Spatial sampling modified by model use

Tomasz Bąk

University of Economics in Katowice, Department of Statistics, Econometrics and Mathematics

Warsaw, July 2018

Result of grant supported by National Science Centre, Poland, no. 2016/21/B/HS4/00666

Introduction to the method

Source of inspiration

- Spatially balanced sample is a sample in which units are well spread throughout the study area. Many authors considered sampling methods which led to spatially balanced sample:
  - Sampling design based on the neighborhood matrix [12]
  - Generalized Random-Tessellation Stratified method (GRTS) [9]
  - Spatially Correlated Poisson Sampling method (SCPS) [4]
  - Local Pivotal method [5]
  - Doubly Balanced Spatial Sampling method (DBSS) [6]
- Sample balanced due to auxiliary variable
  - Cube method [2]
  - Doubly Balanced Spatial Sampling method (DBSS) [6]
- Adaptive sampling designs [11], which are divided into two stages . In the first stage a 'classical' sampling design, i.e. simple random sampling is used. In second stage, some additional elements which fulfilled a certain condition are added to the sample.

Model modification

Adaptive sampling methods give the researcher a chance to control the sample selection. It is possible because after each sampled element the information about this element is used to redefine the probabilities of inclusion. Sampling scheme learns from already sampled elements. The question is how to modify (to learn) the sampling design efficiently during sampling? Quite intuitive solution seems to be to use the statistical models as a learning tool. After each sampled element model is constructed and then it is used to modify sampling design. Here emphasis is put on spatial sampling designs, so spatial modeling (kriging in particular) will be considered as a learning tool.

Notation

- Y is characteristic under study.
- X is additional variable which values are observed during sampling and used on adaptive part of sampling.
- $\bar{x}$ is global average value of X characteristic known before sampling.
- N-population size, n-sample size.

Stages of sampling

Two stages of spatial sampling modified by model use could be distinguished, as well as in adaptive sampling. At the first stage an initial sample $s_0 = \{i_1, \dots, i_{n_0}\}$ size of n_0 is selected. Each element of the initial sample is selected with the probabilities $p_1, \dots, p_{n_0}$, defined as

$$p_j = [p_{j,1}, \dots, p_{j,N}], \quad j = 1, \dots, n_0. \tag{1}$$

What after initial sampling?

Let us assume that the initial sample size of n_0 was selected. Furthermore, n_0 realizations of the X process are obtained. Denote them by $x_{i_1}, \ldots, x_{i_{n_0}}$. It means that for initial sample elements both values and locations are known. Therefore it is possible to use statistical model to predict values of X process for the whole population. Let us assume that $\hat{X}_{n_0}(i)$, $i \in \{1, \ldots, N\}$ is a predictor of X process, constructed on the base of all elements from the initial sample. This predictor is used to change sampling probability of $(n_0 + 1)$-th element. In the second stage, new predictor $\hat{X}_k(i), i \in \{1, \ldots, N\}$, $k > n_0$ is constructed after each newly sampled element.

Sampling elements in the second stage - part 1

Let us consider sampling of k-th element, which is conducted in the second stage ($k > n_0$). It can be made in one of two ways: depending on the predictor $\hat{X}_{k-1}(i)$ or in the same way as selection of the elements $1, \ldots, n_0$. Next, let us introduce the probability d_{k-1}. Then, the condition which defines which way of sampling will be chosen to select k-th element is as follows:

$$\exists_{i \in \{\{1,\ldots,N\} \setminus \{i_1,\ldots,i_{k-1}\}\}} \|\hat{x}_i - \bar{x}\| \leq c, \tag{2}$$

where $\hat{x}_i$, $i \in \{1, \ldots, N\}$ are predictions of X characteristic from the model $\hat{X}_{k-1}(i), i \in \{1, \ldots, N\}$ and c is the coefficient which defines how close the predicted value must be to the global average value $\bar{x}$. The reason behind the condition (2) is a stronger aggregation of X characteristic values around $\bar{x}$ than it could be expected in non-adaptive sampling methods. Depending on the satisfiability of the condition (2), the k-th element is selected in a one of two ways:

Sampling elements in the second stage - part 2

- Condition (2) is false. Then sampling method analogous to the one that has been used in initial sampling stage is used. The k-th element is sampled with probabilities p_k, defined as in (1).
- Condition (2) is true. Then two situations are possible. With probability $1 - d_{k-1}$, the k-th element is sampled in the same way as when condition (2) is false. Otherwise, with probability d_{k-1}, the k-th element is sampled among the elements of the set

$$H_{k-1} = \{i \in \{1, \dots, N\} / \{i_1, \dots, \quad i_{k-1}\} : \quad \|\hat{x}_i - \bar{x}\| \leq c\}. \tag{3}$$

Then vector $p'_k = \left[p'_{k,1}, \dots, p'_{k,N}\right]$ is defined as follows:

$$p'_{k,i} = \begin{cases} \frac{p_{k,i}}{\sum_{i \in H_{k-1}} p_{k,i}}, \text{ when } i \in H_{k-1}, \\ 0, \text{ when } i \notin H_{k-1}. \end{cases} \tag{4}$$

In other words, probabilities in p'_k vector are proportional to probabilities in p_k for the elements of the set H_{k-1} and equal to 0 for other elements of the population. The probabilities p'_k are used to sample the k-th element.

About initial sampling plan

As we can see, sampling plan which was used at the initial stage is used at the second stage of sampling too. Moreover, probabilities p'_k, $k = n_0 + 1, \ldots, N$ are based on the probabilities p_k, $k = n_0 + 1, \ldots, N$. Therefore, the choice of initial sampling has great impact on the spatial sampling modified by model use.

Series d_k

The basic feature of the d_k, $k = n_0, \ldots, n - 1$ is that the higher the value of d_k is, the greater 'adaptability' (ability to learn on already sampled elements) sampling scheme has. On the other hand, the precision of the 'adaptability' is based on the precision of the model, which is mainly conditioned by the number of already sampled elements. It is well known that precision of the spatial model increases with increasing sample size [1]. Therefore, a sequence of probabilities $d_k, k = n_0, \ldots, n - 1$ should be increasing. In principle, the same assumption about the sequence $d_k, k = n_0, \ldots, n - 1$ was made by Thompson in the definition of adaptive web sampling [8].

Estimation method

In spatial sampling modified by model use, first-order probabilities of inclusion are unequal. They depend on the results of modeling and can not be defined explicite. Fattorini [7] proposed the method of using Horvitz-Thompson estimator in case when an explicit derivation of first-order probabilities of inclusion is prohibitive. This approach was later developed in several papers [10, 3].

Let us assume that M samples from the population $\{1, \ldots, N\}$ are selected independently and by repeating the same rules. An invariably positive estimator of first-order probabilities of inclusion $\pi_j, \ j = 1, \ldots, N$ is

$$\hat{\pi}_j = \frac{m_j + 1}{M + 1}, \quad j = 1, \ldots, N, \tag{5}$$

where m_j is the total number of samples in which the j-th element was drawn. Since $M \to \infty$, then asymptotically unbiased modification of the primary Horvitz-Thompson estimator is

$$\hat{T} = \sum_{j=1}^{n} \frac{x_j}{\hat{\pi}_j}. \tag{6}$$

Example

Employees research

Let us consider a spatial research of the average number of employees. Elements under study are districts of a region. An additional characteristic observed during research is the number of inhabitants. In fact, the additional characterstic is auxiliary variable in strict sense. We assume that for all elements number of inhabintants is known before sampling. It is stronger assumption then the one required by spatial sampling modified by model use. Nevertheless, that assumption is desirable from the efficiency testing point of view. It allows to compare the efficiency of spatial sampling modified by model use with the efficiency of some existing methods which require auxiliary variable.

Population under study

Survey population is consisted of 300 districts. For each number of inhabitants is known (X variable) and number of employees is characteristics under study (Y variable). For both characteristics two aggregations of lower values and two aggregations of higher values can be observed [13].

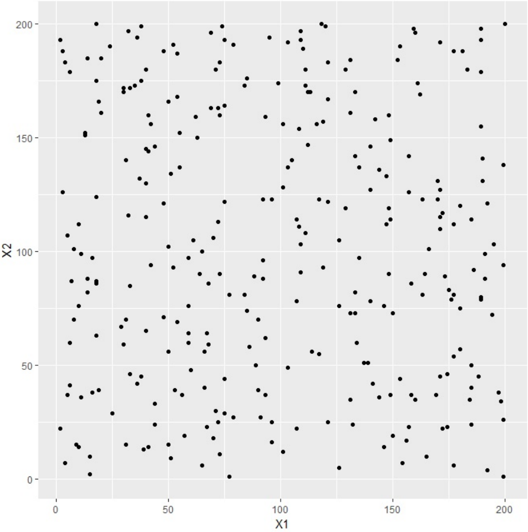
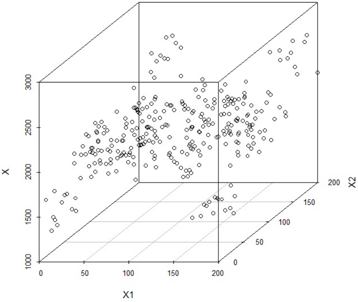
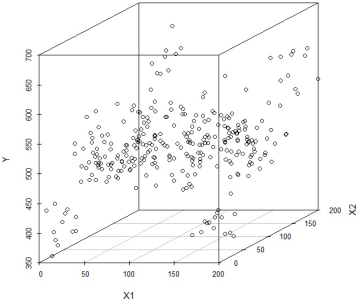


Figure: *Population distribution in the space.*



(a) *Number of inhabitants distribution.*



(b) *Number of employees distribution.*

Initial sampling methods

Five sampling methods were considered as initial sampling method. Three of them were: Spatially Correlated Poisson Sampling method (SCPS) and both Local Pivotal methods (LPM1 and LPM2). Each of them exploits the information about spatial location of the population elements. The fourth of the methods was Cube method (CM) which is example of balanced sampling and fifth was Unequal Probability Sampling (UPS). The last two methods use the information from auxiliary variables. First-order inclusion probabilities were defined proportionally to auxiliary variable for each initial sampling methods.

Modeling method

As a spatial data modeling method ordinary kriging was chosen. Already sampled elements were used to variogram estimation. Spherical, exponential, gaussian and Matern family models were considered as potential shape of the variogram. Finally sampling algorithm picked the variogram model that has the smallest residual sum of squares and used it in kriging modeling.

Sampling assumptions

Sample size of $n = 100$, and initial sample size of $n_0 = 50$ were considered. The sequence $\{d_k\}, k = 50, \ldots, n-1$ was defined as

$$d_k = \frac{k}{100}, k = 50, \ldots, 99. \tag{7}$$

The value of c coefficient was equal to 5, which was about 25% of standard deviation of the X. For each of five initial sampling methods, sampling of 100 elements was repeated 10 000 times to achieve, through Monte Carlo method, estimation of the first-order inclusion probabilities. Two different approaches to first-order probabilities of inclusion definition were considered. First based on on Monte Carlo estimation, second in which the first-order probabilities of inclusion were proportional to the auxiliary variable. For both approaches and each type of initial sampling method 1 000 samples were selected.

Estimation method

In the next step additional 1.000 samples were selected for each approach. For each sample the value of modified Horvitz-Thompson estimator for was calculated.
Results of spatial sampling modified by model use were compared to primary forms of sampling methods (the one which were used as initial sampling methods). Therefore 1 000 samples, each consisted of 100 elements, were selected using all five sampling methods used for initial sampling. Horvitz-Thompson estimator was based on first-order probabilities of inclusion proportional to auxiliary variable.

Sampling results

rRMSE was calculated for each of the approaches.

Table: *rRMSE for different sampling methods, c=5.*

	Spatial sampling modified by model usage - Probabilities of inclusion based on Monte Carlo	Spatial sampling modified by model usage - Probabilities of inclusion proportional to X variable	Spatial sampling method in primary form
CM	0,467%	0,298%	0,371%
LPM1	0,489%	0,306%	0,279%
LPM2	0,464%	0,275%	0,280%
SCPS	0,478%	0,276%	0,324%
UPS	0,494%	0,296%	0,310%

rRMSE for estimators based on Monte Carlo first-order probabilities of inclusion was significantly higher than for the other two approaches. Spatial sampling modified by model use with first-order probabilities of inclusion proportional to auxiliary variable delivered lower rRMSE than primary form for all methods except LPM1.

Summary of the simulation results.

The results shows that for Cube method, Spatially Correlated Poisson Sampling method and Unequal Probability Sampling model modification increases sampling efficiency (in terms of rRMSE reduction) and it is independent from the c value. Modification of UPS method delivered quite stable rRMSE values for different c values. Model modification of SCPS and CM methods had generally higher rRMSE values for higher c values. LPM1 modified by model use achieved better rRMSE value than primary LPM1 only for $c = 10$ and the difference was negligible. As a rule, model modification was unefficient for this sampling method. In case of LPM2, model modification delivered better efficiency for low values of c parameter. However, the gain was smaller than the ones obtained on CM, SCPS and UPS model modifications.

Appendix

Table: *rRMSE for different sampling methods, c=10.*

Method	Spatial sampling modified by model usage - Probabilities of inclusion based on Monte Carlo	Spatial sampling modified by model usage - Probabilities of inclusion proportional to X variable	Spatial sampling method in primary form
CM	0,541%	0,299%	0,371%
LPM1	0,512%	0,276%	0,279%
LPM2	0,547%	0,267%	0,280%
SCPS	0,559%	0,285%	0,324%
UPS	0,491%	0,296%	0,310%

Table: *rRMSE for different sampling methods, c=15.*

Method	Spatial sampling modified by model usage - Probabilities of inclusion based on Monte Carlo	Spatial sampling modified by model usage - Probabilities of inclusion proportional to X variable	Spatial sampling method in primary form
CM	0,507%	0,302%	0,371%
LPM1	0,522%	0,342%	0,279%
LPM2	0,494%	0,270%	0,280%
SCPS	0,483%	0,273%	0,324%
UPS	0,477%	0,295%	0,310%

Table: *rRMSE for different sampling methods, c=20.*

Method	Spatial sampling modified by model usage - Probabilities of inclusion based on Monte Carlo	Spatial sampling modified by model usage - Probabilities of inclusion proportional to X variable	Spatial sampling method in primary form
CM	0,485%	0,336%	0,371%
LPM1	0,549%	0,286%	0,279%
LPM2	0,538%	0,322%	0,280%
SCPS	0,497%	0,316%	0,324%
UPS	0,465%	0,298%	0,310%

Table: *rRMSE for different sampling methods, c=25.*

Method	Spatial sampling modified by model usage - Probabilities of inclusion based on Monte Carlo	Spatial sampling modified by model usage - Probabilities of inclusion proportional to X variable	Spatial sampling method in primary form
CM	0,452%	0,331%	0,371%
LPM1	0,507%	0,326%	0,279%
LPM2	0,408%	0,304%	0,280%
SCPS	0,472%	0,309%	0,324%
UPS	0,441%	0,293%	0,310%

N.A.C. Cressie.
Statistics for Spatial Data.
Wiley, 1993.

J.-C. Deville and Y. Tillé.
Efficient balanced sampling: The cube method.
Biometrika, 91(4), pages 893–912, 2004.

W. Gamrot.
Estimators for the Horvitz-Thompson statistic based on some posterior distributions.
Mathematical Population Studies, 21(1):12–29, 2014.

A. Grafström.
Spatially correlated poisson sampling.
Journal of Statistical Planning and Inference, 142, pages 139–147, 2012.

A. Grafström, P. Lundström, and L. Schelin.
Spatially balanced sampling through the pivotal method.
Biometrics, 68, pages 514–520, 2012.

A. Grafström and Y. Tillé.
Doubly spatial sampling with spreading and restitution of auxiliary totals.
Environmetrics, 24, pages 120–131, 2013.

Fattorini L.
Applying the horvitz-thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities.
Biometrika 93, pages 269–278, 2006.

Thompson S.K.
Adaptive web sampling.
Biometrics, Vol.62, No.4, pages 1224–1234, 2006.

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

M.E. Thompson, , and C. Wu.
Simulation-based randomized systematic PPS sampling under substitution of units.
Survey Methodology 34, pages 3–10, 2008.

S.K. Thompson and G. Seber.
Adaptive Sampling.
John Wiley & Sons, Inc., 1996.

J. Wywiał.
On space sampling.
Statistics in Transition, Vol.7, Nr 7, pages 1185–1191, 1996.

S. Zubrzycki.
Remarks on random stratified and systematic sampling in a plane.
Colloquium Mathematicae 6, pages 251–264, 1958.

Thank you for the attention