \int_0^x (f \circ \Phi^{-1})(u,y)^{1-\alpha} \, \mathrm{d}u,$$
$$\Psi_2(x,y) = \int_0^y h(v) \, \mathrm{d}v,$$

where

$$h(y) = \frac{1}{\Delta_h} \int_0^1 (f \circ \Phi^{-1})(u,y)^{1-\alpha} \, \mathrm{d}u,$$
$$\Delta_h = \int_{[0,1]^2} (f \circ \Phi^{-1})(u,v)^{1-\alpha} \, \mathrm{d}u \, \mathrm{d}v.$$

It is straight forward to show, that this choice of G fulfils (3.2), and we expect that the resulting estimator is more efficient. Unlike G in (3.3) and (3.4), where the transformed points are aligned in one direction, this composition of G results in a completely deformed grid of points, which is not effected much by the order of coordinates. In practice, it may be difficult to work with a transformation of the form $G = \Psi \circ \Phi$ as inverting G can be very hard even numerically.

4 Simulation study

4.1 Motivation

To support our theoretical findings and to investigate the efficiency of the new design compared to existing designs, a simulation study was carried out. As the project was motivated by sampling in microscopy, it is natural to construct set-ups, which resemble data obtained from this field of research. Here, non-uniform sampling is mainly used for counting purposes, for instance for determining the total number of cells in a cell population, but also area estimation is considered, see e.g. Gardi et al. (2008a,b). Both number estimation and area estimation will be studied in the present paper, as well as estimation of the integral of a function. The latter case will be called integral estimation and involves investigation of more smooth functions than the measurement functions in the first two cases, which are based on indicator functions. In microscopy, the value of the measurement function in a point corresponds to complete information from a small observation window, located at this point.

We use a selection of stochastic point processes to generate the centres of the imaginary cells for number estimation, and thereby a selection of measurement functions f. Various choices of the function f_0, which controls the transformation G of the sampling points, is investigated in all three cases of estimation. The function f_0 is constructed from f with different levels of noise including spatial errors. In this

set-up, f_0 corresponds to known auxiliary information, obtainable from automatic image analysis of a tissue section, for instance the amount of a predetermined color identifying a staining of the cells.

4.2 Sampling approximations

The inverse of some non-standard integrals is needed to obtain the samples, but as this is quite time consuming, the integrals are approximated by sums, and only the non-decomposed choice of G given in (3.3) and (3.4) is considered. Furthermore, f_0 is only determined in a finite number of points, corresponding to a pixelation of the image. Let f_0 be determined in a finite number of points in $[0,1]^2$, say in $N \times M$ equidistant points, (x_k, y_l), $k = 1, \ldots, N$, $l = 1, \cdots, M$, where $x_k = (k-1)/N$ and $y_l = (l-1)/M$. These values correspond to an approximation of G using the lower left corners of the 'pixels'. This choice allows for a particularly simple approximation of G^{-1}. We assume that N and M are large compared to the size of the observation window, thus we have a set-up, which resembles a continuous set-up. Replacing f with f_0 in (3.3) and (3.4), we obtain for $x \in [x_k, x_{k+1}) = [\frac{k-1}{N}, \frac{k}{N})$, $y \in [y_l, y_{l+1}) = [\frac{l-1}{M}, \frac{l}{M})$, $k = 1, \ldots, N$, $l = 1, \ldots, M$, discrete approximations given by

$$
\begin{aligned}
\Delta &= \int_{[0,1]^2} f_0(u,v)\,\mathrm{d}u\,\mathrm{d}v \approx \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} f_0(x_i, y_j),\\
g(x) &= \frac{1}{\Delta}\int_0^1 f_0(x,v)\,\mathrm{d}v \approx \frac{1}{\Delta M}\sum_{j=1}^{M} f_0(x_k, y_j),\\
G_1(x) &= \int_0^x g(u)\,\mathrm{d}u = \sum_{i=1}^{k-1}\int_{x_i}^{x_{i+1}} g(u)\,\mathrm{d}u + \int_{x_k}^{x} g(u)\,\mathrm{d}u\\
&\approx \frac{1}{N}\sum_{i=1}^{k-1} g(x_i) + (x - \tfrac{k-1}{N})g(x_k)\\
&\approx \frac{1}{\Delta NM}\sum_{j=1}^{M}\Big(\sum_{i=1}^{k-1} f_0(x_i, y_j) + (Nx - k + 1)f_0(x_k, y_j)\Big),\\
G_2(x,y) &= \frac{1}{\Delta g(x)}\int_0^y f_0(x,v)\,\mathrm{d}v\\
&\approx \frac{1}{\Delta g(x_k)}\Big(\sum_{j=1}^{l-1}\int_{y_j}^{y_{j+1}} f_0(x_k,v)\,\mathrm{d}v + \int_{y_l}^{y} f_0(x_k,v)\,\mathrm{d}v\Big)\\
&\approx \frac{1}{\Delta g(x_k)M}\Big(\sum_{j=1}^{l-1} f_0(x_k, y_j) + (My - l + 1)f_0(x_k, y_l)\Big).
\end{aligned}
$$

In particular,

$$G_1(x_k) \approx \frac{1}{\Delta NM} \sum_{j=1}^{M} \sum_{i=1}^{k-1} f_0(x_i, y_j),$$

$$G_2(x_k, y_l) \approx \frac{1}{\Delta g(x_k) M} \sum_{j=1}^{l-1} f_0(x_k, y_j),$$

for $k = 1, \ldots, N$, $l = 1, \ldots, M$. To approximate the inverse we take for $u, v \in [0, 1)$, $G^{-1}(u, v) \approx (x_k, y_l)$, where $k = \max \{i \mid G_1(x_i) \leq u\}$, $l = \max \{j \mid G_2(x_k, y_j) \leq v\}$. For a given choice of f_0, using the approximations above, it is straight forward to simulate non-uniform samples from uniformly generated random variables, and calculate the estimate from (3.5).

4.3 Sampling designs

In order to investigate the efficiency of the sampling design, the variance or CE^2 (squared coefficient of error) of the design is compared to the ones for standard sampling designs. The designs which are considered are

- 2D continuous: 2D non-uniform systematic sampling, described above and in Section 3.2 (continuous sampling).
- 2D discrete: 2D non-uniform systematic sampling, described in Appendix B, corresponding to proportional-to-size sampling (PPS sampling) with spatial information. This design is a new design proposed in the present paper. It is thought as a compromise between the continuous design above and standard PPS sampling with no spatial information (discrete sampling).
- PPS WR: Proportional-to-size sampling with replacement (discrete sampling).
- SRS WOR: Simple random sampling without replacement (discrete sampling).

When discrete sampling is considered, the section is divided by an equally spaced grid into fields of size $w \times w$ (determined in the section below), thus sampling is performed on a finite number of elements.

4.4 The measurement function

Without loss of generality we assume, that the measurement function is defined on the unit square $[0, 1]^2$. We imitate observations in microscopy by letting the measurement function $f(x, y)$ be proportional to the measurement of interest on a whole observation window $(x, y) + [0, w]^2$, with side-length $0 \leq w \leq 1$. As we shall see, the parameters of interest are expressible as an integral (3.1) of f over $[0, 1]^2$ if the spatial structure is contained in $[w, 1]^2$.

4.4.1 The measurement function for number estimation

Let $Z = \{z_1, z_2, \ldots\}$ be points in $[w, 1]^2$, generated by a pre-chosen point process, indicating centres of a cell population, and assume we want to estimate the number

of points in Z, N_Z. Then, $W_i = z_i - [0,w]^2$ is the set of (x,y) in $[0,1]^2$ for which z_i is counted in the observation window $(x,y)+[0,w]^2$. If we let $f(x,y)$ be the total number of points, counted in $(x,y)+[0,w]^2$, normalized with $1/w^2$,

$$f(x,y) = \sum_{i=1}^{N_Z} \frac{1_{W_i}(x,y)}{w^2}, \tag{4.1}$$

we obtain, as desired, the total number of points N_Z by

$$\begin{aligned} Q &= \int_{[0,1]^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{[0,1]^2} \sum_{i=1}^{N_Z} \frac{1_{W_i}(x,y)}{w^2}\,\mathrm{d}x\,\mathrm{d}y \\ &= \sum_{i=1}^{N_Z} \frac{|W_i|}{w^2} = N_Z. \end{aligned}$$

4.4.2 The measurement function for area or integral estimation

Let λ be a non-negative function on $\mathbb{R}^2$ that is identically 0 outside $[w,1]^2$, and assume we want to estimate

$$\int_{[w,1]^2} \lambda(u,v)\,\mathrm{d}u\,\mathrm{d}v,$$

which for $\lambda(x,y) = 1((x,y) \in A)$ corresponds to the area of the set $A \subseteq [w,1]^2$. If we let $f(x,y)$ be the integral of λ over $(x,y)+[0,w]^2$, normalized with $1/w^2$,

$$f(x,y) = \frac{1}{w^2} \int_{(x,y)+[0,w]^2} \lambda(u,v)\,\mathrm{d}u\,\mathrm{d}v, \tag{4.2}$$

we obtain the complete integral by

$$\begin{aligned} Q &= \int_{[0,1]^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{w^2} \int_{[0,1]^2} \Big(\int_{(x,y)+[0,w]^2} \lambda(u,v)\,\mathrm{d}u\,\mathrm{d}v \Big)\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{w^2} \int_{[0,1]^2} \Big(\int_{[0,w]^2} \lambda(u+x,v+y)\,\mathrm{d}u\,\mathrm{d}v \Big)\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{w^2} \int_{[0,w]^2} \Big(\int_{[0,1]^2} \lambda(u+x,v+y)\,\mathrm{d}x\,\mathrm{d}y \Big)\,\mathrm{d}u\,\mathrm{d}v \\ &= \frac{1}{w^2} \int_{[0,w]^2} \Big(\int_{(u,v)+[0,1]^2} \lambda(x,y)\,\mathrm{d}x\,\mathrm{d}y \Big)\,\mathrm{d}u\,\mathrm{d}v \\ &= \frac{1}{w^2} \int_{[0,w]^2} \Big(\int_{[w,1]^2} \lambda(x,y)\,\mathrm{d}x\,\mathrm{d}y \Big)\,\mathrm{d}u\,\mathrm{d}v \\ &= \int_{[w,1]^2} \lambda(u,v)\,\mathrm{d}u\,\mathrm{d}v. \end{aligned}$$

4.5 Computational details

The simulations have been performed under the following specifications:

- The constructed images consist of 128×128 pixels covering $[w, 1]^2$.
- We let the observation windows consist of 8×8 pixels, thus $|W_i| = w^2$, $w = 8/(128 + 8 - 1)$.
- We have $N = M = 128 + 8 - 1 = 135$.
- The set $[w, 1]^2$ can be covered by 16×16 observation windows, which is considered in the cases of discrete sampling.
- For each set-up we used sample sizes $n \times m$, where $n = m = 3$.
- For each set-up 10 000 samples were simulated to approximate the theoretical variance of the estimator.

4.6 Number estimation

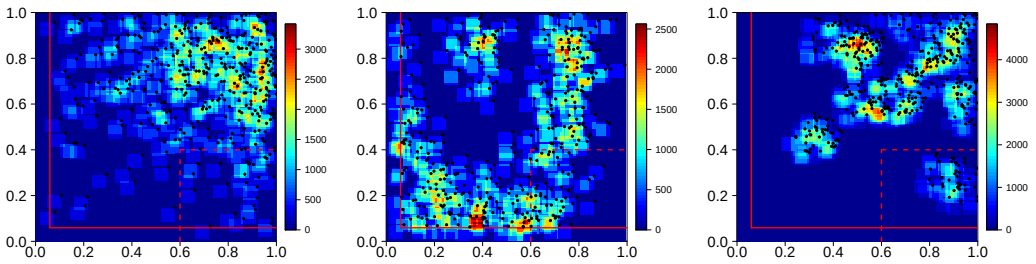

Figure 2: The three cases of the measurement function f, cf. (4.1), together with the points from realizations of the point process Z. The full-drawn red lines indicates the lower and left boundary of $[w, 1]^2$, and the red dashed lines indicates the 'invisible' part of Z, for some choices of f_0, see (4.3) and Table 1.

4.6.1 Choices of f and f_0

Several set-ups have been tested, including different sample sizes (n and m) and population sizes (N_Z). Only some of these will be discussed in detail here. We consider three cases of the measurement function f, which are shown in Figure 2. In all cases, f is given by (4.1), generated by point patterns obtained as realizations of different inhomogeneous point processes, resulting in approximately 400 points. We define the auxiliary information f_0 by the same sum of indicator functions as the measurement function f in (4.1), but with additional spatial error and a constant added, to avoid division by zero. More precisely,

$$f_0(x, y) = \sum_{z_i \in [0,1]^2 \setminus B} \alpha_i \frac{1_{W_i}(x, y)}{w^2} + c, \tag{4.3}$$

with different choices of α_i, B and c. The 9 choices are shown in Table 1 and illustrated in Figure 8 in Appendix A.3. Both α_i and B are used to introduce a spatial error, in particular B is used to make some of the points 'invisible', that is, the points are not expressed in f_0.

Table 1: Parameter choices for f_0 in formula (4.3).

f_0	α_i	B	c
1	$x_i + y_i$	$\emptyset$	1000
2	1	$\emptyset$	1000
3	$x_i + y_i$	$\emptyset$	250
4	1	$\emptyset$	250
5	$x_i + y_i$	$[0.6; 1] \times [0; 0.4]$	1000
6	1	$[0.6; 1] \times [0; 0.4]$	1000
7	$x_i + y_i$	$[0.6; 1] \times [0; 0.4]$	250
8	1	$[0.6; 1] \times [0; 0.4]$	250
9	0	0	1

4.6.2 Results

Figure 3 shows the estimated CE^2 (Row 1) and the estimated relative CE^2 (Row 2) obtained by 10 000 sample simulations, for different choices of the auxiliary information f_0 (x-axis), as detailed in Table 1, with $n = m = 3$.

The values of the three measurements functions f lies in the intervals $[0; 3417]$, $[0; 2563]$ and $[0; 4841]$, thus when constructing f_0 from f with error, the non-uniformity is considerable and large 'smoothing' parameters are needed, e.g. large values of the constant c in Table 1, to prevent extreme estimates. Moreover, extreme non-uniformity results in overlaps between the sample windows in the 2D continuous systematic sampling design and non-fixed effective sample-sizes in the 2D discrete systematic sampling design, even for relative small sample sizes. Due to the extreme non-uniformity, $n = m = 3$ is the upper bound on the sample size in the three cases considered here, and this sample size was therefore chosen.

Figure 3 shows that the main effect on efficiency is from the choice of f_0, and there is almost no difference in efficiency between the three designs, 2D continuous, 2D discrete and PPS WR, which uses non-uniform sampling. The effect from the two more complicated designs, which use spatial information, compared to the much more simple PPS WR design, is therefore negligible. The variances naturally become larger in the non-uniform cases, when many of the cells are placed in the invisible part of the section, but due to the large value of c, this effect is not clearly expressed in the results. In the uniform case, choice 9 of f_0, the effect solely from systematic sampling can be seen. Systematic uniform sampling surprisingly yields in its continuous implementation a higher CE^2 than SRS WOR for the second choice of f, and only little effect is seen for the remaining two choices of f, which may explain why the non-uniform systematic sampling designs do not have the expected effect. In general, it seems that when spatial error is introduced (choices 1,3,5 and 7 of f_0), 2D continuous systematic sampling is either almost as efficient or slightly more efficient than PPS WR, and overall the 2D discrete systematic sampling design performs better than the continuous version and often even slightly better than PPS WR.

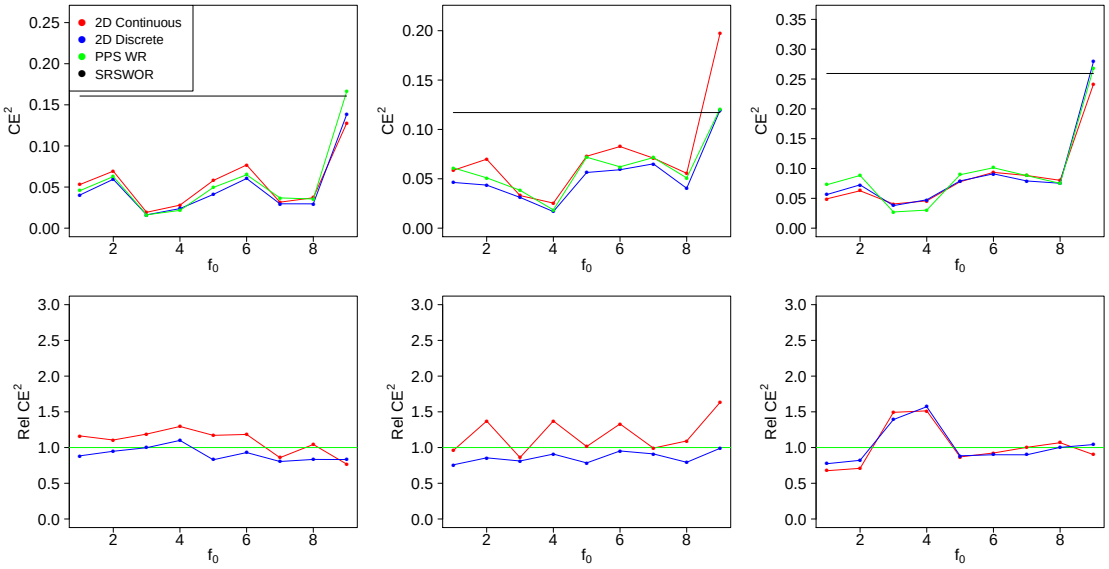

Figure 3: Results for number estimation with $n = m = 3$. Row 1: The estimated CE^2 for the four sampling designs, in the three cases of f. For each choice of f, the nine choices of f_0 are displayed. Row 2: The estimated relative CE^2 for the two 2D non-uniform systematic sampling designs, relative to PPS WR, in the three cases of f. For each choice of f, the nine choices of f_0 are displayed.

4.7 Area estimation

4.7.1 Choices of f and f_0

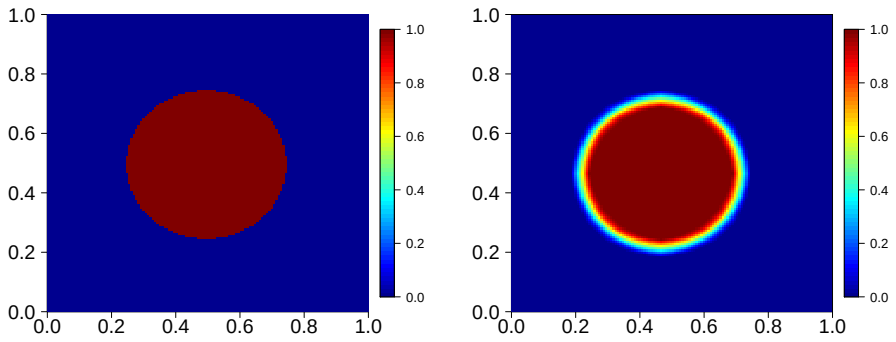

Figure 4: Left: The function λ equal to the indicator function of the set of interest. Right: The corresponding measurement function f, see (4.2).

The integration function λ is in this case an indicator function, indicating where the tissue of interest is located. A simple example is investigated here, see Figure 4, where we consider the area of a circular disc with centre $(0.5, 0.5)$ and radius 0.25, denoted $B((0.5, 0.5), 0.25)$. More precisely, using (4.2), the measurement function f can be written as

$$f(x, y) = |B((0.5, 0.5), 0.25) \cap ((x, y) + [0, w]^2)|/w^2. \tag{4.4}$$

We define the auxiliary information f_0 from the measurement function f by adding spatial error in the following way

$$f_0(x,y) = c_1 f(x,y) + c_2 f(x,y) \\ \times \Gamma(x,y)(\Phi_R(x,y) - \Phi_r(x,y)) + c_3, \tag{4.5}$$

where $\Phi_r(x,y) = 1((x-0.5)^2 + (y-0.5)^2 > r^2)$ and $\Phi_R(x,y) = 1((x-0.5)^2 + (y-0.5)^2 \leq R^2)$, creating a smoother band around the boundary of the circular disk, and with different choices of Γ, c_1, c_2 and c_3. The choices are shown in Table 2 and illustrated in Figure 9 in Appendix A.3. In all cases $r = 0.15$ and $R = 0.35$.

Table 2: Parameter choices for f_0 in formula (4.5).

f_0	Γ	c_1	c_2	c_3
1	$x+y$	1	0.4	0.2
2	1	1	0.4	0.2
3	$x+y$	0	0.4	0.2
4	1	0	0.4	0.2
5	$x+y$	1	1	0.2
6	1	1	1	0.2
7	$x+y$	0	1	0.2
8	1	0	1	0.2
9	$x+y$	1	0.4	0.4
10	1	1	0.4	0.4
11	$x+y$	0	0.4	0.4
12	1	0	0.4	0.4
13	$x+y$	1	1	0.4
14	1	1	1	0.4
15	$x+y$	0	1	0.4
16	1	0	1	0.4
17	0	0	0	1

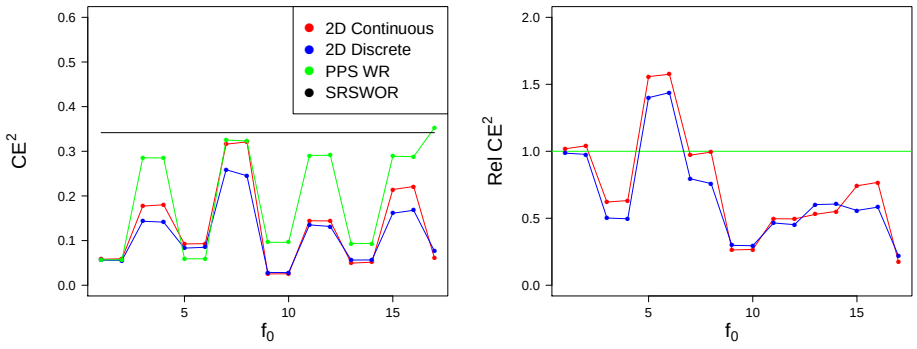

Figure 5: Results for area estimation with $n = m = 3$. Left: The estimated CE^2 for the four sampling designs for the nine choices of f_0 are displayed. Right: The estimated relative CE^2 for the two 2D non-uniform systematic sampling designs, relative to PPS WR.

4.7.2 Results

Figure 5 shows the estimated CE^2 and the estimated relative CE^2 obtained by 10 000 sample simulations, for different choices of the auxiliary information f_0, as detailed in Table 2, with $n = m = 3$.

For most choices of f_0, the non-uniformity is only moderate, hence sample sizes up to $n \times m$, with $n = m = 5$, without non-fixed effective sample sizes are possible in 2D discrete sampling. Nevertheless, in order to minimize the probability of repeated sampling of the same window in PPS WR, we consider only results for $n = m = 3$. Difference between PPS WR and the 2D systematic sampling designs will therefore not solely be due to the re-sampling probability.

The effect from just introducing systematic sampling in the uniform case, choice 17 of f_0, is clearly substantial, and might explain why, in contrast to the case of number estimation, 2D non-uniform systematic sampling (2D continuous and 2D discrete in Figure 5) overall performs better than PPS WR. Although uniform systematic sampling is an efficient design here, the efficiency may still be increased by combining the systematic sampling with non-uniform sampling. The efficiency is most pronounced for moderate non-uniformity, as large values of c_3 reduces the variance of the 2D sampling designs relative to both SRS WOR and PPS WR. There seems to be almost no difference in the results from the discrete and the continuous 2D systematic sample designs.

Clearly the parameter c_1 affects the efficiency relative to SRS WOR, as $c_1 = 1$ results in high agreement between f and f_0, thus non-uniform sampling is close to optimal. The influence of c_1 for the non-uniform designs is less clear. The parameter c_2 controls the magnitude of spatial error, and clearly larger values of c_2 results in a higher variance, but in contrast to what is expected, it also results in less efficiency compared to PPS WR. There seems to be no or negligible effect from the spatial error introduced by Γ.

4.8 Integral estimation

4.8.1 Choices of f and f_0

Here we consider three cases of functions λ, which together with the corresponding measurement functions f, see (4.2), are shown in Figure 6. The integrals have been scaled, such that they integrate to one, which unifies the set-up and choices of parameters. We define the auxiliary information f_0 from the measurement function f by adding spatial error in the following way

$$f_0(x, y) = c_1 f(x, y) + c_2 f(x, y)\Gamma(x, y) + c_3, \tag{4.6}$$

with different choices of Γ, c_1, c_2 and c_3. The choices are shown in Table 3 and illustrated in Figure 10 in Appendix A.3.

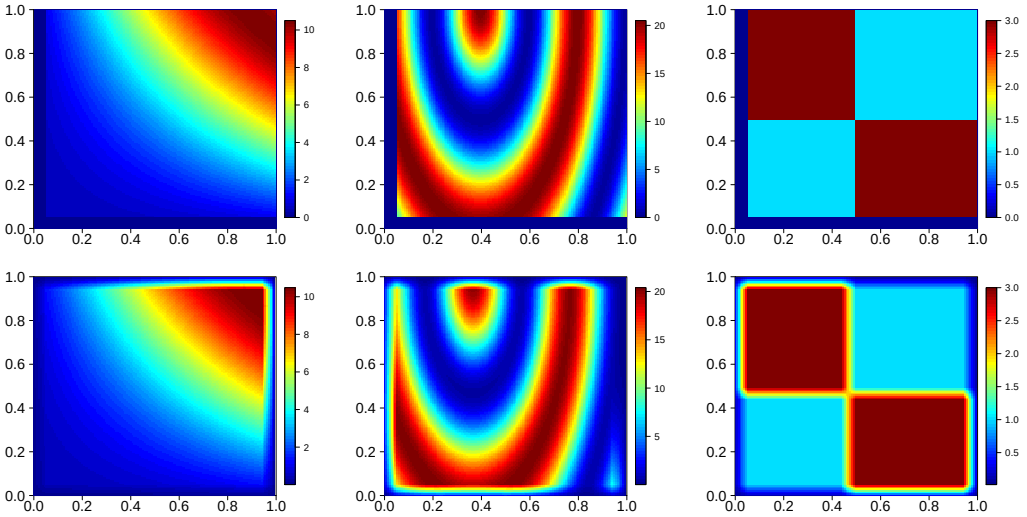

Figure 6: Row 1: The three functions λ. Row 2: The three corresponding measurement functions f.

Table 3: Parameter choices for f_0 in formula (4.6), with $A = [0,1]^2 \backslash ([0.6,1] \times [0,0.4])$.

Sim	Γ	c_1	c_2	c_3
1	$x+y$	1	0.4	0.4
2	1	1	0.4	0.4
3	$x+y$	0	0.4	0.4
4	1	0	0.4	0.4
5	$x+y$	1	1	0.4
6	1	1	1	0.4
7	$x+y$	0	1	0.4
8	1	0	1	0.4
9	$(x+y+0.1)1((x,y) \in A)$	1	0.4	0.4
10	$1((x,y) \in A)$	1	0.4	0.4
11	$(x+y+0.1)1((x,y) \in A)$	0	0.4	0.4
12	$1((x,y) \in A)$	0	0.4	0.4
13	$(x+y+0.1)1((x,y) \in A)$	1	1	0.4
14	$1((x,y) \in A)$	1	1	0.4
15	$(x+y+0.1)1((x,y) \in A)$	0	1	0.4
16	$1((x,y) \in A)$	0	1	0.4
17	0	0	0	1

4.8.2 Results

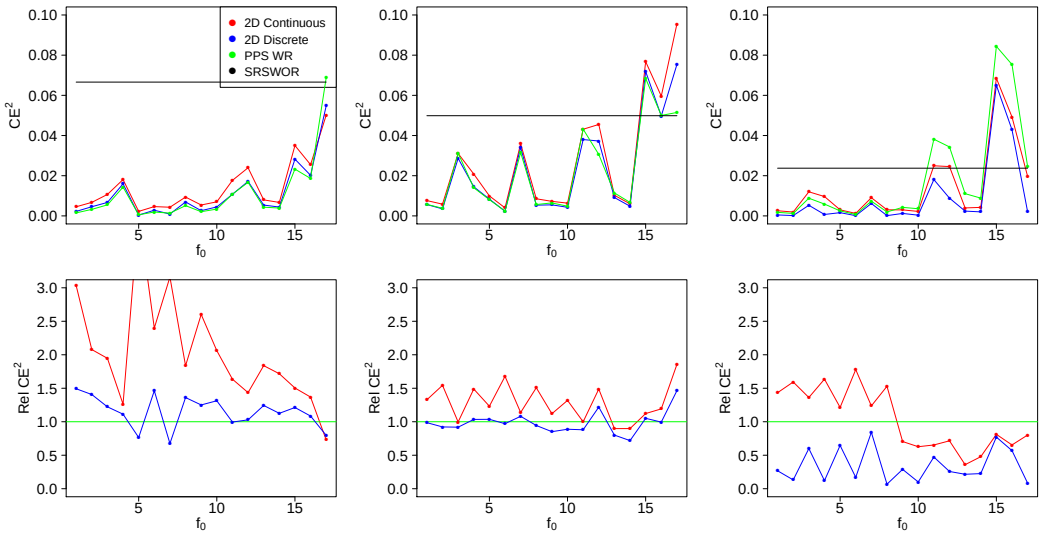

Figure 7: Results for area estimation with $n = m = 3$. Row 1: The estimated CE^2 for the four sampling designs, in the three cases of f. Row 2: The estimated relative CE^2 for the two 2D non-uniform systematic sampling designs, relative to PPS WR, in the three cases of f. For each choice of f, the seventeen choices of f_0 are displayed.

Figure 7 shows the estimated CE^2 and the estimated relative CE^2, obtained by 10 000 sample simulations, for different choices of the auxiliary information f_0, detailed in Table 3, with $n = m = 3$.

In all cases of f_0, the non-uniformity is only moderate, hence allow sample sizes up to $n \times m$, with $n = m = 5$, without non-fixed effective sample sizes in 2D discrete sampling. Nevertheless, with arguments stated in the previous section, we consider only results for $n = m = 3$.

The main effect is again from the choice of f_0, and there is almost no difference between the three designs, which use non-uniform sampling. The relative differences seem on the other hand more significant. In particular the 2D continuous design does not perform well, but the 2D discrete design perform similar or better than PPS WR.

Introducing systematic sampling in the uniform case, choice 17 of f_0, increases the efficiency in the first and third case, where the non-uniformity is most systematic, whereas in the second case the efficiency decreases. Interchanging the sampling order (first on the y-axis, then on the x-axis instead of the opposite order) did not change the results much.

5 Discussion and other perspectives

We have shown the existence of an optimal choice of 2D non-uniform systematic sampling. The design reduces the variance substantially compared to uniform sampling, when the auxiliary information used to construct the sampling inclusion probabilities have a close connection to the measurement function under consideration. More

precisely, when a function proportional to the measurement function is known, the design yields zero variance of the estimator.

To support our theoretical findings and investigate the efficiency of this new design compared to other designs, a simulation study was carried out. As the project was motivated by sampling in microscopy, it was natural to construct a set-up, which resembles data obtained from this field of research. Number estimation, area estimation and general integral estimation was simulated for several choices of the measurement function and the auxiliary information, with different levels of spatial error added.

In most cases considered, the 2D non-uniform systematic designs had similar efficiency as PPS WR. An exception was area estimation where the non-uniform systematic designs were superior in a number of cases considered. Within the non-uniform systematic designs, the new discrete design is more efficient than the continuous design in a number of cases. The discrete design is easy to simulate and with a moderate constant added a robust and efficient choice, if one suspects that the auxiliary information f_0 is not optimal.

In the simulation study, other sample sizes (n and m) were considered such that the range 3 % to 25 % of the total number of observation windows was covered. The conclusion concerning the relative efficiency of the designs was the same as for the sample sizes considered in the present paper.

In the case of number estimation in a Poisson point process, it is expected that 2D non-uniform systematic sampling does not have higher efficiency than the proportionator, simply because the numbers counted in disjoint observation windows are independent. The simulation study shows that in a wider range of sampling situations in microscopy the gain in efficiency, if any, is modest.

One might wonder whether it is possible to construct a theoretical class of sampling situations for which the parameter of interest is more efficiently estimated, using 2D non-uniform systematic sampling compared to independent 2D non-uniform sampling. For this purpose, let f_{hom} be a bounded non-negative function with bounded support $[0,1]^2$. Suppose that

$$\operatorname{Var}\Big(\frac{1}{nm}\sum_{i,j} f_{\text{hom}}(Y_{1i}, Y_{2j})\Big) < \operatorname{Var}\Big(\frac{1}{nm}\sum_{i,j} f_{\text{hom}}(U_{1i}, U_{2j})\Big),$$

where $(Y_{1i}, Y_{2j}) = ((U_1 + i)\frac{1}{n}, (U_2 + j)\frac{1}{m})$ and $U_1, U_2 \sim \text{unif}[0,1)$ independent, while the U_{1i}s and U_{2j}s are all independent and unif[0,1). Furthermore, let $G : [0,1]^2 \to [0,1]^2$ be any diffeomorphism. Then,

$$f_{\text{inhom}}(x,y) = f_{\text{hom}}(G(x,y)) \cdot |\det G'(x,y)|$$

is more efficiently estimated using 2D non-uniform systematic sampling than independent 2D non-uniform sampling, induced by G. More precisely, the variance of

$$\frac{1}{nm}\sum_{i,j} \frac{f_{\text{inhom}}(G^{-1}(Y_{1i}, Y_{2j}))}{|\det G'(G^{-1}(Y_{1i}, Y_{2j}))|}$$

is smaller than the variance of

$$\frac{1}{nm}\sum_{i,j} \frac{f_{\text{inhom}}(V_{ij})}{|\det G'(V_{ij})|},$$

where $V_{ij} = (V_{1ij}, V_{2ij})$ are independent and with common density $|\det G'(v)|$. The practical consequences of this finding are subject of future research.

6 Acknowledgement

The authors would like to thank Eva B. Vedel Jensen for helpful suggestions and constructive comments in the preparation of the manuscript. This research was supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by the Villum Foundation.

References

Baddeley, A. and R. Turner (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software 12*(6), 1–42.

Cruz-Orive, L. (1993). Systematic sampling in stereology. *Proceedings 49th Session Int. Statist. Inst. 2*, 451–468.

Dorph-Petersen, K.-A., H. J. G. Gundersen, and E. B. V. Jensen (2000). Non-uniform systematic sampling in stereology. *Journal of Microscopy 200*(2), 148–157.

Gardi, J. E., J. R. Nyengaard, and H. J. G. Gundersen (2008a). Automatic sampling for unbiased and efficient stereological estimation using the proportionator in biological studies. *Journal of Microscopy 230*(1), 108–120.

Gardi, J. E., J. R. Nyengaard, and H. J. G. Gundersen (2008b). The proportionator: unbiased stereological estimation using biased automatic image analysis and non-uniform probability proportional to size sampling. *Computers in Biology and Medicine 38*(3), 313–328.

Grafström, A. and Y. Tillé (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics 24*(2), 120–131.

Gundersen, H. J. G., E. B. V. Jensen, K. Kiêu, and J. Nielsen (1999). The efficiency of systematic sampling in stereology – reconsidered. *Journal of Microscopy 193*(3), 199–211.

Kołodziej, S. (1998). The complex Monge-Ampere equation. *Acta Mathematica 180*(1), 69–117.

Stevens Jr, D. L. and A. R. Olsen (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association 99*(465), 262–278.

Appendix

A.1 Unbiasedness of $\hat{Q}_{nm}$

The estimator $\hat{Q}_{nm}$ is unbiased, which follows directly by using the distribution of U_1 and U_2. We let $A_n = [0, \frac{1}{n}] \times [0, \frac{1}{m}]$ and obtain

$$\begin{aligned}
\mathbb{E}(\hat{Q}_{nm}) &= \mathbb{E}\Big(\frac{1}{nm}\sum_{i=0}^{n-1}\sum_{j=0}^{m-1}\frac{f(G^{-1}(Y_{1i}, Y_{2j}))}{|\det(G'(G^{-1}(Y_{1i}, Y_{2j})))|}\Big) \\
&= \frac{1}{nm}\sum_{i=0}^{n-1}\sum_{j=0}^{m-1}\mathbb{E}\Big(\frac{f(G^{-1}(\frac{U_1+i}{n}, \frac{U_2+j}{m}))}{|\det(G'(G^{-1}(\frac{U_1+i}{n}, \frac{U_2+j}{m})))|}\Big) \\
&= \frac{1}{nm}\sum_{i=0}^{n-1}\sum_{j=0}^{m-1}\int_{A_n}\frac{nmf(G^{-1}(u+\frac{i}{n}, v+\frac{j}{m}))}{|\det(G'(G^{-1}(u+\frac{i}{n}, v+\frac{j}{m})))|}\,\mathrm{d}u\,\mathrm{d}v \\
&= \sum_{i=0}^{n-1}\sum_{j=0}^{m-1}\int_{[\frac{i}{n}, \frac{i+1}{n}]\times[\frac{j}{m}, \frac{j+1}{m}]}\frac{f(G^{-1}(u,v))}{|\det(G'(G^{-1}(u,v)))|}\,\mathrm{d}u\,\mathrm{d}v \\
&= \int_{[0,1]^2} f(G^{-1}(u,v))\frac{1}{|\det(G'(G^{-1}(u,v)))|}\,\mathrm{d}u\,\mathrm{d}v \\
&= \int_{[0,1]^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= Q.
\end{aligned}$$

A.2 Relation to proportional-to-size sampling

2D systematic sampling proportional-to-size can be described as follows. Let us consider a population of $N \times M$ units $\mathscr{P} = \{(i,j), i = 1, \ldots, N, j = 1, \ldots M\}$, where the numbers refer to the spatial arrangement. For each unit we have an unknown variable of interest x_{ij}, and a known auxiliary variable z_{ij}, which is positively correlated with x_{ij}. The object is to estimate

$$Q = \sum_{(i,j)\in\mathscr{P}} x_{ij},$$

from a systematic sample $S \subseteq \mathscr{P}$ of fixed size $n \times m$. The sampling procedure performed below generates samples, where the probability of including unit (i,j) in S, is proportional to z_{ij}. The procedure uses the (marginal) cumulative values of z_{ij} in two steps, where in each step, systematic sampling is performed analogous to the procedure in 1D. First, the units are divided into N groups according to the values of the first coordinate, from which n groups are sampled, followed by sampling m units, for each of the n groups. Let $z_g = \sum_{j=1}^{M} z_{gj}, g = 1, \ldots, N$, denote the marginal auxiliary variables and $\Delta_0 = 0$, $\Delta_g = \sum_{i=1}^{g} z_i, g = 1, \ldots, N$, the cumulative ones. Then, letting $U_1 \sim \text{unif}[0,1)$, the group g is chosen if there exists an $i = 0, \ldots, n-1$, such that

$$(U_1 + i)\tfrac{1}{n}\Delta_N \in [\Delta_{g-1}, \Delta_g).$$

Assume $z_g < \Delta_N/n, g = 1, \ldots, N$, such that no group is sampled more than once. Next, for each of the n sampled groups g, let $\Delta_{g0} = 0$, $\Delta_{gk} = (\Delta_{gM}/z_g)\sum_{j=1}^{k} z_{gj}, k = 1, \ldots, M$, denote the (scaled) cumulative auxiliary variables within the group g. Then, letting $U_2 \sim \text{unif}[0, 1)$, unit k is chosen if there exists an $j = 0, \ldots, m-1$, such that

$$(U_2 + j)\tfrac{1}{m}\Delta_{gM} \in [\Delta_{g(k-1)}, \Delta_{gk}).$$

Assume again that $z_{gk} < \Delta_{gM}/m, k = 1, \ldots, M$. It can be shown (see calculations below), that $\pi_{ij} = \mathbb{P}((i, j) \in S) = nmz_{ij}/\Delta_N$ from which we get the Horvitz-Thompson estimator

$$\hat{Q}_{nm} = \sum_{(i,j)\in S} \frac{x_{ij}}{\pi_{ij}} = \frac{1}{nm}\Delta_N \sum_{(i,j)\in S} \frac{x_{ij}}{z_{ij}}.$$

The following calculation verifies that the scaling of the auxiliary variables results in the correct inclusion probabilities, using the distribution of U_1, U_2 and the assumptions $z_g < \Delta_N/n$ and $z_{gk} < \Delta_{gM}/m$. We have with $A_g = [\Delta_{g-1}, \Delta_g)$, $A_{gk} = [\Delta_{g(k-1)}, \Delta_{gk})$ and $B_g(i, j) = [i, i + 1)\frac{1}{n}\Delta_N \times [j, j + 1)\frac{1}{m}\Delta_{gM}$

$$\begin{aligned}
\pi_{gk} &= \mathbb{P}((g, k) \in S) \\
&= \sum_{i=0}^{n-1}\sum_{j=0}^{m-1} \mathbb{P}((U_1 + i)\tfrac{1}{n}\Delta_N \in A_g, (U_2 + j)\tfrac{1}{m}\Delta_{gM} \in A_{gk}) \\
&= \sum_{i=0}^{n-1}\sum_{j=0}^{m-1} \int_{A_g \times A_{gk} \cap B_g(i,j)} \frac{nm}{\Delta_N \Delta_{gM}} \,\mathrm{d}u \,\mathrm{d}v \\
&= \frac{nm}{\Delta_N \Delta_{gM}} |A_g \times A_{gk}| \\
&= \frac{nm}{\Delta_N \Delta_{gM}} z_g \frac{\Delta_{gM} z_{gk}}{z_g} \\
&= nm\frac{z_{gk}}{\Delta_N}.
\end{aligned}$$

The relation to the set-up with integrals of a measurement function follows when we let

$$\begin{aligned}
f(x, y) &= NMx_{ij}, \\
f_0(x, y) &= NMz_{ij}, \qquad \text{when} \\
(x, y) &\in [i - 1, i)\tfrac{1}{N} \times [j - 1, j)\tfrac{1}{M}
\end{aligned}$$

and

$$
\begin{aligned}
G_1(x,y) &= \int_0^x g(u)\,\mathrm{d}u \\
&= \frac{1}{\Delta}\Big(\sum_{k=1}^{i-1}\sum_{j=1}^{M} z_{kj} + (Nx - i + 1)\sum_{j=1}^{M} z_{ij}\Big) \\
G_2(x,y) &= \frac{1}{g(x)\Delta}\int_0^y f_0(x,v)\,\mathrm{d}v \\
&= \frac{1}{\sum_{j=1}^{M} z_{ij}}\Big(\sum_{k=1}^{j-1} z_{ik} + (My - j + 1)z_{ij}\Big),
\end{aligned}
$$

where

$$
\begin{aligned}
g(x) &= \frac{\int_0^1 f_0(x,y)\,\mathrm{d}y}{\Delta} = \frac{N\sum_{j=1}^{M} z_{ij}}{\Delta}, \\
\Delta &= \int_{[0,1]^2} f_0(x,y)\,\mathrm{d}x\,\mathrm{d}y = \sum_{(i,j)\in\mathscr{P}} z_{ij},
\end{aligned}
$$

when $(x,y) \in [i-1,i)/N \times [j-1,j)/M$, for $i = 1\ldots,N$ and $j = 1,\ldots,M$.

A.3 Figures

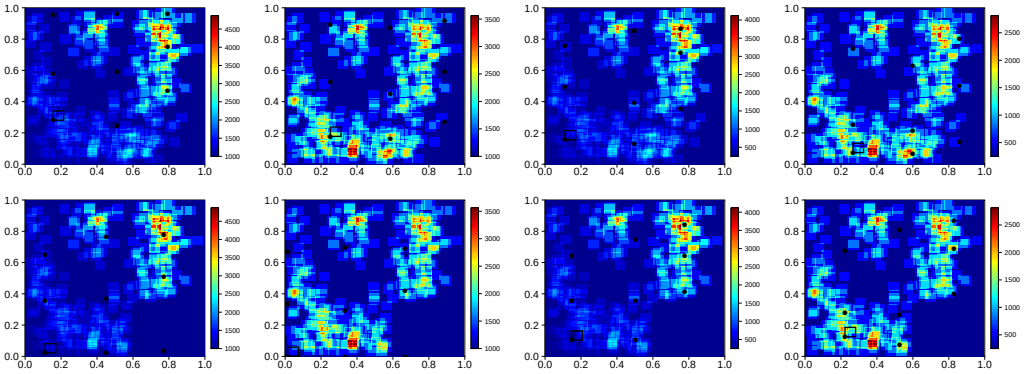

Figure 8: The eight cases of non-uniform auxiliary information f_0 for number estimation, for the second measurement function, together with one example of sample points.

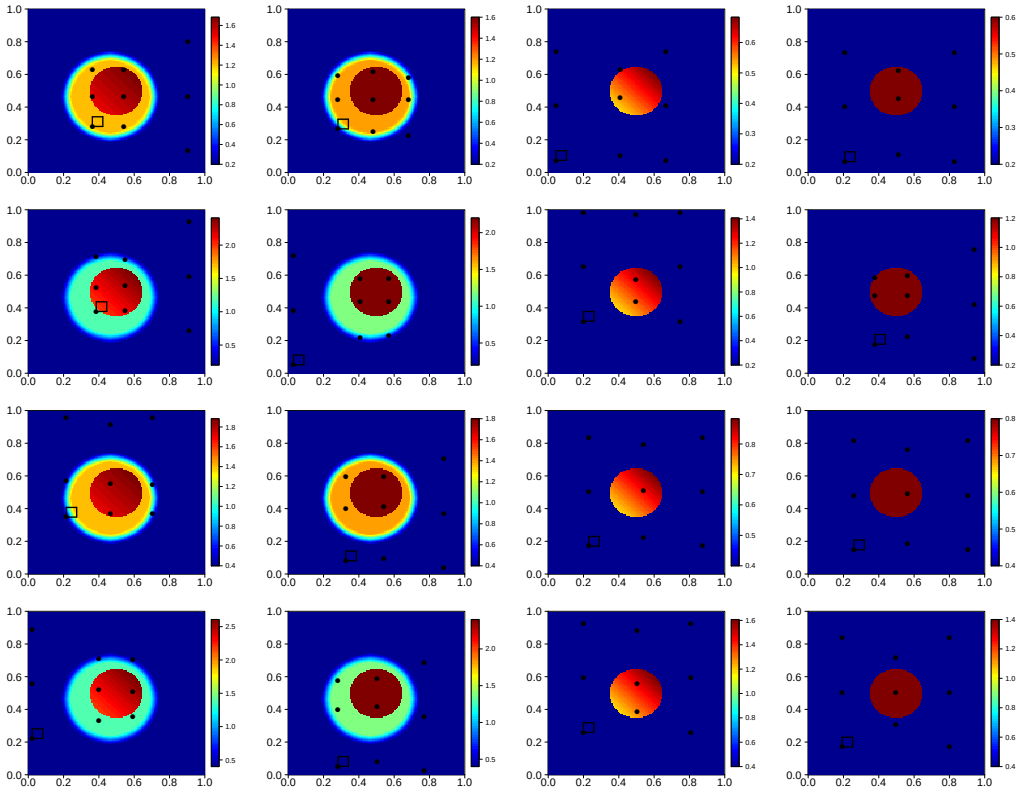

Figure 9: The sixteen cases of non-uniform auxiliary information f_0 for area estimation, together with one example of sample points.

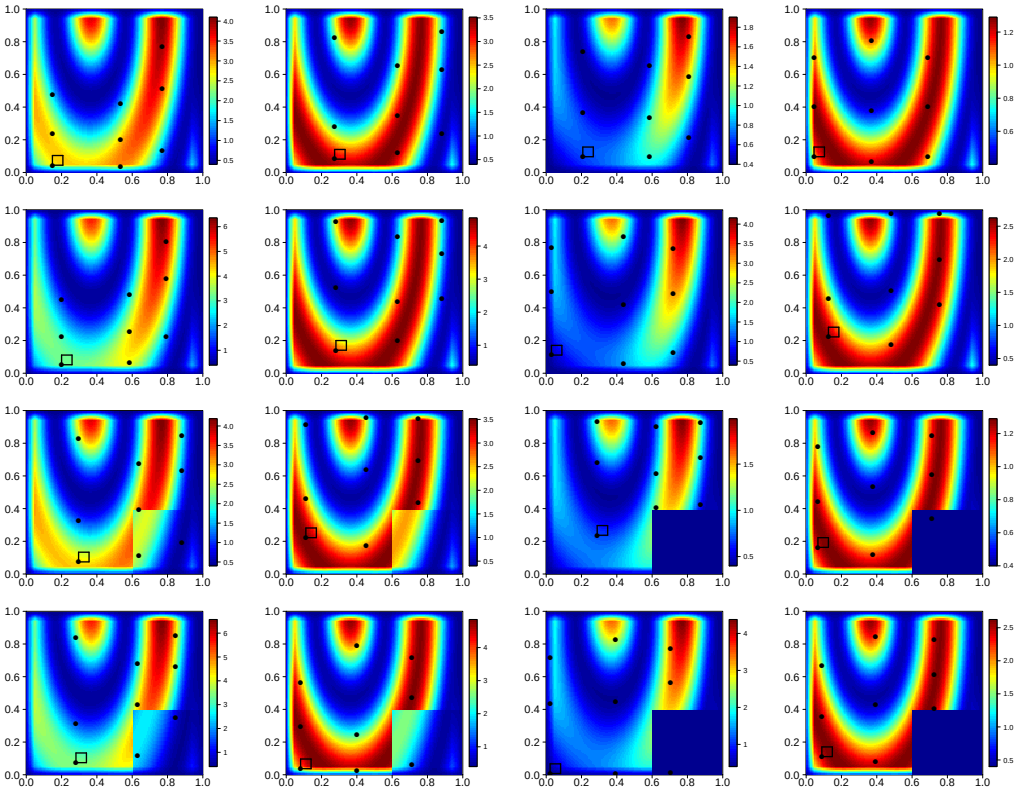

Figure 10: The auxiliary information in the sixteen cases of f_0 for integral estimation, for the second measurement function, together with one example of sample points.