Enhancing Small Area Estimation through Spatially Balanced Designs

M. Simona Andreano
Universitas Mercatorum

Francesco Pantalone
University of Perugia

UNIVERSITÀ DEGLI STUDI FIRENZE
INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS
SIS Società Italiana di Statistica
S2G Survey Sampling Group of the Italian Statistical Society
ITACOSM 2019
FLORENCE 5-7 June 2019 | 6th ITAlian COnference on Survey Methodology

Contents of the presentation

- Review of Spatially balanced sampling (SBS):

  - *Generalized Random Tesselation Stratified*

  - *Spatially correlated Poisson Sampling*

  - *Local Pivotal Method*

  - *Product within Distance*

- Small area estimation and SBS
- Simulation and results
- References

Spatially Balanced Sampling: an overview

- Usually units in spatial population exhibit spatial dependence. In particular, units close together tend to be similar, because influenced by the same set of factors.
- In order to improve estimator efficiency, it is desirable taking into account this information.
- In the design phase, that would mean to move from traditional sampling designs, which do not consider the spatial information of the population, to spatial sampling designs.
- A good strategy would be to select sample well spread over the population of interest, or *spatially balanced samples*, in order to capture spatial heterogeneity.

Spatially Balanced Sampling: an overview

- Finite spatial population $U = \{1, 2, \ldots, N\}$
- Response variable y_i
- Coordinates $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_j, \ldots, \mathbf{c}_h\}$
- Auxiliary variables $\mathbf{x}_i = \{x_1, \ldots, x_q\}$
- Target of inference $t_y = \sum_{i \in U} y_i$
- *Design-based* approach: *y* is considered fixed and the only source of randomness is coming from the selection of the sample, i.e. *p(s)*
- First-order inclusion probability: $\pi_i = \sum_{i \in S} p(s)$
- Second-order inclusion probability: $\pi_{ij} = \sum_{(i,j) \subset S} p(s)$
- Horvitz-Thompson estimator: $HT(t_y) = \sum_{i \in S} \frac{y_i}{\pi_i}$

Spatially Balanced Sampling: an overview

To take into account spatial information when desingning a sample, we introduce a spatial model:

$$\begin{cases} y_i = \mathbf{x}_i^{'}\beta + \varepsilon_i \\ E_m(\varepsilon_i) = 0 \\ Var_m(\varepsilon_i) = \sigma_i^2 \\ Cov_m(\varepsilon_i, \varepsilon_j) = \sigma_i \sigma_j \rho_{ij} \end{cases}$$

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

$$AV(t_{y,HT}) = E_s E_m \left(t_{y,HT} - t_y\right)^2 =$$

$$= E_s \left[\left(\sum_{i \in S} \frac{x_i}{\pi_i} - \sum_{i \in U} x_i \right)' \beta \right]^2 + \sum_{i,j \in U} \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$$

Spatially Balanced Sampling: an overview

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 $\longrightarrow$ if ρ_{ij} decrease with respect to distance between units, then selecting units far apart reduces this term

Generalized Random Tesselation Stratified (GRTS)

Mapping two-dimensional spatial population into one-dimensional population, by preserving some spatial order:

1. Sampling units are sorted according to a recursive, hierarchical randomization process, which tries to preserve the spatial relationship of the units;

2. Sampling units are ordered according to a function *f*, which maps the two-dimensional space of the population into a one-dimensional space, by defining an ordered spatial address;

3. The one-dimensional space of units is then divided into a number of equal-lenght segments (the line is divede in n sub segments).

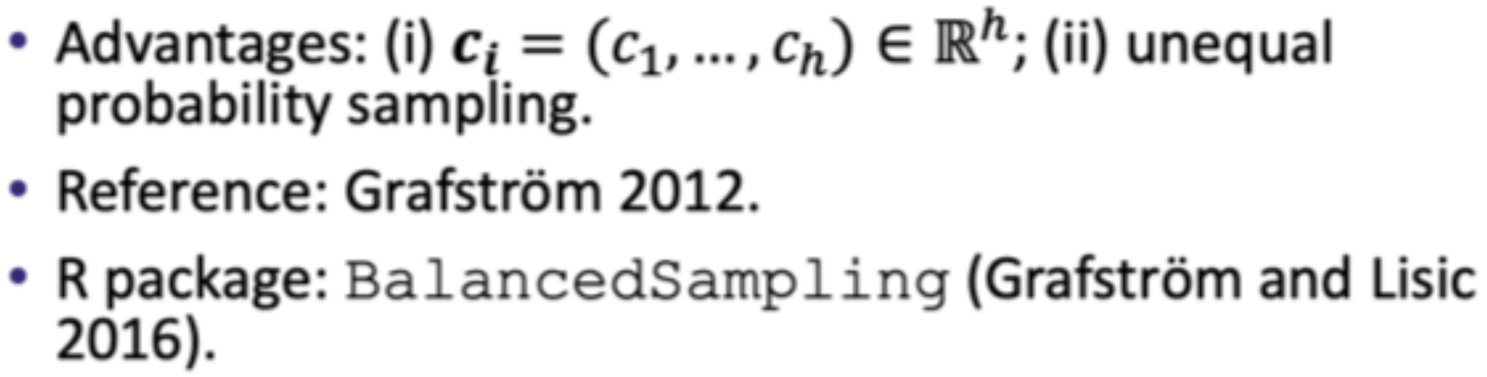
Spatially Balanced Sampling: an overview

- *Generalized Random Tesselation Stratified (GRTS)*
- Advantages: (i) spatial balance; (ii) it can be used for sampling point, linear features and not contiguous phenomena; (iii) possibility to sample with unequal probability; (iiii) practical and can be applied even in problematic situations like poor frame information and irregular space pattern.
- Disadvantages: (i) only applicable over units with a pair of coordinates ($\boldsymbol{c_i} = (c_1, c_2) \in \mathbb{R}^2$); (ii) possibility to lose some spatial relationship during the use of f.
- Reference: Steven and Olsen 2004.
- R package: `spsurvey` (Kincaid and Olsen 2016).

Spatially Balanced Sampling: an overview

Spatially Correlated Poisson Sampling (SCPS)

- is a modification of Correlated Poisson Sampling (CPS), introduced by Bondesson & Thorburn (2008);
- design based on a list sequential criterion;
- the probability function for SCPS can be written as Poisson;
- At each step *t* the outcome of the *t*-th unit in the list is decided and inclusion probabilities are updated through the use of weights creating negative correlations between close units;



- Advantages: (i) $\boldsymbol{c_i} = (c_1, \ldots, c_h) \in \mathbb{R}^h$; (ii) unequal probability sampling.
- Reference: Grafström 2012.
- R package: `BalancedSampling` (Grafström and Lisic 2016).

Spatially Balanced Sampling: an overview

- *Local Pivotal Method (LPM)*
- Sampling mechanism: modification of the Pivotal Method (). At each step, two nearby randomly units are chosen and the sampling outcome is decided for at least one of them. The procedure continue until a decision for each unit in the population has been reached. The sample is obtained in N steps.
- Advantages: (i) $\boldsymbol{c_i} = (c_1, \dots, c_h) \in \mathbb{R}^h$; (ii) unequal probability sampling.
- Reference: Grafström et al 2012.
- R package: `BalancedSampling` (Grafström and Lisic 2016).

Spatially Balanced Sampling: an overview

Product Within Distance (PWD)

- define a design *p*(*S*) with a selection probability of each sample proportional to some synthetic index M(D_s) of the within sample distance matrix D_s.
- is based on a Markov Chain Monte Carlo (MCMC) procedure whose aim is to generate samples *s* directly from the distribution *p*(*S*) without assumptions on the first- or second- order inclusion probabilities.
- Sampling mechanism: MCMC-based algorithm. At each step a Markov-chain is run and, given the actual configuration, a new one is selected or rejected according to an acceptance rule.

Spatially Balanced Sampling: an overview

Product Within Distance (PWD)

- Presence of a tuning parameter β that controls the spreading of the sample: the higher is β, the more spread will be the sample;
- Deciding if updating the sample in accordance to:

$$p = \min\left[1, \left(\frac{\mathrm{M}\left(D_{s_e^{(t)}}\right)}{\mathrm{M}\left(D_{s^{(t)}}\right)}\right)^{\beta}\right]$$

- Advantage: unequal probability sampling, applied on merging different surveys or surveys in dfferent time
- Reference: Benedetti and Piersimoni 2017
- Rpackage: Spbsampling (Pantalone, Benedetti, Piersimoni 2019)

Small Area Estimation: an overview

Suppose we have a population of interest, i.e. a population for which the survey was designed. In this case *direct estimators*, e.g. Horvitz-Thompson, should be reliable for this population.

What happens when we change our target population in some sub-population, or *domain*, of the original one?

Two problems:

1. the survey is not designed for the new population;
2. we can end up with domain with few (or even zero) observations.

In this situation, the direct estimators could be not reliable and could have low efficiency.

Solution (a possible one): *Small Area Estimation (SAE)*.

Small Area Estimation: an overview

Two extreme cases:

(i) domain with zero observation: we need a *model*;

(ii) domain with high number of observations: the *direct estimator* works fine.

Intermediate case:

(iii) domain with a few observations: the direct estimator still works, but the reliability is lower.

Idea:

- use of both: *composite estimator*.

Small Area Estimation: an overview

A composite estimator is a linear combination between different estimators.

In our case, we want to combine the direct estimator with a model estimator, in order to face the trade-off between the instability of the direct estimator with the potential bias of the model estimator.

The weight of the linear combination should account for all the situations we have seen previously:

Situation (i): we can use only the model estimator;

Situation (ii): we want to use the direct estimator;

Situation (iii) we want to use both.

Small Area Estimation: an overview

- *Fay-Harriot Model – Area Level Approach*
- Area level model: $\theta_i = \boldsymbol{x}_i\boldsymbol{\beta} + z_i u_i$

with θ_i parameter of interest, $\boldsymbol{x}_i$ area specific covariates, $\boldsymbol{\beta}$ regression parameters vectors, z_i known positive constant and $u_i \sim \mathrm{N}(0, \sigma_u^2)$

- Assumption about the direct estimator: $\hat{\theta}_i = \theta_i + e_i$

with e_i independent sampling error with mean 0 and know variance σ_e^2

At the end, the Fay-Harriot Model is:

$$\hat{\theta}_i = \boldsymbol{x}_i\boldsymbol{\beta} + z_i u_i + e_i$$

Small Area Estimation: an overview

Fay – Harriot Model – Area level Approach

- Empirical BLUP (EBLUP) for the parameter of interest θ

$$\tilde{\theta}_i = \hat{\Phi}_i\hat{\theta}_i + \left(1 - \hat{\Phi}_i\right)\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + z_i'\hat{\boldsymbol{u}}\right)$$

with $\hat{\Phi}_i = \frac{z_i^2\hat{\sigma}_u^2}{z_i^2\hat{\sigma}_u^2+\sigma_e^2}$ *shrinkage factor.*

The shrinkage factor is used to assign the weights to the estimators according to the number of observations of the domain: the higher it is, the more the weight of the direct estimator will be. At the opposite, the lower the number of observarions, the higher the weight of the model estimator will be.

Small Area Estimation: an overview

- *Fay-Harriot Model – Area Level Approach*

- MSE of the Fay-Harriot estimator

$$MSE\left(\tilde{\theta}_i\right) = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2)$$

where $g_{1i}(\sigma_u^2)$ is due to random errors, $g_{2i}(\sigma_u^2)$ is due to $\widehat{\boldsymbol{\beta}}$ and $g_{3i}(\sigma_u^2)$ is due to $\hat{\sigma}_u^2$.

- Approximately correct estimates of the MSE

$$\widehat{MSE}\left(\tilde{\theta}_i\right) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2)$$

SBS & SAE

- *Why using SBS with SAE?*

We can reduce the first term of the MSE since it is due to the independent sampling errors. In fact

$$g_{1i}(\sigma_u^2) = \widehat{\Phi}_i \sigma_e^2$$

and since the use of a spread sample reduces the sampling variance, the $g_{1i}(\sigma_u^2)$ reduces as well, which in turn reduces the $MSE\left(\tilde{\theta}_i\right)$.

SBS & SAE

- *Why using SBS with SAE?*

Theorem (Grafstrom 2012)

Let the population consist of two separated regions, A and B, such that the within region distances are always less than the distance between units in different regions. If $\sum_{s \in A} \pi_i = n_A$ and $\sum_{s \in B} \pi_i = n_B$, where n_A and n_B are positive integers, then the maximal weight strategy produces samples of fixed sizes, n_A and n_B respectively.

SBS & SAE

- *Why using SBS with SAE?*

Theorem

If the population is partitioned in $d = 1, \dots, D$ domains such that every unit $i \in d$ can be included only in the Voronoi polygon of a selected unit $j \in d$, then if $\sum_{i \in d} \pi_i = n_d$ the sample size in each domain will be equal to the size of the domain divided by the average size of its Voronoi polygons

$$n_d^* = \frac{n_d}{M_d(v_i)} \text{ thus } E[n_d^*] = n_d \text{ and } V(n_d^*) \approx K(n_d)V_d(v_i)$$

If the sample is spatially balanced then $V_d(v_i) = 0$ and any not planned domain regular shaped will have a controlled sample size.

Simulation

Objective of the simulation: testing the efficiency of the EBLUP under different SBS, in order to understand the impact of them on the final estimates.

Target area: Italian region “Piemonte”.

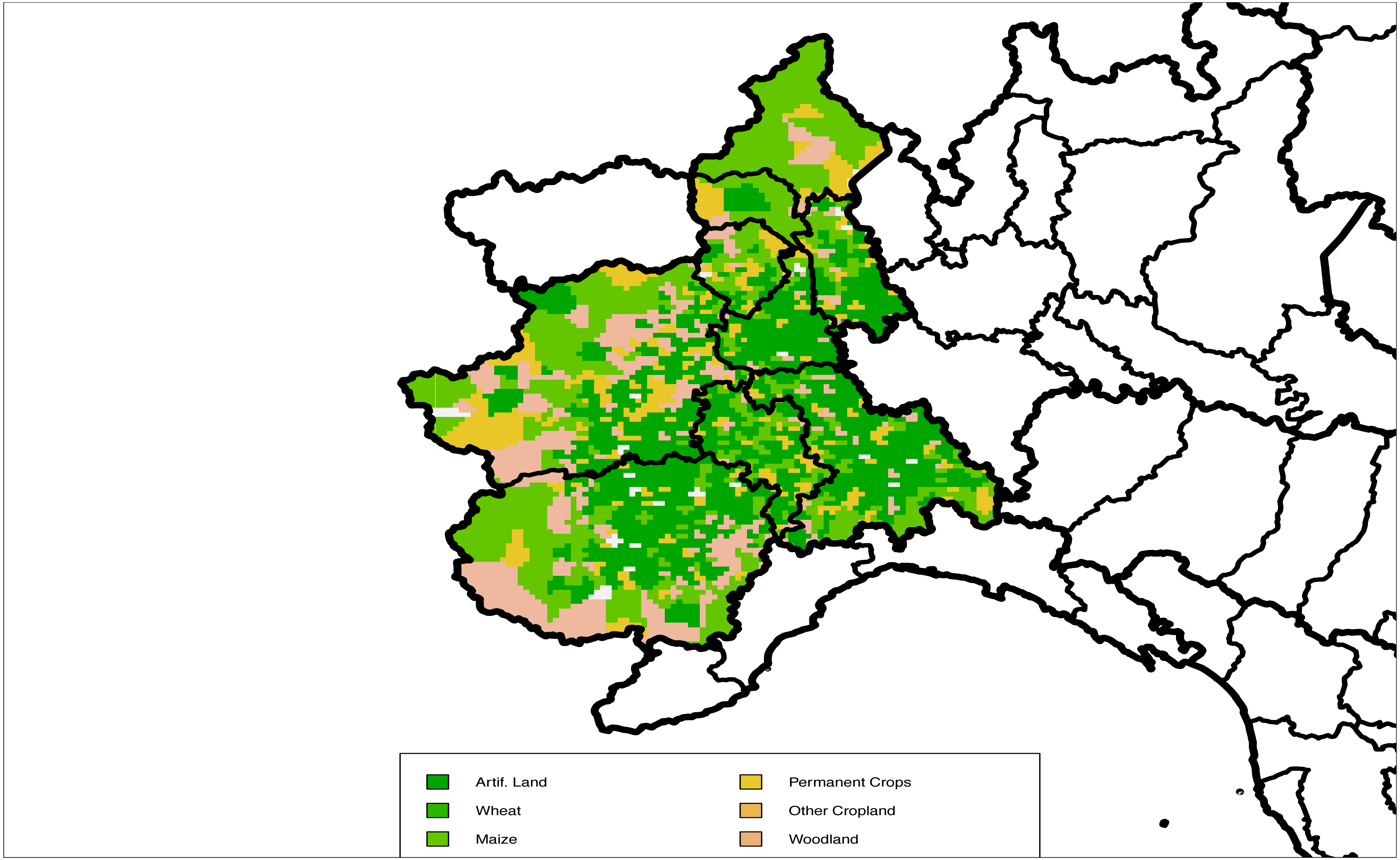
Small areas: 8 provinces, Verbania, Vercelli, Novara, Biella, Torino, Alessandria, Asti, Cuneo.

Target of inference: total of area dedicated to maize at the province level.

Sample data: Land Use/Cover Area Frame Survey (LUCAS).

Auxiliary information: satellite data.

Simulation


Artif. Land
Wheat
Maize
Permanent Crops
Other Cropland
Woodland

Simulation

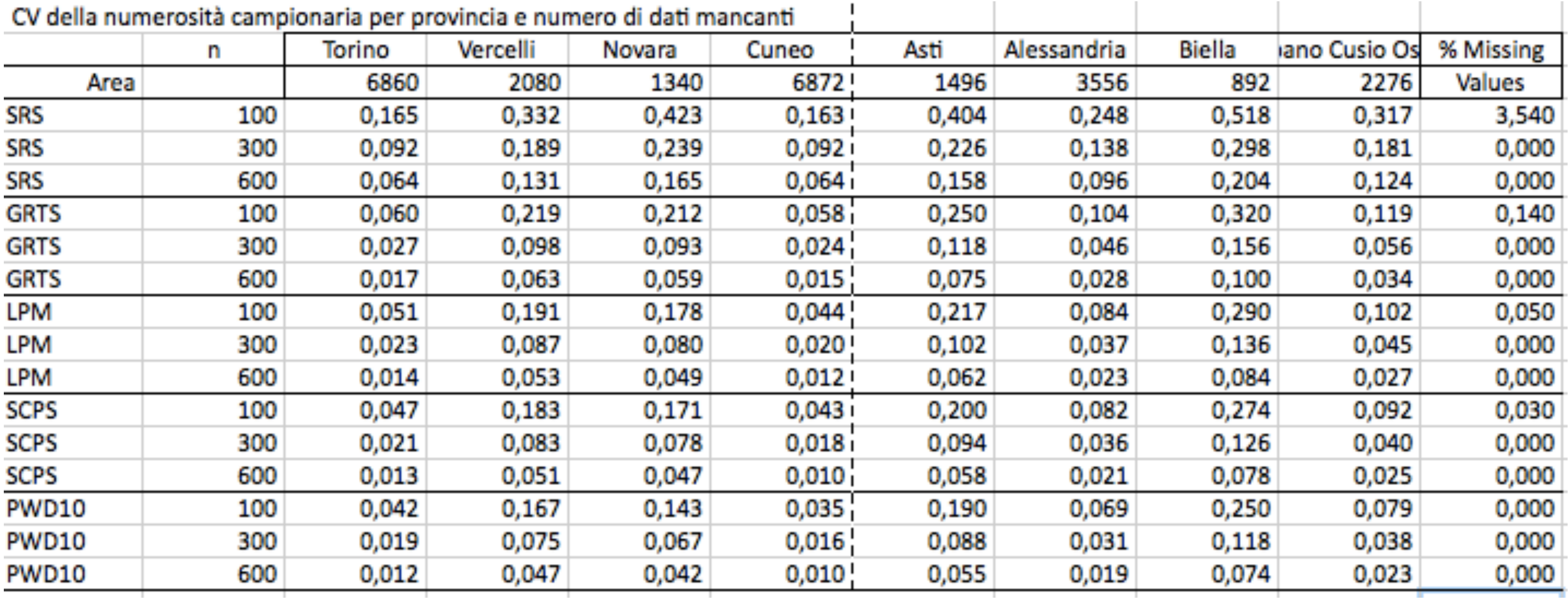
-

Per each design considered, we selected 1000 samples and computed the relative EBLUPs for the small areas. Next, we computed the RMSE and compared with the RMSE of the EBLUPs computed by SRS.

Designs considered:

- GRTS;
- LPM;
- SCPS;
- PWD with parameter of spread $\beta = 10$.

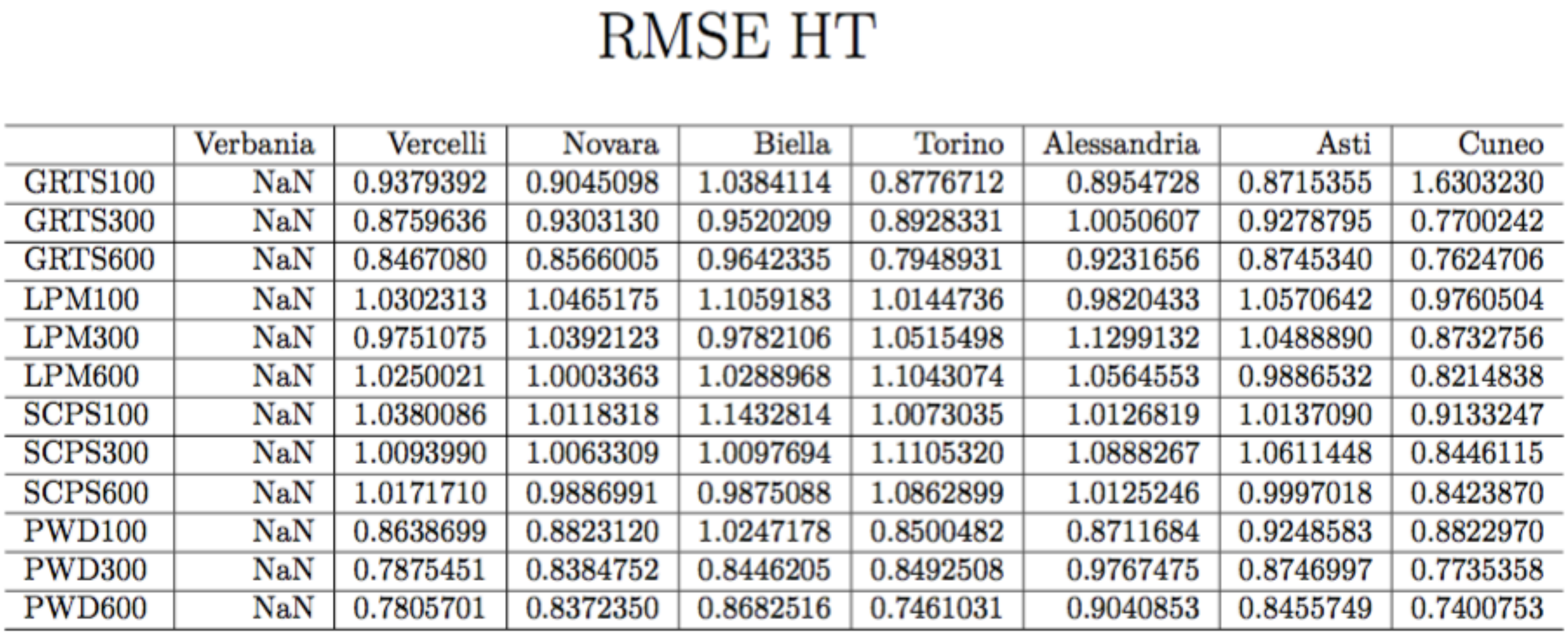
Simulation



CV della numerosità campionaria per provincia e numero di dati mancanti

	n	Torino	Vercelli	Novara	Cuneo	Asti	Alessandria	Biella	ano Cusio Os	% Missing
Area		6860	2080	1340	6872	1496	3556	892	2276	Values
SRS	100	0,165	0,332	0,423	0,163	0,404	0,248	0,518	0,317	3,540
SRS	300	0,092	0,189	0,239	0,092	0,226	0,138	0,298	0,181	0,000
SRS	600	0,064	0,131	0,165	0,064	0,158	0,096	0,204	0,124	0,000
GRTS	100	0,060	0,219	0,212	0,058	0,250	0,104	0,320	0,119	0,140
GRTS	300	0,027	0,098	0,093	0,024	0,118	0,046	0,156	0,056	0,000
GRTS	600	0,017	0,063	0,059	0,015	0,075	0,028	0,100	0,034	0,000
LPM	100	0,051	0,191	0,178	0,044	0,217	0,084	0,290	0,102	0,050
LPM	300	0,023	0,087	0,080	0,020	0,102	0,037	0,136	0,045	0,000
LPM	600	0,014	0,053	0,049	0,012	0,062	0,023	0,084	0,027	0,000
SCPS	100	0,047	0,183	0,171	0,043	0,200	0,082	0,274	0,092	0,030
SCPS	300	0,021	0,083	0,078	0,018	0,094	0,036	0,126	0,040	0,000
SCPS	600	0,013	0,051	0,047	0,010	0,058	0,021	0,078	0,025	0,000
PWD10	100	0,042	0,167	0,143	0,035	0,190	0,069	0,250	0,079	0,000
PWD10	300	0,019	0,075	0,067	0,016	0,088	0,031	0,118	0,038	0,000
PWD10	600	0,012	0,047	0,042	0,010	0,055	0,019	0,074	0,023	0,000

Simulation



RMSE HT

	Verbania	Vercelli	Novara	Biella	Torino	Alessandria	Asti	Cuneo
GRTS100	NaN	0.9379392	0.9045098	1.0384114	0.8776712	0.8954728	0.8715355	1.6303230
GRTS300	NaN	0.8759636	0.9303130	0.9520209	0.8928331	1.0050607	0.9278795	0.7700242
GRTS600	NaN	0.8467080	0.8566005	0.9642335	0.7948931	0.9231656	0.8745340	0.7624706
LPM100	NaN	1.0302313	1.0465175	1.1059183	1.0144736	0.9820433	1.0570642	0.9760504
LPM300	NaN	0.9751075	1.0392123	0.9782106	1.0515498	1.1299132	1.0488890	0.8732756
LPM600	NaN	1.0250021	1.0003363	1.0288968	1.1043074	1.0564553	0.9886532	0.8214838
SCPS100	NaN	1.0380086	1.0118318	1.1432814	1.0073035	1.0126819	1.0137090	0.9133247
SCPS300	NaN	1.0093990	1.0063309	1.0097694	1.1105320	1.0888267	1.0611448	0.8446115
SCPS600	NaN	1.0171710	0.9886991	0.9875088	1.0862899	1.0125246	0.9997018	0.8423870
PWD100	NaN	0.8638699	0.8823120	1.0247178	0.8500482	0.8711684	0.9248583	0.8822970
PWD300	NaN	0.7875451	0.8384752	0.8446205	0.8492508	0.9767475	0.8746997	0.7735358
PWD600	NaN	0.7805701	0.8372350	0.8682516	0.7461031	0.9040853	0.8455749	0.7400753

Simulation

	Verbania	Vercelli	Novara	Biella	Torino	Alessandria	Asti	Cuneo
GRTS100	0.6360868	1.5635796	0.9888364	0.6794555	1.4223646	0.8675475	1.5691916	1.5296811
GRTS300	0.8353760	0.9034418	1.0170633	0.9447852	0.6173188	1.2353202	0.8641661	0.5553207
GRTS600	0.9346582	0.6455000	0.8420095	0.9188386	0.4561037	1.1307870	0.6653687	0.4170726
LPM100	0.8859967	1.2638340	1.1547011	0.9282241	1.0655236	1.2626772	1.2646547	0.9607005
LPM300	0.9325531	0.8393133	1.3136664	1.0491140	0.6977828	1.5577502	0.8550787	0.5814019
LPM600	1.0621174	0.5987429	1.0063838	1.0596980	0.5833469	1.3774114	0.5750544	0.4045299
SCPS100	0.8796000	1.2495079	1.0724639	0.9267192	1.0214755	1.2282016	1.1788845	0.9027995
SCPS300	0.9661380	0.8831110	1.2819012	1.1082960	0.6821520	1.5427569	0.8075880	0.5480727
SCPS600	1.1021186	0.6043376	0.9777239	1.0584320	0.5777138	1.3890846	0.5685249	0.4033101
PWD100	0.7943272	1.0715450	0.8741273	0.8035085	0.9479421	1.0528199	1.2049230	0.8970312
PWD300	0.7985164	0.8704397	0.9690116	0.8564637	0.6110313	1.2165179	0.8545213	0.5523522
PWD600	0.9059844	0.6511020	0.7414879	0.8171829	0.4222790	1.0944463	0.6687905	0.4105362

	Verbania	Vercelli	Novara	Biella	Torino	Alessandria	Asti	Cuneo
Total area maize	0	47	37	14	136	51	49	145

Simulation

- Apart for the province of “Alessandria”, the EBLUPs obtained by SBS are generally more efficient than the EBLUPs obtained by SRS.
- In some cases, we reach a gain in the efficiency around the 60%.

Conclusions

- SBS: taking into account the spatial information when selecting a sample.
- SAE: face the challenge of obtain reliable estimates for small areas (likely with few observations) through the use of a composite estimator.
- SBS + SAE: the results show a possible gain in the efficiency of the final estimates if SBS and SAE are used together. This is due to the possible reduction of the first term of the MSE of the EBLUP.

References

- Benedetti R, Piersimoni F (2017) A spatially balanced desing with probability function proportional to the within sample distance, , Biometical Journal, 59, 1067-1084.
- 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.
- 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, 2, 514-520.
- Stevens Jr. DL, Olsen AR (2004). Spatially balanced sampling of natural resources. JASA, 99, 262–278.