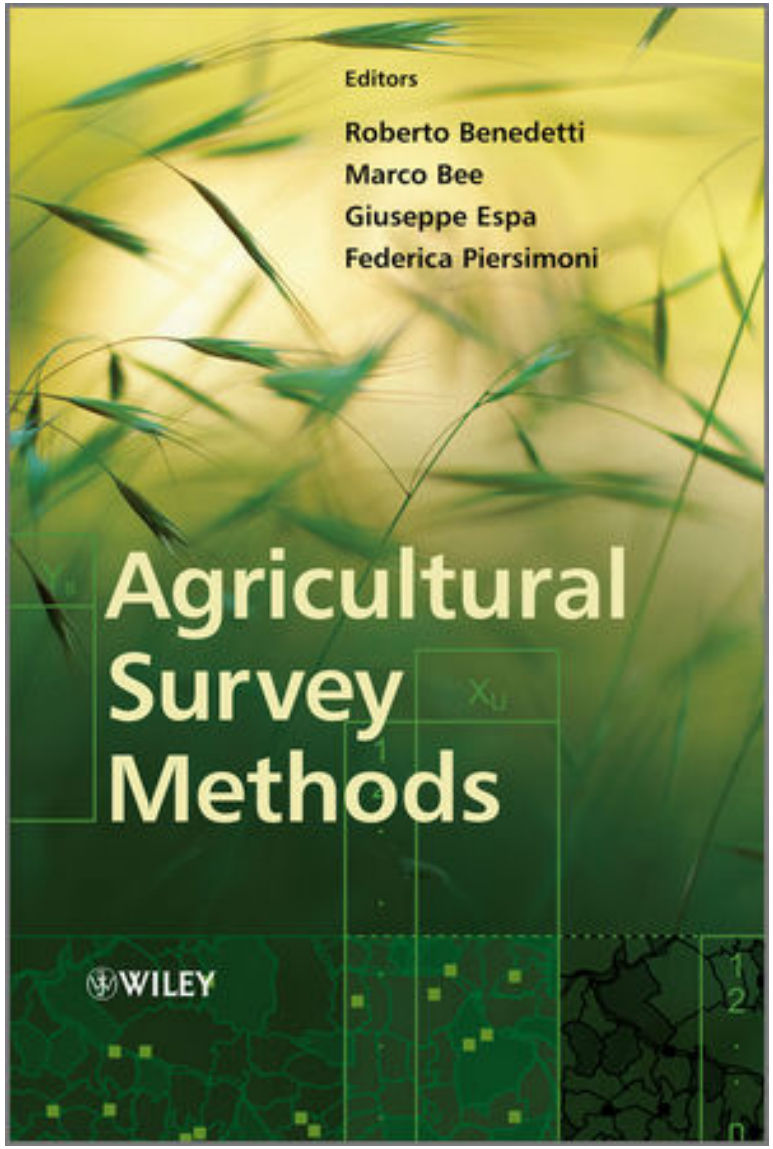
Spatially Balanced Designs

Roberto Benedetti
University of Chieti-Pescara, Italy

Federica Piersimoni
ISTAT, Italian National Statistical Institute, Rome, Italy

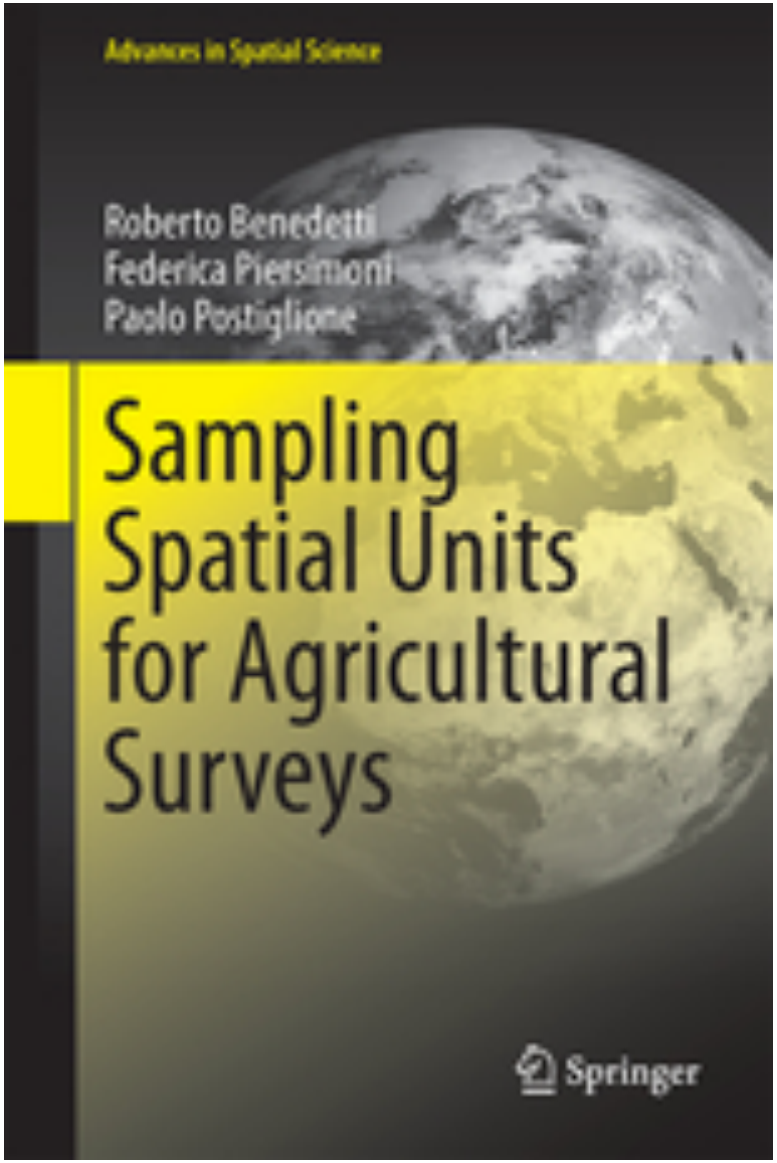
Francesco Pantalone
University of Perugia, Italy

Some references



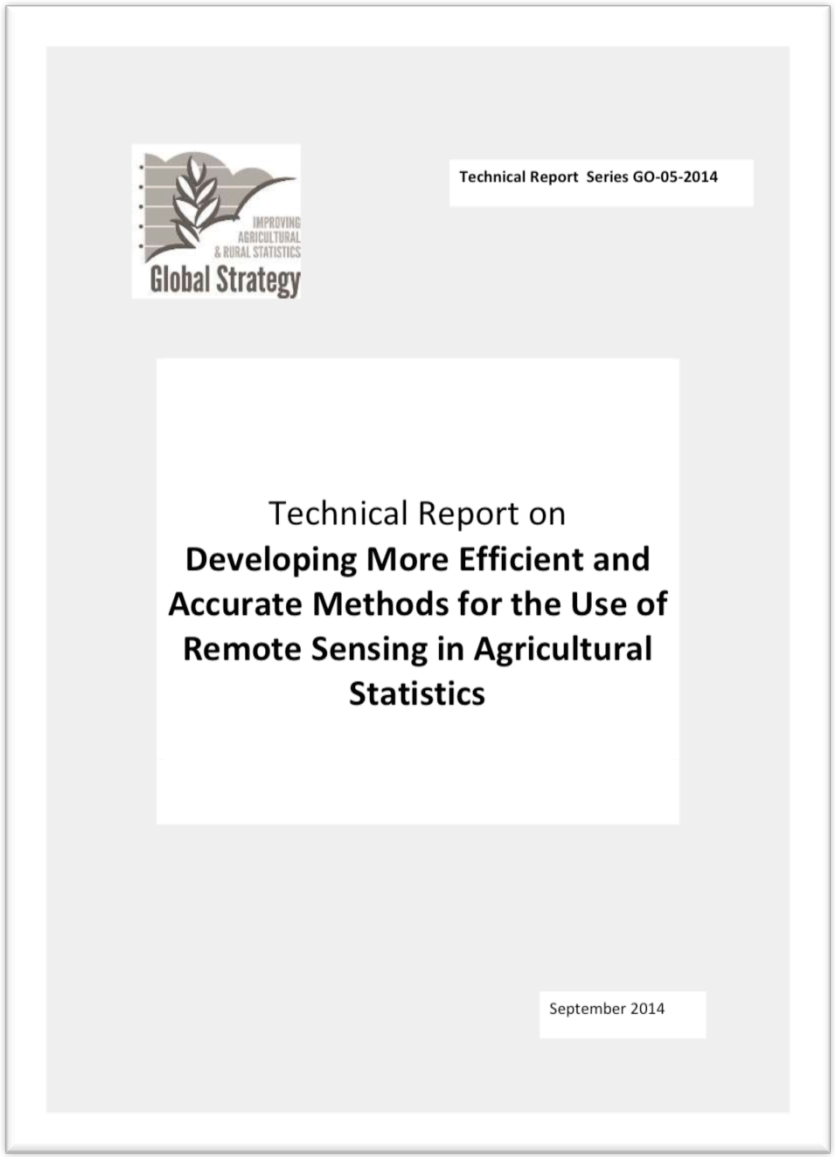
Contents

Some references



Contents

Some references



Contents

Spatially Balanced Sampling

Agricultural and environmental surveys

In agricultural and environmental surveys the main feature of the population is to be geo-referenced.

In these circumstances, usually the units exhibit spatial dependence between them.

Therefore, it is important to consider the spatial distribution of the units as information for selecting samples.

Spatially Balanced Sampling: *Set-up*

- Finite population $U = \{1,\ldots,N\}$.
- Set of q auxiliary variables, $\boldsymbol{X} = \{\boldsymbol{x}_1,\ldots,\mathbf{x}_q\}$.
- Set of h coordinates, commonly obtained by the geo-coding of each unit and usually with $h = 2$, $\boldsymbol{C} = \{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_h\}$.
- Response variable y_i.
- Target of inference $t_y = \Sigma_{i \in U} y_i$.

From the matrix $\boldsymbol{C}$ it is always possible derive, according to any distance function, a matrix that specifies how far all the pairs of units in the population are $\boldsymbol{D}_U = \{d_{kl}; i = 1,\ldots,N, j = 1,\ldots,N\}$.

Spatially Balanced Sampling: *Set-up*

A finite population is a collection of a finite number of identifiable objects or units.

$$U = \{1, 2, \ldots, k, \ldots, N\}$$

A sampling design, p(s), is a probability distribution on Ω (collection of all possible samples) that satisfies:

$$p(s) \geq 0, \text{ all } s \in \Omega,$$

$$\sum_{\Omega} p(s) = 1$$

where s is the outcome of a random variable S.

Spatially Balanced Sampling: *Set-up*

First-order inclusion probability:

$$\pi_k = \Pr(k \in S) = \Pr(I_k = 1) = \sum_{s \ni k} p(s)$$

Second-order inclusion probability:

$$\pi_{kl} = \Pr(k \,\&\, l \in S) = \Pr(I_k I_l = 1) = \sum_{s \ni k \,\&\, l} p(s)$$

Spatially Balanced Sampling: *Set-up*

HT estimator for the population total $t = \sum_U y_k$ is:

$$\hat{t}_{HT} = \sum_s \frac{y_k}{\pi_k}$$

The estimator $\hat{t}_{HT}$ is unbiased for $t = \sum_U y_k$

An unbiased estimator of $V_{HT}(\hat{t}_{HT})$ is given by:

$$\hat{V}_{HT}(\hat{t}_{HT}) = \sum\sum_s \breve{\Delta}_{kl} \breve{y}_k \breve{y}_l \quad \text{where} \quad \breve{\Delta}_{kl} = \Delta_{kl} / \pi_{kl}$$

represents the expanded Δ value, for all $k, l \in U$. Alternatively

$$\hat{V}_{HT}(\hat{t}_{HT}) = \sum\sum_s \frac{1}{\pi_{kl}} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \right) y_k y_l$$

Spatially Balanced Sampling: *Set-up*

Alternative formula for the variance of estimator $\hat{t}_{HT}$ obtained when p(s) is a fixed size sampling design

Yates and Grundy (1953), and Sen (1953):

$$V_{YGS}(\hat{t}_{HT}) = Var_{YGS}(\hat{t}_{HT}) = -\frac{1}{2}\sum\sum_{U}\Delta_{kl}\left(\breve{y}_k - \breve{y}_l\right)^2$$

Spatially Balanced Sampling: *Model-Based*

In the design-based approach, the uncertainty is ensured by $p(s)$, while in the superpopulation approach, the randomness is provided from the model ξ. For this reason, this approach is also called model-based survey sampling.
Using the superpopulation approach, we estimate the population total as follows. The population total, t, can be decomposed into:

$$t = \sum_s y_k + \sum_{\bar{s}} y_k = t_{y_s} + t_{y_{\bar{s}}}$$

In other words, the population total is the sum of the sample total t_{y_s} and the corresponding non-sample total $t_{y_{\bar{s}}}$. Obviously, after the sample has been drawn, the sum of the sample total t_{y_s} is known, and the estimation problem is reduced to predicting $t_{y_{\bar{s}}}$ given t_{y_s}. Given the superpopulation model ξ, the aim is to choose the best predictor $\hat{t}_{y_{\bar{s}}}$ of $t_{y_{\bar{s}}}$, and a sample s, so that we minimize the sample error, $\hat{t} - t = \hat{t}_{y_{\bar{s}}} - t_{y_{\bar{s}}}$.

Spatially Balanced Sampling: *Model-Based*

The best predictor $\hat{t}$ is the member of the acceptable predictors that has the smallest value of $E_\xi\left[\left(\hat{t}-t\right)^2 \middle| s\right]$. Besides, we aim to choose s that minimizes $E_\xi\left[\left(\hat{t}-t\right)^2 \middle| s\right]$ across the set of all possible s that are practical and satisfy the resource constraints. The optimal predictor ($\hat{t}$) and optimal sample (s) constitute the optimal strategy for t under the assumed superpopulation model ξ.

Spatially Balanced Sampling: *Model-Based*

Consider an un-sampled location $\mathbf{z}_0$. The main aim of geostatistics is to predict $y(\mathbf{z}_0)$. It would seem reasonable to estimate $y(\mathbf{z}_0)$ using a weighted average of the values at observed locations $y(\mathbf{z}_i)$, i=1,2,..,n, with weights given by some decreasing function of the distance between the unobserved and observed sites. So, the predictor of $y(\mathbf{z}_0)$ can be defined as:

$$\hat{y}(\mathbf{z}_0) = \sum_i \lambda_i y(\mathbf{z}_i)$$

A simple and popular spatial prediction method is kriging. This method uses a model of spatial continuity, or dependence.

The main purpose of kriging is to optimally determine the weights λ_i. A predictor is defined by first constructing a function that measures the loss sustained by using $\hat{y}(\mathbf{z}_0)$ as a predictor of $y(\mathbf{z}_0)$. The squared loss function is most often used in practical applications.

Generally, the theory aims at finding estimators that minimize the average loss. In this case, the loss can be expressed in terms of the mean squared prediction error (MSPE) as:

Spatially Balanced Sampling: *Model-Based*

$$E\left[y(\mathbf{z}_0) - \hat{y}(\mathbf{z}_0)\right]^2$$

Kriging computes the best linear unbiased predictor (BLUP), $\hat{y}(\mathbf{z}_0)$, based on a stochastic model of the spatial dependence defined by the expectation, μ(z), and covariance function, C(**h**), of the random field.

Spatially Balanced Sampling: *Model-Based*

Consider a continuous variable. A spatial model that satisfies the first-order Markov property

$$\Pr\left\{y\left(\mathbf{z}_i\right)\middle|y\left(\mathbf{z}_j\right), j\in D, j\neq i\right\} = \Pr\left\{y\left(\mathbf{z}_i\right)\middle|y\left(\mathbf{z}_j\right), j\in N(i), j\neq i\right\}$$

is the auto-normal or conditional autoregressive model (CAR, Besag 1974). It assumes that the conditional density functions of each random variable with respect to the others is Gaussian and can be expressed as:

$$f\left[y\left(\mathbf{z}_i\right)\middle|y\left(\mathbf{z}_j\right)\right] = f\left[y\left(\mathbf{z}_i\right)\middle|y\left(\mathbf{z}_j\right), j\in N(i), j\neq i\right] =$$

$$= \left(2\pi\sigma_i^2\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma_i^2}\left[y\left(\mathbf{z}_i\right)-\mu_i-\sum_{i\neq j}c_{ij}\left(y\left(\mathbf{z}_j\right)-\mu_j\right)\right]^2\right\}$$

where $\mu_i = E\left(y(\mathbf{z}_i)\right)$, and c_{ij} denote spatial dependence parameters that are only non-zero if $\mathbf{z}_j \in N(i)$.

Spatially Balanced Sampling: *Model-Based*

From these definitions it follows that:

$$E\left[y\left(\mathbf{z}_i\right)\middle|y\left(\mathbf{z}_j\right), j\in N(i), j\neq i\right]=\mu_i+\sum_{i\neq j}c_{ij}\left(y\left(\mathbf{z}_j\right)-\mu_j\right)$$

and:

$$Var\left[y\left(\mathbf{z}_i\right)\middle|y\left(\mathbf{z}_j\right), j\in N(i), j\neq i\right]=\sigma_i^2$$

Spatially Balanced Sampling: *Model-Based*

Let $\boldsymbol{\varepsilon} \propto N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$, where $\varepsilon(\mathbf{d}_i)$ is the variable associated with site $\mathbf{z}_i$. A random field is said to be Gaussian SAR (Whittle 1954) if:

$$y(\mathbf{z}_i) = \mu_i + \sum_{i \neq j} b_{ij}\left[y(\mathbf{z}_j) - \mu_j\right] + \varepsilon(\mathbf{z}_i)$$

where $b_{ii}=0$. In a matrix notation model, the above equation can be written as:

$$(\mathbf{I} - \mathbf{B})(\mathbf{y} - \boldsymbol{\mu}) = \boldsymbol{\varepsilon}$$

If $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ **Y** is multivariate normal such that:

$$\mathbf{y} \propto MVN\left(\boldsymbol{\mu}, \sigma^2\left[(\mathbf{I} - \mathbf{B})^{-1}(\mathbf{I} - \mathbf{B}^t)^{-1}\right]\right)$$

References

Survey Sampling

- Chambers RL, Clark RG (2012). An introduction to model-based survey sampling with applications. Oxford University Press, Oxford.
- Cochran WG (1977). Sampling techniques. John Wiley & Sons, Inc., New York.
- Fuller WA (2009). Sampling statistics. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Lohr SL (2010), Sampling: Design and Analysis, 2nd Ed., Brooks/Cole, Boston, Mass.
- Lumley T (2010). Complex surveys. A guide to analysis using R. John Wiley & Sons, Inc, Hoboken, New Jersey.
- Särndal CE, Swensson B, Wretman J (1992). Model assisted survey sampling. Springer, New York.

Spatial Statistcs

- Anselin L (1988). Spatial econometrics, methods and models. Kluwer Academic, Boston.
- Besag J (1974). Spatial interaction and the statistical analysis on lattice systems. Journal of the Royal Statistical Society, Series B, 36: 192-236.
- Cressie N (1993). Statistics for spatial data. John Wiley & Sons, Inc., New York.
- Diggle PJ (2003). Statistical analysis of spatial point patterns. Arnold publishers, London.
- Haining (2003). Spatial data analysis: theory and practice. Cambridge University Press, Cambridge.
- Ripley BD (1981) Spatial Statistics, Wiley
- Whittle P (1954). On stationary processes in the plane. Biometrika, 41: 434–449.

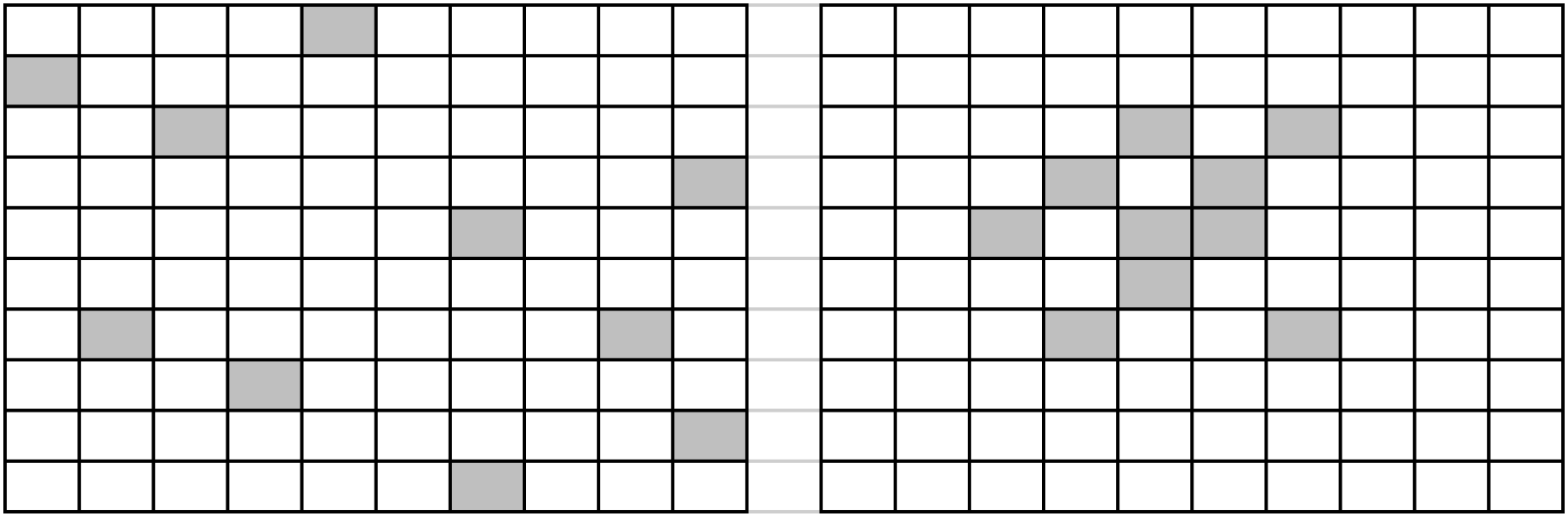
Spatially Balanced Sampling

*Some thoughts about **X** and **C** in spatial surveys.*

- If *U* is a list of regularly or irregularly shaped polygons defined *ad hoc*, ***C*** is always available and ***X*** can be constructed summarizing within each polygon a classification of remotely sensed data (unless an overlay of ***C*** with a cadaster is possible).
- If *U* is a list of points, ***X*** can be only represented by a design matrix of codes of a land use classification of remotely sensed data.
- If *U* is a list of economic or social units, ***C*** is rarely obtainable (it depends on the availability of accurate cadastral maps) and should be made by a map of polygons representing parcels of land used by each holding, while ***X*** is usually filled with administrative data sources.

Spatially Balanced Sampling

Consider the two following samples obtained by SRSWOR



The *p(s)* of these two samples are exactly the same, *p(S) = 1/C(N,n)*

Spatially Balanced Sampling

Spatially Balanced Samples

How to take into account the spatial information while designing a sample?

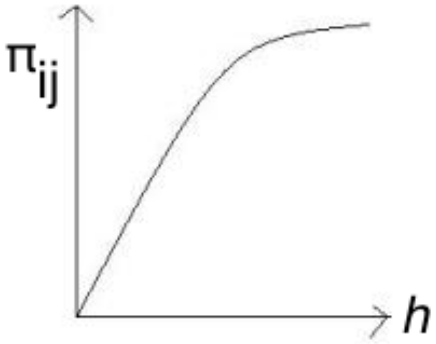
Spatially balanced samples: samples well-spread over the population of interest. In this way, it could be possible capture the spatial heterogeneity of the population.

Some theoretical motivations:

- Yates-Grundy-Sen formulation of the HT variance.
- Anticipated Variance.
- Lemma decomposition.

Spatially Balanced Sampling: motivation A

The variogram (or semi variogram) $\gamma_y(h)$ whose shape is a valuable information to choose on how and to what extent the variance of y is or not a function of the distance between the statistical units.

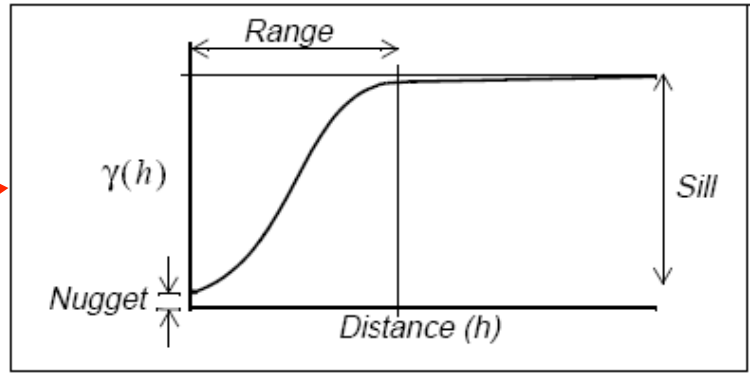


- Yates-Grundy-Sen formulation of the HT variance:

$$V\left(\hat{Y}_{HT}\right) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\left(\pi_{ij} - \pi_i\pi_j\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Semivariogram

$$\frac{1}{2}Var\left[y(\mathbf{z}) - y(\mathbf{z}+\mathbf{h})\right] = \gamma(\mathbf{h})$$



Spatially Balanced Sampling: motivation B

We wish to derive a model that relates each $\mathbf{y}_v$ with the **X** observed in past surveys or other data sources. We assume that our *prior* knowledge on the finite population can be viewed as if it were a sample from an infinite superpopulation and that a model ξ defines its characteristics. To design a survey, we should thus search for the optimal anticipated variance (*AV*) of the estimator of the population total. This can be defined as the variance of the random variable $(\hat{t} - t)$ under both the design and the model

$$AV\left(\hat{t}-t\right) = E_\xi\left\{E_s\left[\left(\hat{t}-t\right)^2\right]\right\} - \left[E_\xi\left\{E_s\left(\hat{t}-t\right)\right\}\right]^2$$

A typical assumption is a linear model that relates a target **y** and an auxiliary **x**

$$\begin{cases} y_k = \mathbf{x}_k^t \boldsymbol{\beta} + \varepsilon_k \\ E_\xi\left(\varepsilon_k\right) = 0 \\ V_\xi\left(\varepsilon_k\right) = \sigma_k^2 \\ E_\xi\left(\varepsilon_k \varepsilon_l\right) = \sigma_k \sigma_l \rho_{kl} \quad k \neq l \end{cases}$$

Spatially Balanced Sampling: motivation B

Anticipated Variance

$\boldsymbol{x}_i$ is a vector of auxiliary variables, β is a vector of coefficient regression, ρ_{ij} is the autocorrelation coefficient and E_m, Var_m and Cov_m denote, respectively, expectation, variance and covariance with respect to the model.

The Anticipated Variance (Isaki and Fuller 1982) of HT estimator under the model is (Grafstrom and Tillè 2013)

$$AV\left(\hat{t}_{HT} - t\right) = E_s\left[\left(\sum_{k\in s}\frac{x_k}{\pi_k} - \sum_{k\in U} x_k\right)^T \beta\right]^2 + \sum_{k\in s}\sum_{l\in s}\sigma_k\sigma_l\rho_{kl}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l}$$

Spatially Balanced Sampling: motivation B

Uncertainty can be splitted into two terms:

1. $E_s\left[\left(\sum_{i\in S}\frac{x_i}{\pi_i}-\sum_{i\in U}x_i\right)'\beta\right]^2$ can be reduced through the use of balanced sampling (Deville and Tillè 2004)

2. $\sum_{i,j\in U}\sigma_i\sigma_j\rho_{ij}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}$ can be reduced exploiting spatial information → if ρ_{ij} decrease with respect to distance between units, then selecting units far apart reduces this term

Spatially Balanced Sampling: motivation C

The decomposition lemma

states that (Knottnerus, 2003, p. 87):

$$\sigma^2_{\bar{y}} = V_s\left(\bar{\bar{y}}_s\right) + \frac{n-1}{n} E_s\left(S^2_{\bar{y},s}\right)$$

It can be seen that the HT estimator can be more efficient by setting the first-order inclusion probabilities in such a way that y_k/π_k is approximately constant and/or by defining a design $p(s)$ that increases the expected within sample variance. The intuitive explanation for this is that if a sample s contains as much information as possible, the uncertainty in the estimation process is clearly reduced to zero. This consideration suggests that we should find a rule that makes the probability $p(s)$ of selecting a sample s proportional, or more than proportional, to its variance S^2. This variance is unknown, because it is relative to the target, unobserved variable **y**. Thus, this is a purely theoretical topic unless we can find auxiliary information for s.

Spatially Balanced Sampling: motivation C

Decomposition Lemma

This consideration suggests that we should find a rule that makes the probability *p(s)* of selecting a sample *s* proportional, or more than proportional, to its variance S^2. This variance is unknown, because it is relative to the target, unobserved variable **y**. Thus, this is a purely theoretical topic unless we can find auxiliary information for *s*.

When dealing with spatially distributed populations, a promising candidate for this rule is the distance between units, as evidenced in spatial interpolation literature (Ripley 1981, Cressie 1993). This is because it is often highly related to the variance of variables observed on a set of geo-referenced units.

Spatially Balanced Sampling: practical motivations

There could be a lot of different reasons why it is appropriate to select samples which are spatially well distributed:

1. y has a linear or monotone spatial trend;
2. there is spatial autocorrelation, i.e. close units have data more similar than distant units;
3. the y shows to follow zones of local stationarity of the mean and/or of the variance, i.e. a spatial stratification exists in observed phenomenon;
4. the units of the population have a spatial pattern which can be clustered, i.e. the intensity of the units varies across the study region.

The Index of Spatial Balance

The Voronoi polygon for unit *k* of a generic sample *s* includes all the population units closer to *k* than to any other unit in the sample. Let

$$v_k = \sum_{i \in VP(k)} \pi_i$$

be the sum of the inclusion probabilities of the units in the *k*-th Voronoi polygon VP(*k*). Then, for any sample unit, we will have an expected value E(v_k)=1. Additionally, all the v_ks should be close to 1 for a spatially balanced sample (Steven and Olsen 2004). Thus, the index V(v_k) (the variance of the v_k) can be used as a measure of the *spatial balance* of a sample. Obviously, a lower value of V(v_k) implies a *good* spatially balanced sample.

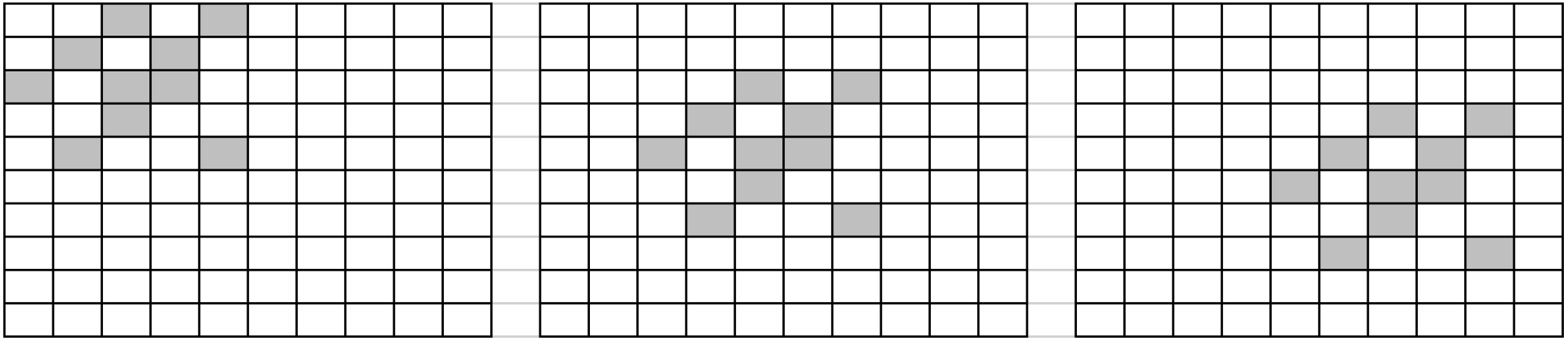
$$V\left(v_k\right) = \frac{\sum_{k \in s} \left(v_k - 1\right)^2}{n}$$

Contents

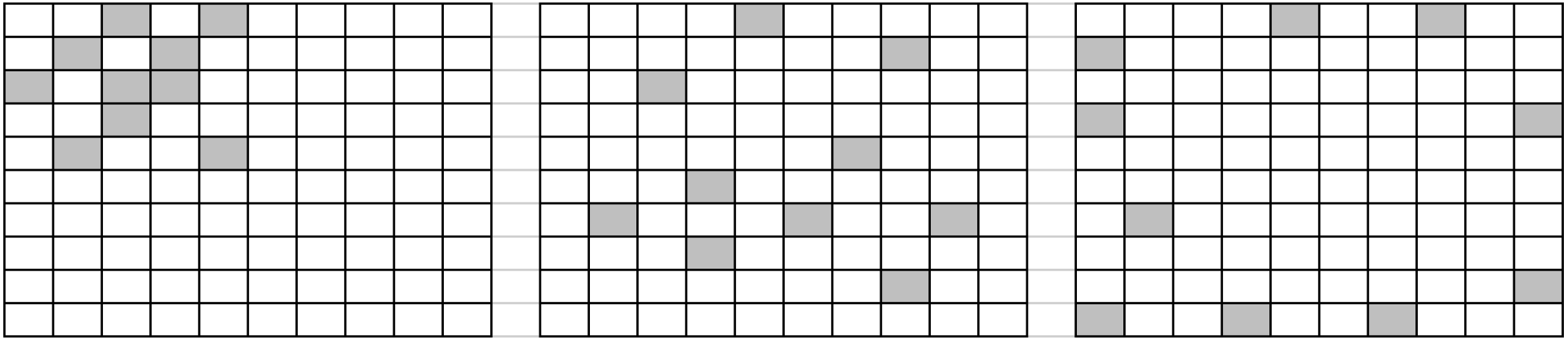
- *Balanced Sampling CUBE*
- *Systematic Sampling*
- *Maximal Stratification*
- *Optimal Sampling Designs*
- *DUST design*
- *Sampling Plans that Exclude Adjiacent Units*
- *Generalized Random Tessellation Sampling (GRTS)*
- *Spatially Correlated Poisson Sampling (SCPS)*
- *Local Pivotal Method (LPM)*

The Balanced Sampling and Cube Method

Which is the “best” sample ?



Which is the “best” sample ?



OK, we like the position in the middle, but why ?

The Balanced Sampling and Cube Method

Select samples with the important property :

$$\sum_{k\in s} d_k x_{kj} = \hat{t}_{HT,x_j} = t_{x_j} = \sum_{k\in U} x_{kj} \;\; \forall \; j = 1,\ldots,q$$

Note that many sampling designs can be viewed as particular cases of balanced sampling. For example, stratified sampling can also be defined as a design respecting the constraint :

$$\sum_{k\in s} d_k \varphi_{kh} = \sum_{k\in U} \varphi_{kh} = N_h, \;\; \forall \; h = 1,\ldots,H$$

Where φ_{kh}s are indicator variables equal to 1 if the unit *k* is in the stratum *h,* and 0 otherwise.

The algorithm consists of two main procedures: the *flight* and *landing* phases. During the first phase, the constraints are always exactly satisfied. The objective is to randomly round-off almost all the π_ks to 0 or 1. The landing phase addresses the fact that the constraint cannot always be exactly satisfied.

The Balanced Sampling and Cube Method

1. Generate a vector $\mathbf{u}(t)=\{u_k(t)\}\neq 0$, not necessarily random, such that $\mathbf{u}(t)$ belongs to the kernel[1] of $\mathbf{A}$ (i.e., ker($\mathbf{A}$)) and $u_k(t)=0$ if $\pi_k(t)$ is an integer.
2. Compute the largest values of $\lambda_1(t)$ and $\lambda_2(t)$ (λ_1^* and λ_2^*) such that $0\leq\boldsymbol{\pi}(t)+\lambda_1(t)\mathbf{u}(t)\leq 1$ and $0\leq\boldsymbol{\pi}(t)-\lambda_2(t)\mathbf{u}(t)\leq 1$, obviously $\lambda_1(t)>0$ and $\lambda_2(t)>0$.
3. Compute the next $\boldsymbol{\pi}$ using

$$\boldsymbol{\pi}(t+1)=\begin{cases}\boldsymbol{\pi}(t)+\lambda_1^* u(t) & \text{with probability } \delta(t)\\ \boldsymbol{\pi}(t)-\lambda_2^* u(t) & \text{with probability } 1-\delta(t)\end{cases}$$

where $\delta(t)=\lambda_2^*/\left(\lambda_1^*+\lambda_2^*\right)$.

The three steps are iterated until we cannot perform Step 1. In the flight phase, finding a vector in ker($\mathbf{A}$) can be quite computationally expensive. To overcome this difficulty, Chauvet and Tillé (2006) developed a faster algorithm for implementing the three steps. The idea consists of replacing $\mathbf{A}$ with a smaller matrix $\mathbf{B}$, where $\mathbf{B}$ is a sub-matrix of $\mathbf{A}$ containing only $q+1$ columns of $\mathbf{A}$.

The Balanced Sampling and Cube Method

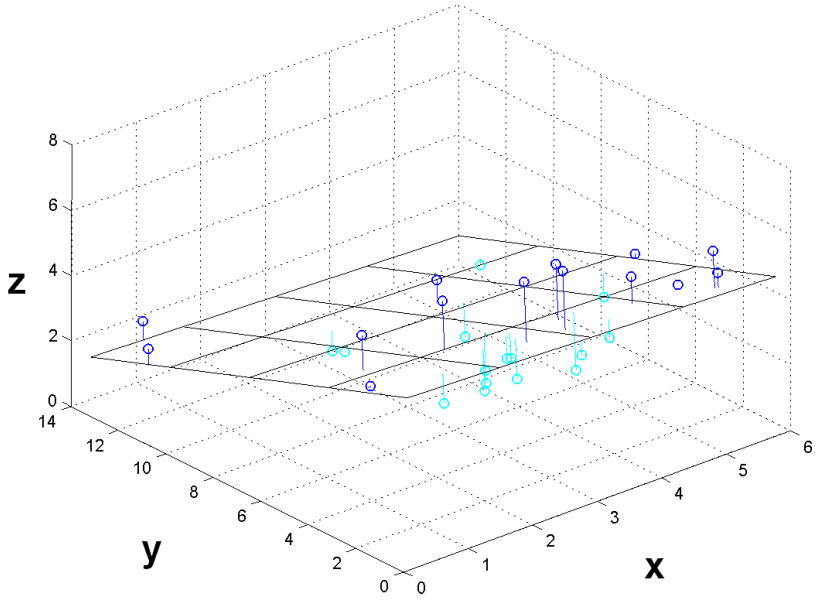
Following Tillé (2011), we can summarize the main features of balanced sampling as follows:

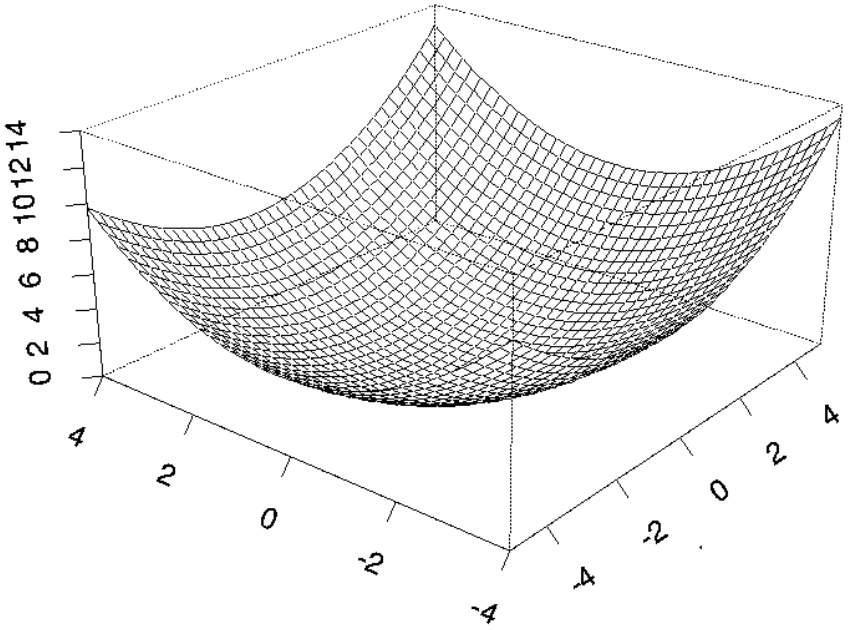
- It increases the accuracy of the HT estimator, because its variance depends only on the regression residuals of the variable of interest by the balancing variables.
- It protects against large sampling errors, because the most unfavorable samples have a null probability of being selected.
- It protects against a misspecification of the model within a model-based inference.
- It can ensure that the sample sizes in planned domains are not too small, or even equal to zero. By adding the indicator variables of the planned domains to the list of balanced auxiliaries, we can fix the sample size for each domain.

The Balanced Sampling and Cube Method

Constraint on the 1st moment =

Assume a linear spatial trend





Constraint on the 1st and 2nd moments = assume a quadratic spatial trend

Constraint on a Penalized Spline that fit the spatial trend (Breidt and Chauvet, 2012)

The Balanced Sampling and Cube Method

```
> library(sampling)
> n <- 100
> N <- 1000
> set.seed(200694)
> par(mar=c(1,1,1,1), xaxs="i",yaxs="i")
> plot(framepop$xc,framepop$yc,
  +     axes=F,cex=0.5,pch=19,
  +     xlim=c(0,1),ylim=c(0,1))
> box()
> set.seed(200694)
> pik <- rep(n/N,N)
> X   <- as.matrix(cbind(framepop$xc,framepop$yc))
> bal <- samplecube(X,pik,comment=TRUE,method=1)
```

The Balanced Sampling and Cube Method

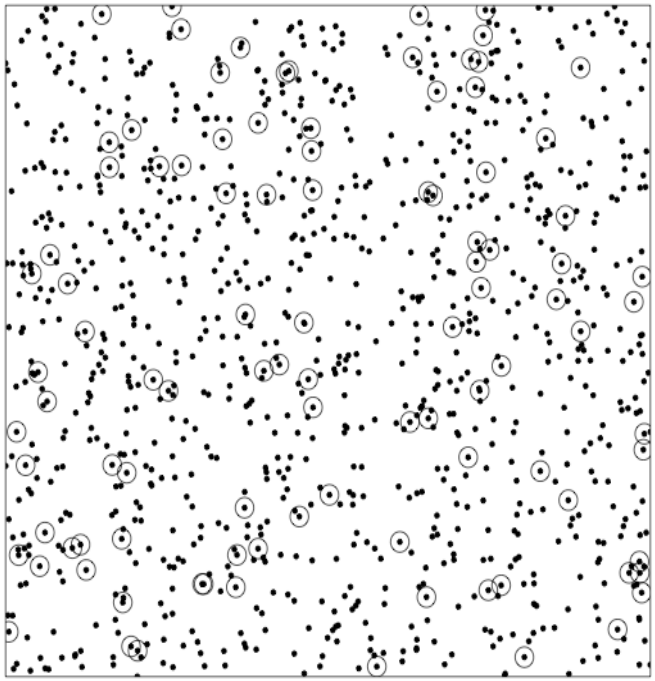
```
BEGINNING OF THE FLIGHT PHASE
The matrix of balanced variable has 2  variables and  1000  units
The size of the inclusion probability vector is  1000
The sum of the inclusion probability vector is  100
The inclusion probability vector has  1000  non-integer elements
Step 1
BEGINNING OF THE LANDING PHASE
At the end of the flight phase, there remain  2 non integer
probabilities
The sum of these probabilities is  0.6969337
This sum is  non-integer
The linear program will consider  3  possible samples
The mean cost is  0.0003727694
The smallest cost is  0.0001680449
The largest cost is  0.0006208612
```

The Balanced Sampling and Cube Method

```
The cost of the selected sample is
0.0006208612

QUALITY OF BALANCING

    TOTALS HorvitzThompson_estimators
Relative_deviation

1 494.0807                          492.2161
-0.3773813

2 494.8565                          496.1754
0.2665137


>sum(bal)

[1] 105
```

The Balanced Sampling and Cube Method

```
>set.seed(200694)
>X   <- as.matrix(cbind(pik,framepop$xc,framepop
$yc))
>bal1 <- samplecube(X,pik,comment=TRUE,method=1)
BEGINNING OF THE FLIGHT PHASE
QUALITY OF BALANCING
      TOTALS HorvitzThompson_estimators
Relative_deviation
pik 100.0000                    100.0000
1.406875e-12
2   494.0807                    491.0461
-6.141966e-01
3   494.8565                    495.9795
2.269404e-01
>sum(bal1)
[1] 100
>framebal <- framepop[bal1==1,]
>points(framebal$xc,framebal$yc, pch=1, cex=2)
>sbi(ds,rep(n/N,N),(1:1000)[bal1==1])
[1] 0.3242424
```



Some references

Some references about Balanced Sampling.

- Breidt FJ, Chauvet G (2012). Penalized balanced sampling. Biometrika, 99, 4, 945–958.
- Chauvet G, Tillé Y (2006). A fast algorithm of balanced sampling. Computational Statistics, 21, 53-62.
- Deville J-C, Tillé Y (2004). Efficient balanced sampling: The cube method. Biometrika, 91, 4, 893–912.
- Tillé Y (2006). Sampling algorithms. Springer series in statistics. Springer, New York.
- Tillé Y (2011). Ten years of balanced sampling with the cube method: An appraisal. Survey Methodology, 37: 215–226.
- Tillé Y, Favre AC (2005). Optimal allocation in balanced sampling. Statistics & Probability Letters, 74: 31–37.

Systematic Sampling

Systematic sampling has a long tradition in survey sampling. When applied to a list frame of individuals or families, it can be referred to as the *every r*-th *rule*. The main parameter of the method is *r*, which is the number of units between each unit selected from the sample, according to a given ordering of the population. The randomization principle is typically retained by using a random starting point and a fixed interval *r*.

This scheme is a widely used technique in survey sampling because of its simplicity, particularly when the units are selected with equal probability, but also with probabilities proportional to an auxiliary size measure.

Systematic sampling is also a common design for spatially distributed populations. If the ordering uses the coordinate system that geo-codes the population frame, it has the additional advantage that it has a good spatial coverage. It is an efficient method for sampling autocorrelated populations.

Systematic Sampling

Disadvantages

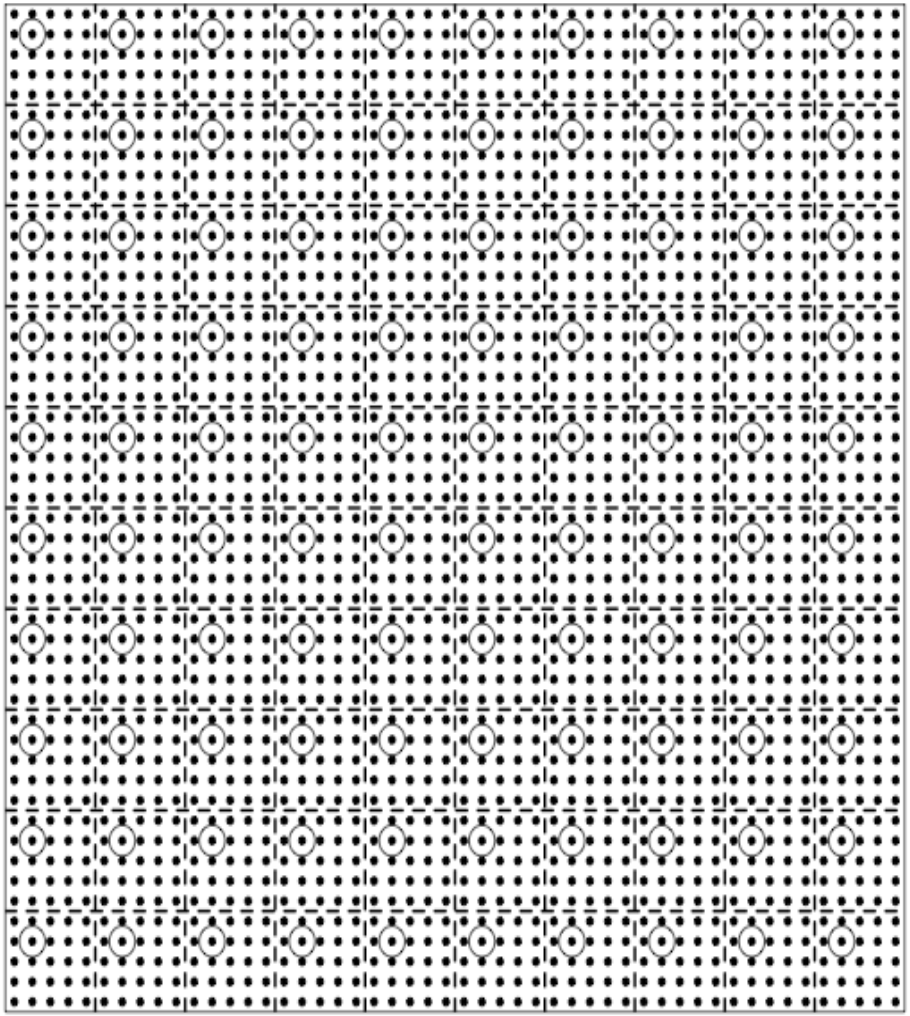
- Because all the second-order probabilities are equal to zero within each step *r*, there is no unbiased method for estimating the sampling variance.
- The ratio *N*/*n* is not typically an integer, so it is often impossible to find a step *r* that is suitable for finding exactly *n* sampling units. This practical difficulty may become relevant when the selection should be repeated in groups of homogeneous units of the population, or in spatial frames where we need at least a pair (r_x, r_y) of step parameters (one for each dimension).

Systematic Sampling

```
> set.seed(200694)
> startx <- sample(1:5,1)
> starty <- sample(1:5,1)
> datasys <- matrix(0,2500,3)
> init <- 0
> for (xc in seq(0.01,0.99,0.02))
+ {
+   for (yc in seq(0.01,0.99,0.02))
+   {
+     init <- init + 1
+     datasys[init,1] <- xc
+     datasys[init,2] <- yc
+     datasys[init,3]=ifelse((abs((xc %% 0.1)-(startx/50-0.01))
+               +abs((yc %% 0.1)-(starty/50-0.01)) < 0.001),1,0)
+   }
+ }
```

Systematic Sampling

```
> par(mar=c(1,1,1,1), xaxs="i",
+ yaxs="i")
> plot(datasys[,1],datasys[,2],
+ axes=F,cex=0.5,
+ pch=19,xlim=c(0,1),ylim=c(0,1))
> for (i in seq(0.1,0.9,0.1))
+ {
+   abline(h=i,lty=2,lwd=2)
+   abline(v=i,lty=2,lwd=2)
+ }
> points(datasys[datasys[,3]
+      ==1,1],datasys[datasys[,3]
+      ==1,2],pch=1, cex=2)
> box()
```



Maximal Stratification

An intuitive way to produce samples that are well spread over the population, widely used by practitioners, is to stratify the units of the population on the basis of their location.

A maximal stratification, i.e. partitioning the study in as many strata as possible and selecting one or two units per stratum. The basic principle is to extend the use of systematic sampling to two or more dimensions. The problems arising from such a fragmentation of the population is often reflected in the following unfavorable issues:

- If we allocate less than a fixed threshold (say T_h) to a generic stratum *h,* we typically have $n_h = T_h$ with a consequent fictitious increase of the sample size.

Maximal Stratification

- If the effective number of observed units r_h in a stratum is less than 2 because of non-responses, it is no longer possible to estimate the variability and accuracy of the sample. If r_h=0, it is not even possible to produce point estimates.
- It is difficult to manage panel rotations or more general sample coordination (between multiple surveys) in an under-represented stratum.

In these cases, little can be done unless we accept solutions that are not methodologically desirable, but that introduce as few bias as possible. A practical solution is the *posterior* aggregation of similar strata.

Maximal Stratification

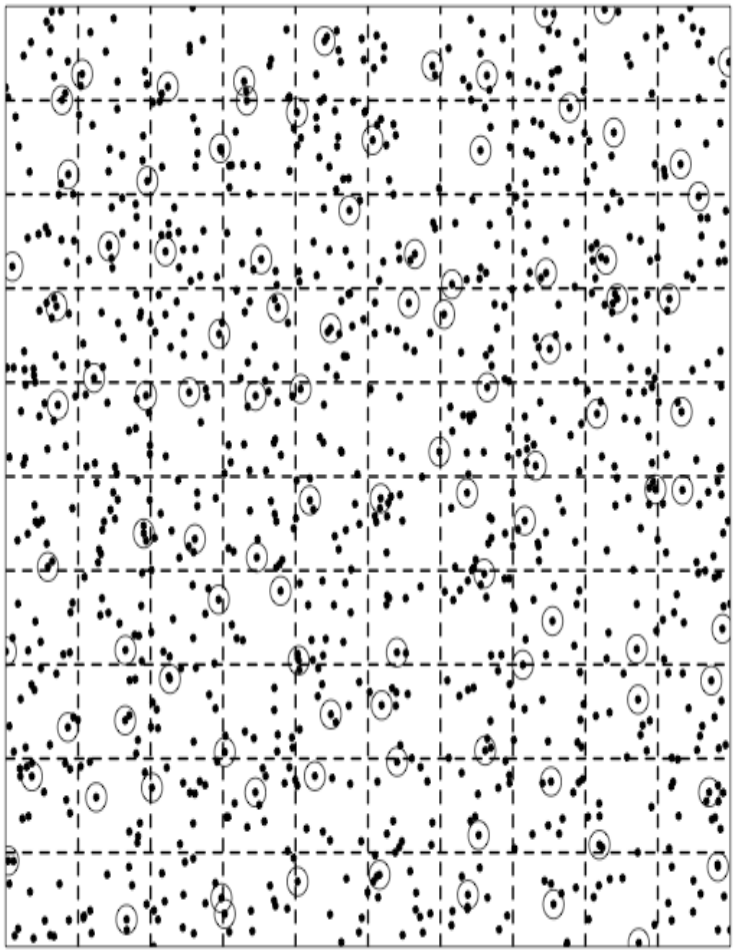
```
> library(sampling)
> library(survey)
> n <- 100
> N <- 1000
> set.seed(160964)
> framepop <- data.frame(id = 1:N,
+ xc = runif(N), yc = runif(N))
> yobs<-(exp((framepop$xc-0.5)^2)+exp((framepop$yc-0.5)^2))
> yobs <- 100-((yobs-min(yobs))/(max(yobs-
        + min(yobs))))*100+(rnorm(N)+5)*5
> q1obs <- sample(1:3,N,replace=T)
```

Maximal Stratification

```
> q2obs <- as.numeric(cut(yobs,quantile(yobs,probs =
+          seq(0, 1, 0.2))))
> q2obs[is.na(q2obs)] <- 1
>framepop<-cbind(framepop,strataid2
+                    = floor(framepop$xc*10)*10
+                    + floor(framepop$yc*10))
>table(framepop$strataid2)
>set.seed(200694)
>str <- strata(framepop,"strataid2",size=rep(1,100),
+       method="srswor")
>str <- getdata(framepop,str)
>table(str$strataid2)
```

Maximal Stratification

```
> par(mar=c(1,1,1,1),
+      xaxs="i",yaxs="i")
> plot(framepop$xc,framepo$yc,
+  axes=F,cex=0.5,
+  pch=19,xlim=c(0,1),
+  ylim=c(0,1))
> for (i in seq(0.1,0.9,0.1))
+ {
+   abline(h=i,lty=2,lwd=2)
+   abline(v=i,lty=2,lwd=2)
+ }
> box()
> points(str$xc,str$yc,pch=1,cex=2)
```



Some references

Some references about Maximal Stratification & Systematic Sampling.

- Breidt FJ (1995). Markov chain designs for one-per-stratum sampling. Survey Methodology, 21: 63-70.
- Brewer KRW (1963) A model of systematic sampling with unequal probabilities. Australian Journal of Statistics, 5: 5-13.
- Christman MC (2000). A review of quadrat-based sampling of rare, geographically clustered populations. Journal of Agricultural, Biological, and Environmental Statistics, 5: 168–201.
- Dunn R, Harrison A (1993). Two-dimensional systematic sampling of land use. Applied statistics, 42: 585–601.
- Zhang LC (2008). On some common practices of systematic sampling. Journal of Official Statistics, 24: 557– 569.

Text Books

- Cochran WG (1977). Sampling Techniques. John Wiley & Sons, Inc., New York.
- Fuller WA (2009). Sampling statistics. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Särndal CE, Swensson B, Wretman J (1992). Model assisted survey sampling. Springer, New York.

Optimal Designs

Sampling schemes for spatial units can be reasonably treated by introducing a suitable model of spatial dependence within a model-based. However under this assumption the concern consist necessarily in finding the sample configuration that is the best representative of the whole population and leads to define our selection as a combinatorial optimization problem (Benedetti and Palma 1995). If we define a model $\xi : E_{\xi}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$

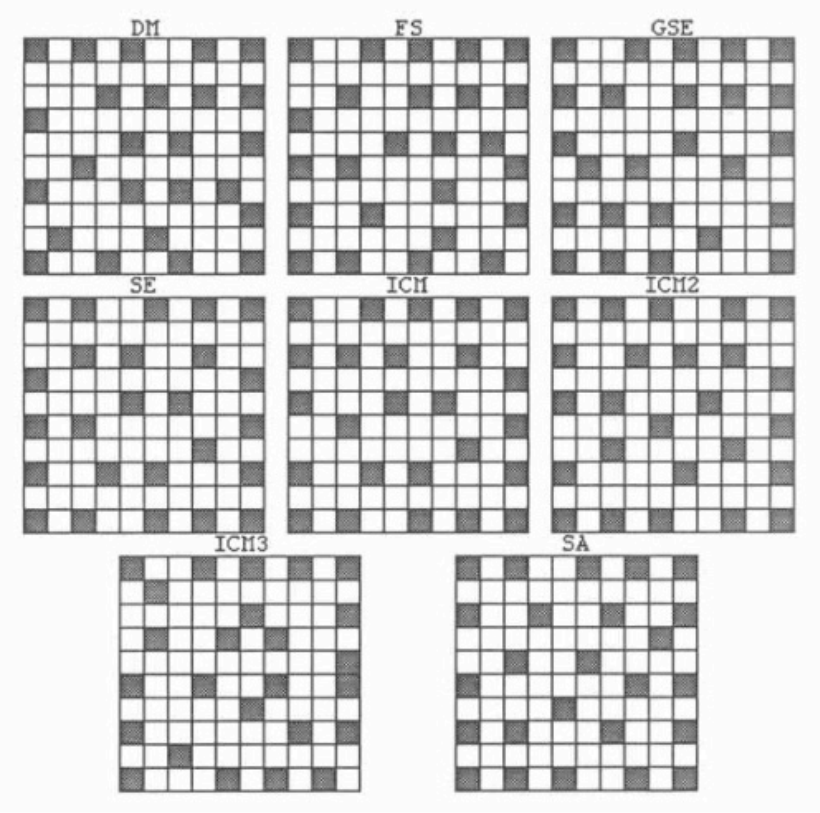
$$Var_{\xi}(\mathbf{Y}) = \mathbf{V}$$

We can choose the sample *s* that minimize the MSE of the model-based prediction :

$$\min_{s}\left\{Var_{\xi}\left(\hat{\theta}-\theta\right)=\boldsymbol{\gamma}_{\bar{s}}^{t}\left(\mathbf{V}_{\bar{s}\bar{s}}+\mathbf{X}_{\bar{s}}\mathbf{A}_{s}^{-1}\mathbf{X}_{\bar{s}}^{-1}\right)\boldsymbol{\gamma}_{\bar{s}}\right\}$$

Usually it is the sample that maximize the distance between units.

Optimal Designs



Optimal Designs: Some References

- Benedetti R, Palma D (1995). Optimal sampling designs for dependent spatial units. Environmetrics, 6: 101-114.
- Delmelle EM (2013). Spatial sampling. In: Fischer MM, Nijkamp P (eds) Handbook of regional science, Springer, Berlin, pp. 1385-1399.
- Ver Hoef JM (2002) Sampling and geostatistics for spatial data. Ecoscience 9:152–161
- Ver Hoef JM (2008), Spatial methods for plot-based sampling of wildlife populations, Environmental and Ecological Statistics 15:3–13
- Wang JF, Stein A, Gao BB, Ge, Y (2012). A review of spatial sampling. Spatial Statistics, 2: 1-14.

Sampling Plans that Exclude Adjacent Units

If there exists some ordering of the units, and contiguous units are anticipated to provide similar data, Hedayat *et al*. (1988b) suggested that more information could be obtained if the sample avoids pairs of contiguous units. It is interesting to note that this feature is considered so important that it was suggested by Hedayat *et al*. (1988b) as a practical solution. In fact, they observe that

"if in any observed sample contiguous (or close to each other in some sense) units occur, they may be collapsed into a single unit with the corresponding response as the average observed response over these units. An estimate of the unknown parameter is then made on the basis of such a reduced sample".

The basic design was suggested by Hedayat *et al*. (1988a) and called balanced sampling design excluding contiguous units (BSEC). It is a fixed size n design where $\pi_{kl} = 0$ if the units k and l are contiguous, and all other π_{kl}s are equal to an appropriate constant.

Sampling Plans that Exclude Adjacent Units

A theoretical comparison of the variance of this design with the classical benchmark represented by SRS shows that, when using the HT estimator for the total, BSEC represents a better strategy if and only if

$$\rho_1 > -\frac{1}{N-1}$$

where ρ_1 is the first-order circular serial correlation coefficient between the units and is given by

$$\rho_1 = \sum_{k \in U} \left(y_k - \mu_y\right)\left(y_{k+1} - \mu_y\right) \Big/ N\sigma^2$$

It is interesting to note that a similar role is played by the *sample autocorrelation coefficient* defined as $\rho_{\tilde{y}} = C_{k \neq l \in U}(\tilde{y}_k, \tilde{y}_l) \big/ s^2_{\tilde{y}}$, where *C* is the covariance and *S* is the variance of the survey variable *y*. Using the decomposition, this can be shown to have the bounds (Knottnerus 2003, p. 89)

$$-\frac{1}{n-1} \leq \rho_{\tilde{y}} \leq 1$$

DUST design

Arbia (1993) was inspired by purely model-based assumptions on the dependence of the stochastic process that generates the data, according to the algorithm types identified by Tillé (2006). Arbia (1993) suggested a draw-by-draw scheme called the dependent areal units sequential technique (DUST). The properties of DUST can be also analyzed within a design-based framework, because it respects the randomization principle.

The main argument for this method was that

"it is intuitively clear that, when we have a clue of the spatial correlation structure underlying the spatial phenomenon to be sampled, it is desirable to exploit this information in the sampling design. In this way we could avoid duplicate information partly contained in areas already sampled and we can economize sampling costs without loosing reliability of the estimates"

DUST design

The DUST algorithm starts by randomly selecting a unit *k*. Then, at every step *t*<*n*, the algorithm updates the selection probabilities of any other unit (*l*) of the population according to the rule

$$\pi_l^{(t)} = \pi_l^{(t-1)} \left(1 - \lambda^{d_{kl}}\right)$$

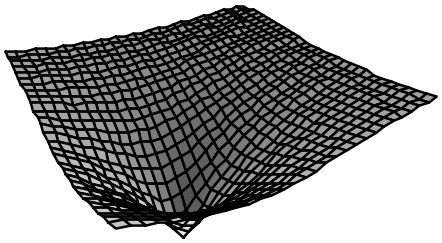
where λ is a tuning parameter used to control the distribution of the sample over the study region, and d_{kl} is a measure of distance between unit *k* and *l*.

This algorithm, or at least the sampling design that it implies, can easily be interpreted and analyzed in a design-based perspective, with particular reference to a detailed empirical assessment of the first and second-order inclusion probabilities because they are theoretically unknown.

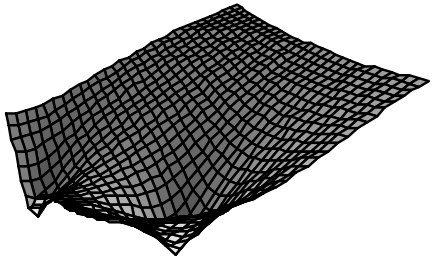
Problem : are the obtained π_k^* equal to the design π_k? **NO**

DUST design

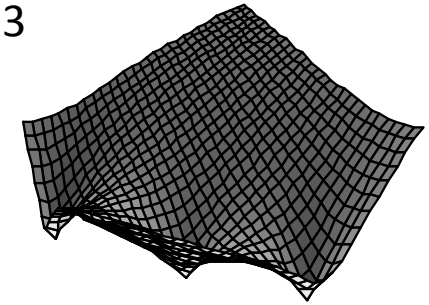
1



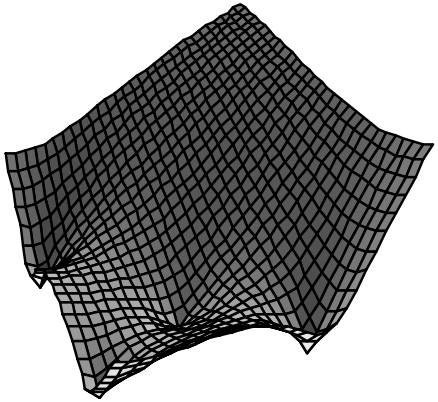
2



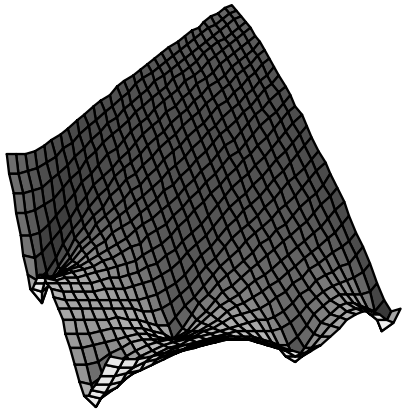
3



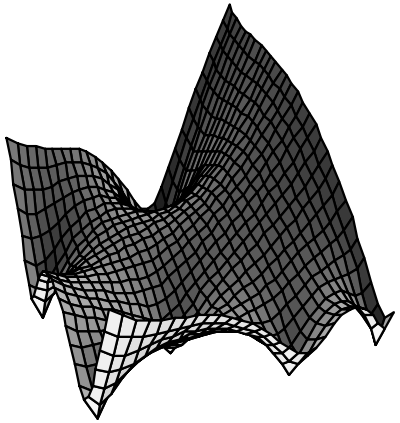
4



5



6

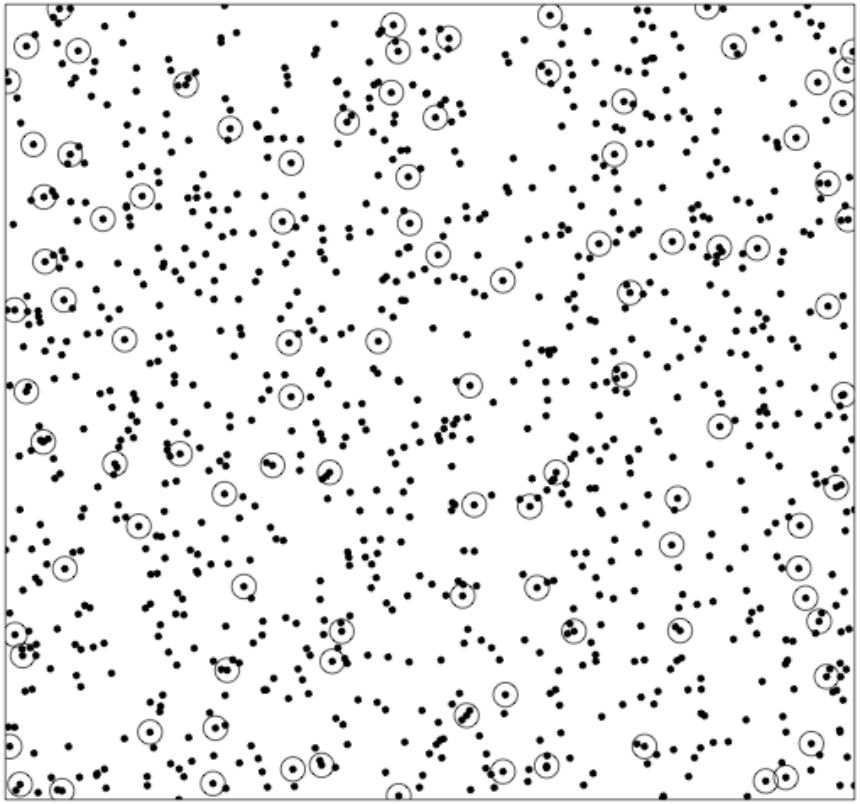


DUST design

```
> DUST <- function (matraux,nsamp,bdis=2,nrepl=1)
+ {
+   selez<-matrix(0,nsamp*nrepl,2)
+   npo<-nrow(matraux)
+   dis<-as.matrix(dist(matraux))
+   dis<-1-exp(-bdis*(dis))
+   for (cc in 1:nrepl)
+   {
+     psel<-rep(1/npo,npo)
+     for (j in 1:nsamp)
+     {
+       selez[(cc-1)*nsamp+j,1]<-cc
+       selez[(cc-1)*nsamp+j,2]<-sample(1:npo,1,prob=psel)
+       psel=psel*dis[selez[(cc-1)*nsamp+j,2],]
+       psel=psel/sum(psel)
+     }
+   }
+   selez
+ }
```

DUST design

```
> n <- 100
> N <- 1000
> set.seed(200694)
> X <- cbind(framepop$xc,framepop$yc)
> DUSTsel <- DUST(X,n,bdis=10)
> nrow(DUSTsel)
[1] 100
>sbi(ds,rep(n/N,N),DUSTsel[,2])
[1] 0.2830303
>par(mar=c(1,1,1,1), xaxs="i",
+ yaxs="i")
>plot(framepop$xc,framepop$yc,
+ axes=F, cex=0.5,pch=19,
+       xlim=c(0,1),ylim=c(0,1))
>points(framepop$xc[DUSTsel[,2]],
+ framepop$yc[DUSTsel[,2]],
+         pch=1, cex=2)
>box()
```

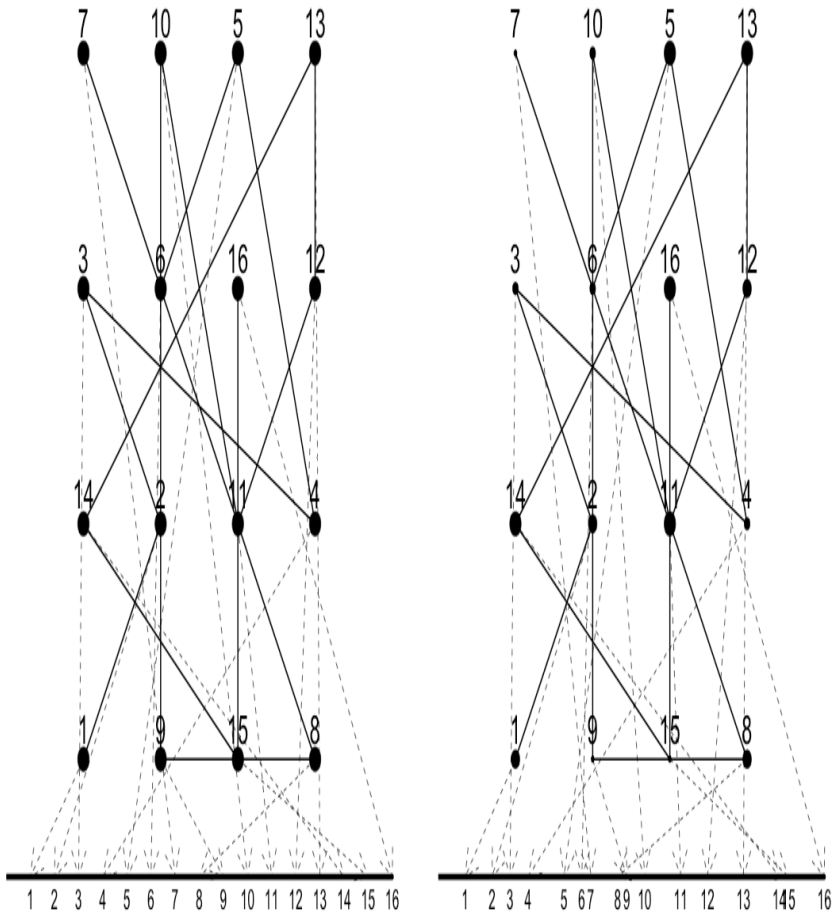


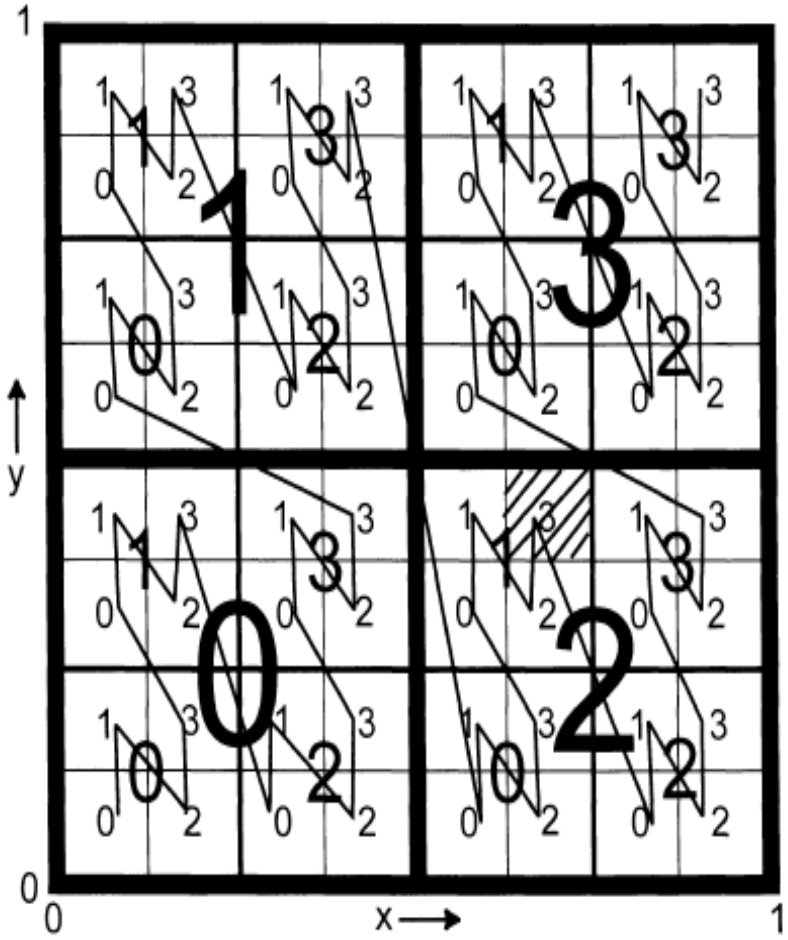
Generalized Random Tessellation Sampling

- Idea: map two-dimensional spatial population into one-dimensional population while preserving some spatial order.
- Sampling mechanism:

1. the sampling units are sorted according to a recursive, hierarchical randomization process, which tries to preserve the spatial relationship of the units;
2. the sampling units are ordered by means of a function f, which maps the two-dimensional space of the population into one-dimensional space, then defining an ordered spatial address;
3. the one-dimensional space of units (i.e. a line) obtained by the previous steps is then divided into a number of equal-length segments. This division depends on the request on the requested sample size, since one unit is selected randomly from each segment (hence, the line is divided in n segments).

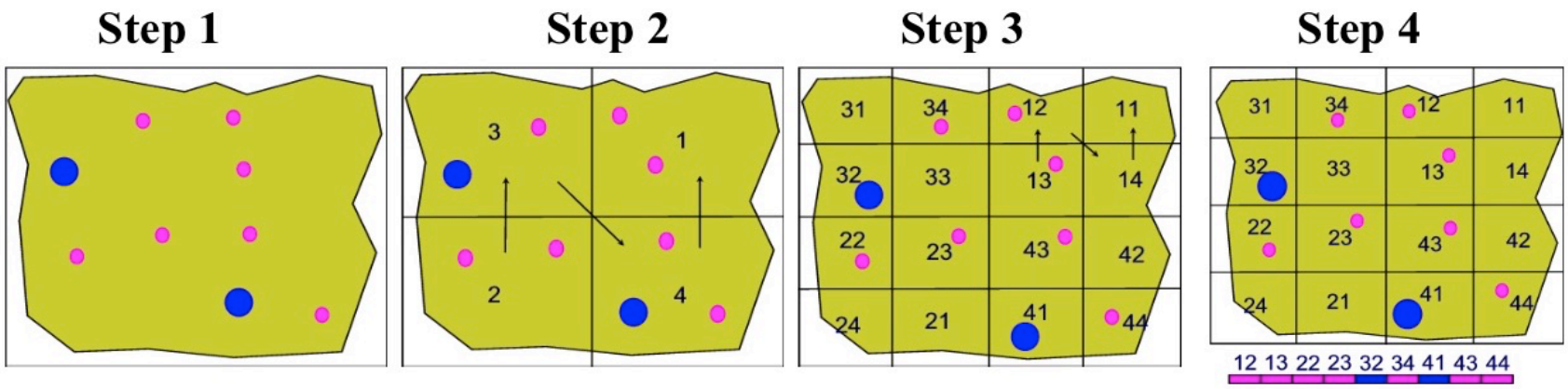
Generalized Random Tessellation Sampling





Generalized Random Tessellation Sampling

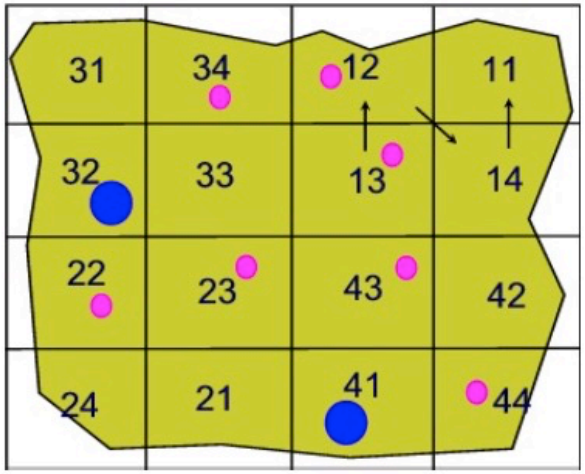
Example, Equal probabilities π_i

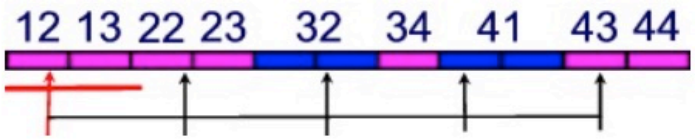


- Step 1: Frame: Large lakes: blue; Small lakes: pink; Randomly place grid over the region
- Step 2: Sub-divide region and randomly assign numbers to sub-regions
- Step 3: Sub-divide sub-regions; randomly assign numbers independently to each new sub-region; create hierarchical address. Continue sub-dividing until only one lake per cell.
- Step 4: Identify each lake with cell address; assign each lake length 1; place lakes on line in numerical cell address order.

Generalized Random Tessellation Sampling

Example, Unequal probabilities π_i





- Assume want large lakes to be twice as likely to be selected as small lakes
- Instead of giving all lakes same unit length, give large lakes twice unit length of small lakes
- To select 5 sites divide line length by 5 (11/5 units); randomly select a starting point within first interval; select 4 additional sites at intervals of 11/5 units
- Same process is used for points and areas (using random points in area)

Generalized Random Tessellation Sampling

Advantages

- Spatial balance.
- It can be used for sampling point, linear features and not contiguous phenomena.
- Possibility to sample with unequal probability.
- Practical and can be applied even in problematic situations like poor frame information and irregular space pattern.

Generalized Random Tessellation Sampling

Disadvantages

- Only applicable over units with a pair of coordinates (.
- Possibility to lose some spatial relationship during the use of .

Reference

- Reference: Steven and Olsen 2004.

R package

- `spsurvey` (Kincaid and Olsen 2016).

Some references

Some references about GRTS.

- Barabesi L, Franceschi S (2011). Sampling properties of spatial total estimators under Tessellation Stratified Designs. Environmetrics, 22: 271–278.
- Stevens DL Jr (1997). Variable density grid-based sampling designs for continuous spatial population. Environmetrics, 8: 167-195.
- Stevens DL Jr, Olsen AR (1999). Spatially restricted surveys over time for aquatic resources. Journal of Agricultural, Biological, and Environmental Statistics, 4: 415-428.
- Stevens DL Jr, Olsen AR (2003). Variance estimation for spatially balanced samples of environmental resources. Environmetrics, 14: 593–610.
- Stevens DL Jr, Olsen AR (2004). Spatially balanced sampling of natural resources. Journal of the American Statistical Association, 99: 262–278

Spatially Correlated Poisson Sampling

It is difficult to modify the second-order inclusion probabilities while preserving fixed π_k. Bondesson and Grafström (2011) extended Sampford's method to address this issue by defining a procedure that appears to be only in one dimension, but that actually explores two dimensions of spatial units. On the basis of this result, Grafström (2012) proposed a method called spatially correlated Poisson sampling (SCPS).

The sequential nature of the list means that we first decide the sampling outcome for the first unit of a (possibly previously randomly sorted) list, then for the second unit, and so on until *n* units have been selected.

Spatially Correlated Poisson Sampling

If Unit 1 is included with probability $\pi_1^{(0)} = \pi_1$, we set I_1 =1, otherwise I_1=0. After each step, the inclusion probabilities for the remaining units in the list are updated according to a specific rule. We start with $\pi_k^{(0)} = \pi_k$, for k≥1.

At step t, the values of I_1, I_2,…, I_{t-1} are known, and we select unit t with probability $\pi_t^{(t-1)}$. We update the generic unit k≥t+1 according to $\pi_k^{(t)} = \pi_k^{(t-1)} - \left(I_t - \pi_t^{(t-1)}\right) w_{k-t}^{(t)}$, where $w_{k-t}^{(t)}$ are weights that depend on I_1, I_2,…, I_{t-1} but not on I_t .

To preserve the fixed first-order inclusion probabilities, the weight that we can give to a unit is limited by

$$-\min\left(\frac{1-\pi_k^{(t-1)}}{1-\pi_t^{(t-1)}}, \frac{\pi_k^{(t-1)}}{\pi_t^{(t-1)}}\right) \le w_{k-t}^{(k)} \le \min\left(\frac{\pi_k^{(t-1)}}{1-\pi_t^{(t-1)}}, \frac{1-\pi_k^{(t-1)}}{\pi_t^{(t-1)}}\right)$$

Spatially Correlated Poisson Sampling

One very interesting property is that the maximal weights strategy locally balances the sample size, like a form of loose spatial maximal stratification without fixed and accurate borders. This local property can be better appreciated by showing that, if the study region is partitioned into two strata, *A* and *B*, so that units within the same stratum are always closer than units belonging to different strata,

$$\sum_{l \in A} \pi_l \text{ and } \sum_{l \in B} \pi_l$$

are asymptotically equal to n_A and n_B, respectively. Then the maximal weights method will approximately select units from *A* and from *B*. Therefore, it locally satisfies the theoretical basis of the *spatial balance* index.

Spatially Correlated Poisson Sampling

Advantages

- Unequal Probability Sampling

Reference

- Reference: Grafstrom 2012.

R package

- `BalancedSampling (Grafstrom and Lisic 2016`

Local Pivotal Method

Using a similar technique, Grafström *et al*. (2012) derived two alternative procedures for selecting samples with fixed π_k and correlated inclusion probabilities, as an extension of the pivotal method for selecting *πps* samples (Deville and Tillé 1998). They are essentially based on an updating rule for the probabilities π_k and π_l. At each step, the rules state that the sum of the updated probabilities is as locally constant as possible, and that they differ from each other in the way that the two nearby units *k* and *l* are chosen. These two methods are referred to as the local pivotal method 1 (LPM1), which the authors suggest is better *spatially balanced*, and the local pivotal method 2 (LPM2), which is simpler and faster.

A sample is obtained in *N* steps. At each step, the inclusion probabilities for two units are updated, and the sampling outcome is decided for at least one the units.

Local Pivotal Method

Using a similar technique, Grafström *et al*. (2012) derived two alternative procedures for selecting samples with fixed π_k and correlated inclusion probabilities, as an extension of the pivotal method for selecting *πps* samples (Deville and Tillé 1998). They are essentially based on an updating rule for the probabilities π_k and π_l. At each step, the rules state that the sum of the updated probabilities is as locally constant as possible, and that they differ from each other in the way that the two nearby units *k* and *l* are chosen. These two methods are referred to as the local pivotal method 1 (LPM1), which the authors suggest is better *spatially balanced*, and the local pivotal method 2 (LPM2), which is simpler and faster.

A sample is obtained in *N* steps. At each step, the inclusion probabilities for two units are updated, and the sampling outcome is decided for at least one the units.

Local Pivotal Method

Deville and Tillé (1998) suggested randomly choosing a pair of units at each step to maximize the entropy of the selected units. Grafström *et al*. (2012) introduced LPMs that update the inclusion probabilities according to the same updating rule of Deville and Tillé (1998) but for two nearby units, improving the *spatial balance*.

LPM1 randomly chooses the first unit *k,* and then the closer unit *l* (if two or more units are the same distance from *k*, the method randomly chooses between them). If *k* is the nearest neighbor of *l*, then the inclusion probabilities are updated as follows.

Local Pivotal Method

If $\pi_k + \pi_l < 1$, then

$$\left(\pi_k^*, \pi_l^*\right) = \begin{cases} \left(0, \pi_k + \pi_l\right) & \text{with probability } \dfrac{\pi_l}{\pi_k + \pi_l} \\ \left(\pi_k + \pi_l, 0\right) & \text{with probability } \dfrac{\pi_k}{\pi_k + \pi_l} \end{cases}$$

or, if $\pi_k + \pi_l \geq 1$, then

$$\left(\pi_k^*, \pi_l^*\right) = \begin{cases} \left(1, \pi_k + \pi_l - 1\right) & \text{with probability } \dfrac{1 - \pi_l}{2 - \pi_k - \pi_l} \\ \left(\pi_k + \pi_l - 1, 1\right) & \text{with probability } \dfrac{1 - \pi_k}{2 - \pi_k - \pi_l} \end{cases}$$

The expected number of computations for this algorithm is at worst proportional to N^3, and at best proportional to N^2.

Local Pivotal Method

Advantages

- Unequal Probability Sampling

Reference

- Reference: Grafstrom 2012.

R package

- `BalancedSampling (Grafstrom and Lisic 2016)`

The Doubly Balanced Sampling

Grafström and Tillé (2013) combined their techniques (i.e., the LPM and the CUBE), proposing a new method that aims to achieve a double property of balancing. This new method ensures that the sample is well-spread avoiding the selection of selecting neighboring units (i.e., as the LPM). Besides, the method also allows satisfying balancing equations on auxiliary variables that are available on all the sampling spatial units (i.e., as the CUBE). This method is denoted as doubly balanced spatial sampling (DBSS).

The Doubly Balanced Sampling

Advantages

- Unequal Probability Sampling, Balanced on a set of covariates **X**

Reference

- Reference: Grafstrom and Tillé 2013.

R package

- `BalancedSampling (Grafstrom and Lisic 2016)`

Halton Numbers, Continuous Populations

Halton sequences are sequences used to generate points in space for numerical methods such as Monte Carlo simulations. Although these sequences are deterministic, they are of low discrepancy, that is, appear to be random for many purposes (pseudo random). They were first introduced in 1960 and are an example of a quasi-random number sequence. They generalise the one-dimensional van der Corput sequences.

Halton Numbers

Advantages

- Very simple, continuous populations

Reference

- Reference: Robertson et al. 2013.

R packages

- `randtoolbox(Chalabi, Dutang, Savicky and Wuertz, 2016)`
- `SDraw (McDonald, T. L. 2016)`

Some references

- Arbia G (1993). The use of GIS in spatial statistical surveys. International Statistical Review, 61: 339–359.
- Benedetti, R, Piersimoni, F and Postiglione, P. (2015). Sampling spatial units for agricultural surveys. Advances in Spatial Science Series. Springer.
- Benedetti, R, Piersimoni, F and Postiglione, P. (2017). Spatially Balanced Sampling: A Review and A Reappraisal. International Statistical Review, 85, 3, 439–454.
- Bondesson L, Grafström A (2011). An extension of Sampford's method for unequal probability sampling. Scandinavian Journal of Statistics, 38: 377-392.
- Bondesson, L. Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. Scand. J. Stat., 35, 466–483.
- Deville, J.-C., Tillé, Y. (2004) Efficient balanced sampling: The cube method, *Biometrika*, 91, 893-912
- Grafström A (2012). Spatially correlated Poisson sampling. Journal of Statistical Planning and Inference, 142: 139–147.
- Grafström A, Lundström NLP, Schelin L (2012). Spatially balanced sampling through the pivotal method. Biometrics, 68: 514-520.
- Grafström A, Tillé Y (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics, 24: 120-131.
- Grafström A, Lundström NLP(2013). Why well spread probability samples are balanced. Open Journal of Statistics, 3: 36–41.
- Grafström A, Schelin L. (2014). How to select representative samples. Scandinavian Journal of Statistics.
- Robertson, B. L., Brown, J. A., McDonald, T. L. and Jaksons, P. (2013). BAS: Balanced acceptance sampling of natural resources, Biometrics 69, 776-784.
- **Variance Estimation**
- Benedetti, R. Espa, G. and Taufer, E (2017) "Model-based variance estimation in non-measurable spatial designs", *Journal of Statistical Planning and Inference* 181, 52-61
- Grafström, A., and Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41, 2, 277-290.
- Stevens DL Jr, Olsen AR (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14: 593–610.

See you in the afternoon for the 3rd and final episode of the series…