# An introduction to 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

# 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 $\xi$. 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 $\xi$, 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 $\lambda_i$. 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.

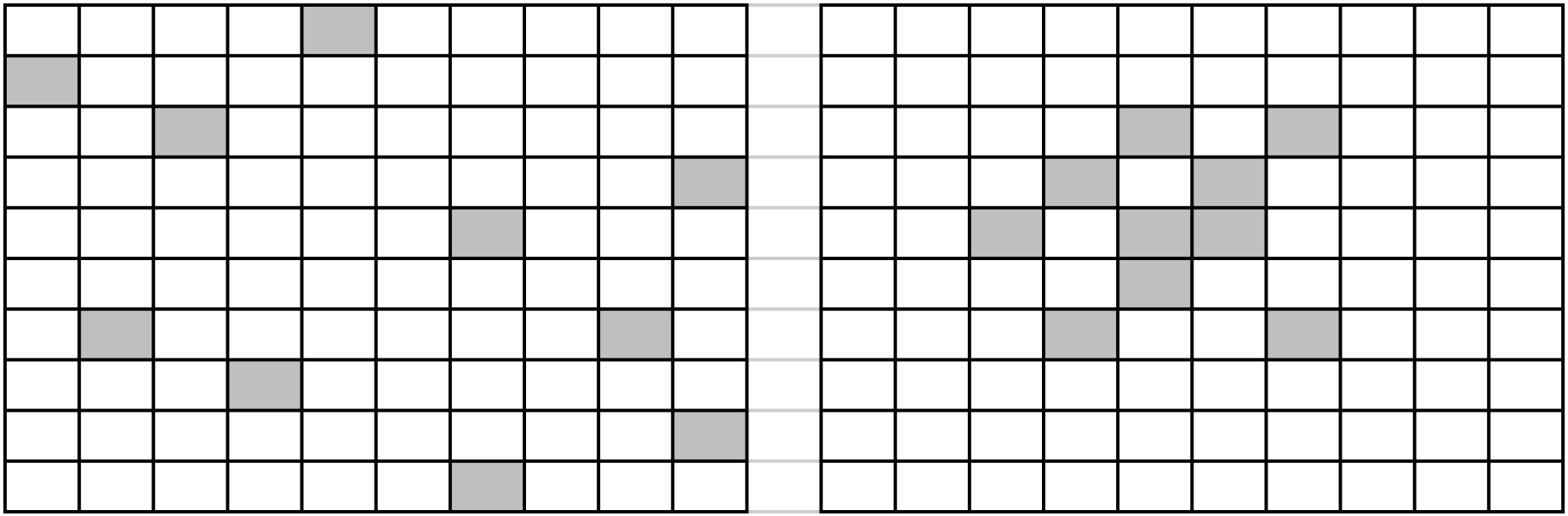
# Spatially Balanced Sampling

*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:

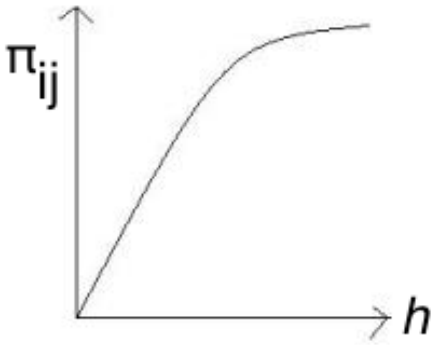
- Yates-Grundy-Sen formulation of the HT variance.
- Anticipated Variance.
- Lemma decomposition.

# 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.

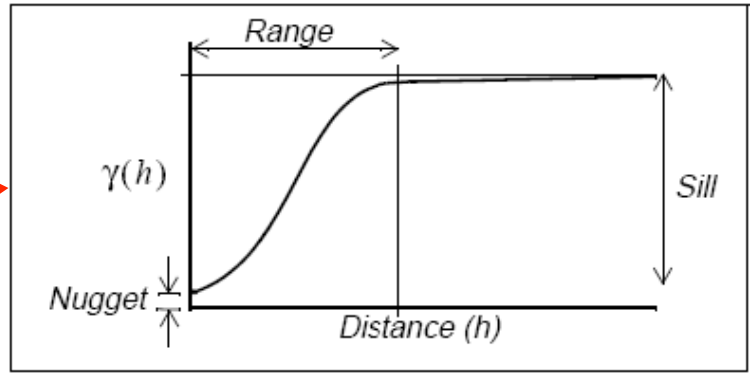
- Yates-Grundy-Sen formulation of the HT variance:

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




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 $\xi$ 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, $\beta$ is a vector of coefficient regression, $\rho_{ij}$ 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 $\rho_{ij}$ 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 $y_k/\pi_k$ 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

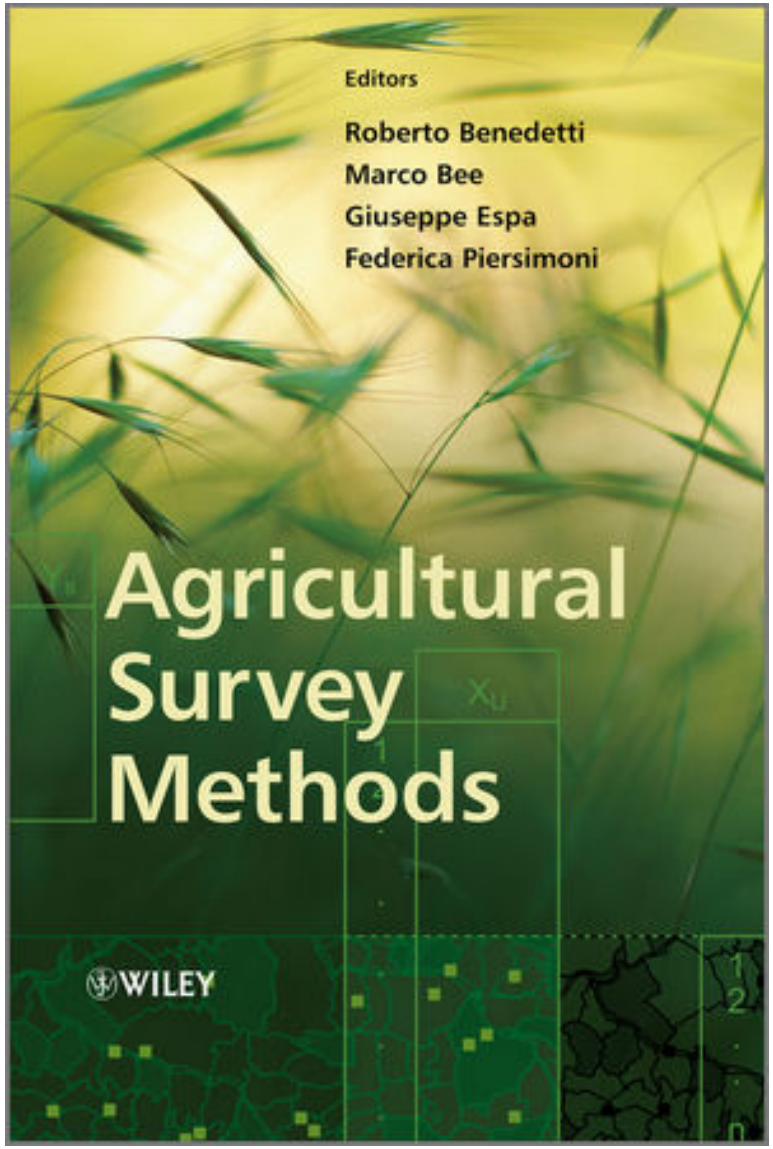
The Voronoi polygon for unit *k* of a generic sample *s* includes all the population units closer to *k* than to any other unit in the sample. Let

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

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 $v_k$s 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.

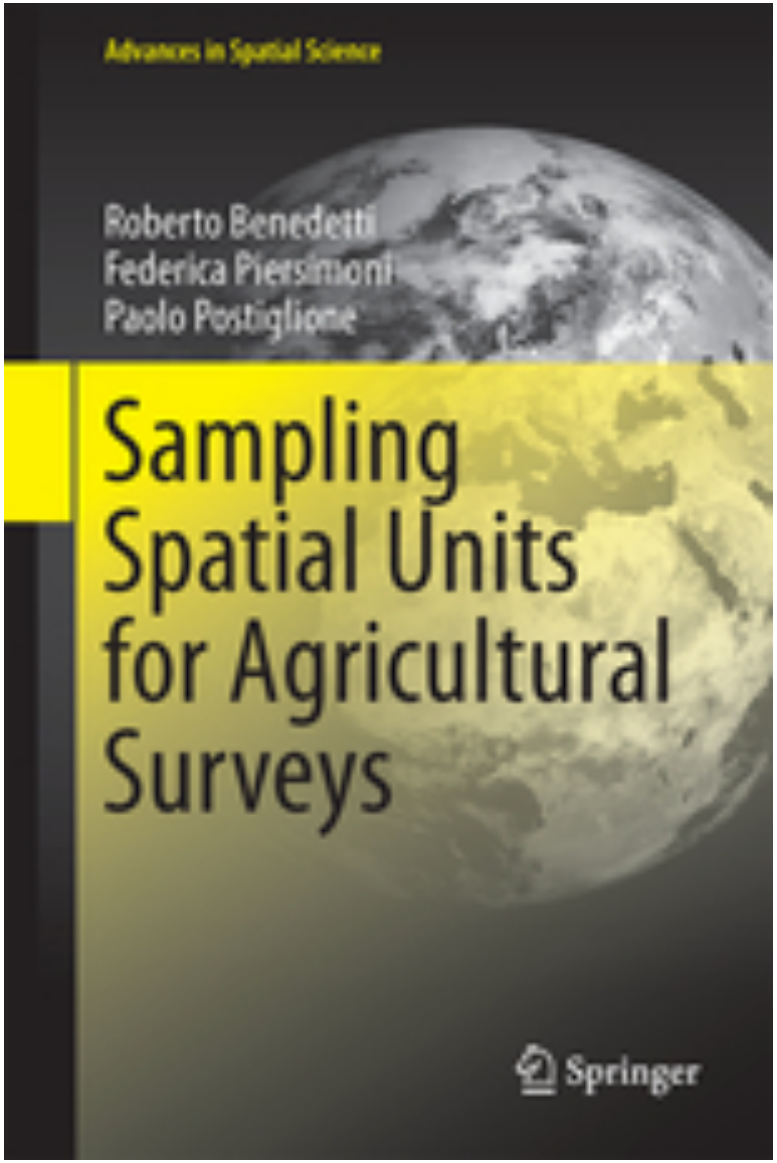
$$V\left(v_k\right) = \frac{\sum_{k \in s} \left(v_k - 1\right)^2}{n}$$

# Some references



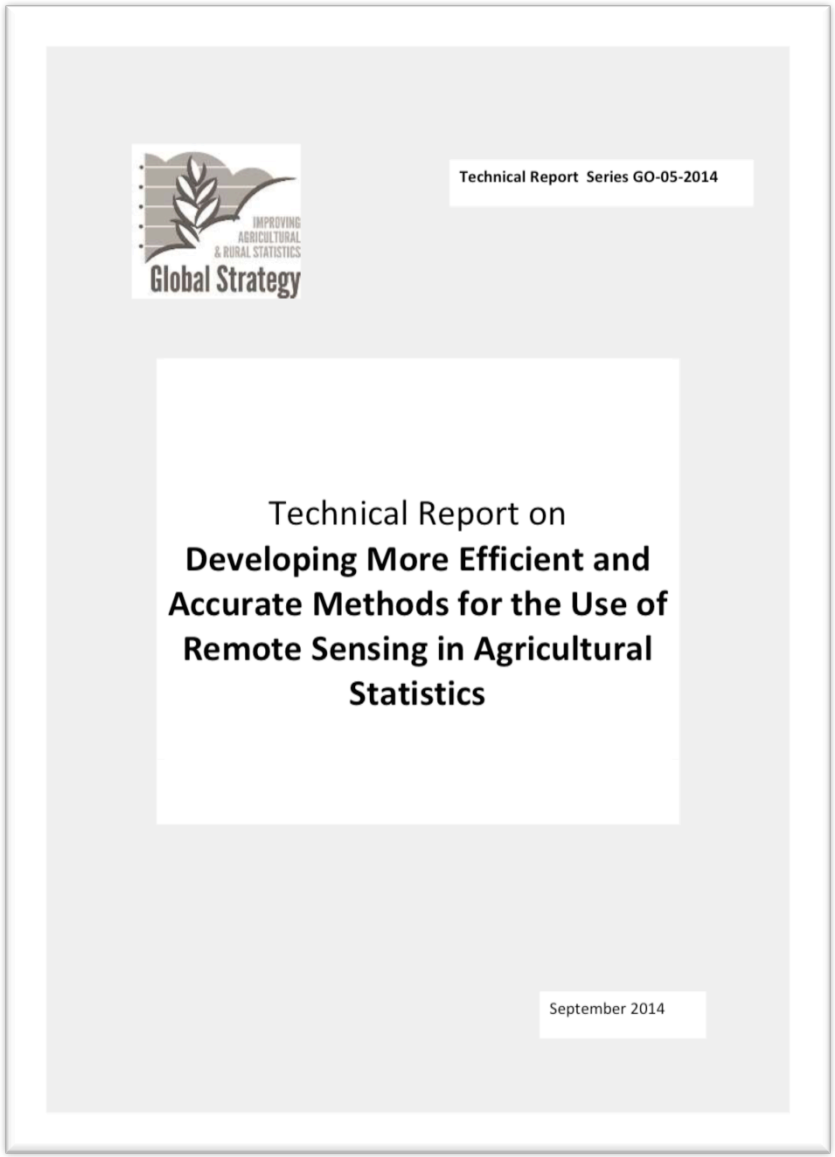

**Contents**

# Some references




**Contents**

# Some references




**Contents**

# Spatially Balanced Sampling

**Thank you for your attention!**