PROCEEDINGS

Edited By: Giada Adelfio and Antonino Abbruzzo

GRASPA 2023

GRASPA-SIS BIENNAL CONFERENCE

The Researcher Group for Environmental Statistics of The Italian Statistical Society

TIES EUROPEAN REGIONAL MEETING

The International Environmetrics Society

Palermo, 10-11 July, 2023

Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo

Proceedings of the GRASPA 2023 Conference
Palermo, 10-11 July 2023
Edited by: Giada Adelfio and Antonino Abbruzzo
–
Palermo: Università degli Studi di Palermo.

ISBN: 979-12-210-3389-2

Sponsored by:

Some empirical results on nearest neighbour pseudo-populations for resampling from spatial populations

Di Biase RM[1,2,*], Marcelli A[3], Franceschi S[1] and Fattorini L[1]

[1]*Department of Economics and Statistics, University of Siena, Italy; rosa.dibiase@unisi.it, sara.franceschi@unisi.it, lorenzo.fattorini@unisi.it*
[2] *NBFC, National Biodiversity Future Center, Palermo, Italy*
[3] *Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia, Italy; agnese.marcelli@unitus.it*
[*]*Corresponding author*

Abstract. *In finite populations, pseudo-population bootstrap is the sole method preserving the spirit of the original bootstrap performed from iid observations. In spatial and environmental sampling, the issue of creating pseudo-populations able to mimic the characteristics of real populations is challenging because spatial trends, relationships, and similarities among neighboring locations are invariably present. In this paper we propose the use of the nearest-neighbor interpolation of spatial populations for constructing pseudo-populations that converge to real populations under mild conditions. The effectiveness of these proposals with respect to traditional pseudo-populations is empirically checked by a simulation study.*

Keywords. *Spatial surveys, Horvitz-Thompson estimator, Spatially balanced sampling, Pseudo-population bootstrap, Nearest-neighbour criterion.*

---

1 Introduction

In finite population sampling, pseudo-population bootstrap (PPB) is the sole method preserving the spirit of the original bootstrap, where samples are iid data from an unknown distribution and bootstrap samples, like the original ones, are iid data from the empirical distribution.
The effectiveness of any PPB technique rests on the pseudo-population (PP) from which the bootstrap resampling is performed. Indeed, the crucial issue is the creation of PPs able to mimic the characteristics of the real population from which the sample has been selected. This issue is even more challenging in environmental and spatial surveys, where spatial trends, relationships, and similarities among neighboring locations are invariably present.
To solve this problem, we propose the use of PPs suitable to work in spatial and environmental surveys. In particular, we propose the use of the nearest-neighbor interpolation (NNI) of spatial populations for constructing PPs that were proven to converge to real populations under mild conditions. In this way, for population and sample sizes sufficiently large, the resulting PPs are likely to be good pictures of the actual spatial populations.
We empirically evaluated the effectiveness of the PPs based on the NNI with respect to other commonly used PPs to approximate the distribution of the Horvitz-Thompson (HT) estimators of total in spatial

sampling. Comparison is performed with respect to several spatial populations showing different spatial patterns, different sizes and two spatially balanced sampling schemes: local pivotal method (LPM) and doubly balanced spatial sampling (DBSS).
The paper is organized as follows: section 2 describes the problem of estimating totals and averages in spatial surveys and introduces some criteria to construct PPs suitable to work with spatial populations, including the NNI criterion. In section 3 the PPs considered in section 2 are compared by a simulation study in terms of the performance of the bootstrap distributions to fit the actual distribution of the HT estimators of total under two schemes usually adopted in spatial sampling, as well as in terms of the coverage of bootstrap confidence intervals. Concluding remarks are provided in section 4.

2 Spatial pseudo-populations

Design-based inference has been widely adopted for estimating totals and averages of finite spatial populations, especially in environmental studies, where the estimation is usually performed via the HT criterion or its modifications able to exploit auxiliary information. In environmental and spatial studies, spatially balanced schemes, such as the local pivotal method (LPM) by [6] and the doubly balanced spatial sampling (DBSS) by [5], are widely used. These schemes not only ensure that the sampling locations are spread out throughout the area but also ensure balance with respect to a set of auxiliary variables. Using spatially balanced schemes, the selection of neighbouring units is avoided, which means that a large portion of second-order inclusion probabilities is null. However, the exclusion of neighbouring units in the samples precludes the use of a finite-population central limit theorem and, consequently, traditional confidence intervals using the standard normal quantiles cannot be constructed. Therefore, in spatial surveys the use of PPB may be a suitable solution for making inference on the distribution of the HT estimator and for constructing confidence intervals.
As stated in the Introduction, the effectiveness of any PPB rests on the capacity of constructing PPs able to well depict the real population from which the sample has been selected. However, some PPs commonly adopted in PPB are not suitable for mimic spatial populations, such as those that provide random population sizes.
Consider a study region $\mathcal{A}$ which is supposed to be a connected and compact set of $\mathbb{R}^2$. As it is customary in the finite population asymptotic framework (e.g. [7]), in spatial surveys we suppose $\mathcal{V} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \dots\}$ be an infinite sequence of locations onto $\mathcal{A}$ and $y(\mathcal{V}) = \{y_1, y_2, \dots\}$ be the corresponding sequence of the interest variable Y values, where for brevity $y_j = y(\boldsymbol{p}_j)$. Moreover, let $x(\mathcal{V}) = \{x_1, x_2, \dots\}$ be the corresponding sequence of a strictly positive size variable X, where for brevity $x_j = x(\boldsymbol{p}_j)$.
By a little abuse of notation, we shall denote locations by their labels. A sequence $\{U_k\}$ of spatial populations is considered, in such a way that $\{U_k\}$ turns out to be a sequence of nested populations of increasing sizes within $\mathcal{A}$. Finally suppose a sequence of spatial designs $\{P_k\}$ to select spatially balanced samples S_k from U_k of fixed and increasing size $n_k = pN_k$ for a fixed sampling fraction $0 < p < 1$. For each k and for each location $j \in U_k$, denote by $\pi_{j(k)}$ the first-order inclusion probability induced by the kth design P_k that is taken proportional to the size variable X. The key motivation for this choice is the efficiency of the HT estimator of population totals and averages when there is a strong direct relationship between X and Y. For each k and for each pair of locations $j, h \in U_k$, denote by $\pi_{j,h(k)}$ the second-order inclusion probability induced by the kth design.
In this paper, we considered three PPs: the multinomial PP, the hot-deck PP and the PP based on NNI.
The Multinomial PP (MPP) by [8] is one among the "traditional" PPs that can be applied also in spatial sampling. For each location $j \in U_k$, MPP independently assigns to j the values $\hat{y}_j = y_h$ and $\hat{x}_j = x_h$ with

probability

$$\Pr\left\{\hat{y}_j = y_h\right\} = \frac{x_h^{-1}}{\sum_{l\in S_k} x_l^{-1}} \ , \ h \in S_k$$

The resulting PP has obviously the same size N_k of the real population. Even if virtually applicable in spatial surveys, MPP takes into consideration only the information provided by the size variable X, completely neglecting the spatial locations, while in most cases the Y values strictly depend on locations. As such, MPP is likely to provide poor representations of real populations.
The hot-deck PP (HDPP) is another PP that can be applied in spatial surveys. HDPP has been proposed by [1] based on the idea that the values of the size variable are good proxies for the Y values. Then, for each location $j \in U_k$, HDPP assigns to j the values $\hat{x}_j = x_j$ and

$$\hat{y}_j = Z_{k,j} y_j + \left(1 - Z_{k,j}\right) y_{nn_x(j)} \tag{1}$$

where $Z_{k,j}$ is the sample indicator variable that is equal to 1 if $j \in S_k$ and it is equal to 0 otherwise and $nn_x(j) = \text{argmin}_{h\in S_k} \left|x_j - x_h\right|$, i.e., HDPP predicts the Y value at any unsampled location j by the Y value of the sampled location that is nearest to j in the space of the X values. Practically speaking, prediction (1) constitutes a NNI in the X space. Even if HDPPs do not directly consider the information provided by locations, locations enter in the predictions by the fact that any $x_j = x(\boldsymbol{p}_j)$ is actually a function of its location.
Alternatively, we here propose the use of NNI in which predictions are completely based on the spatial coordinates. NNI exploits the well-known Tobler's first law of geography, for which the Y values at locations that are close in space tend to be more similar than those at locations that are far apart [10]. Therefore, for each location $j \in U_k$, this criterion, henceforth referred to as nearest-neighbour pseudo population (NNPP), assigns to j the values $\hat{x}_j = x_j$ and

$$\hat{y}_j = Z_{k,j} y_j + \left(1 - Z_{k,j}\right) y_{nn_g(j)} \tag{2}$$

where in this case $nn_g(j) = \text{argmin}_{h\in S_k} \left|\boldsymbol{p}_j - \boldsymbol{p}_h\right|$, i.e., we predict the Y value at any unsampled location j by the Y value of the sampled location that is nearest to j in the geographical space. Practically speaking, NNPP assigns the value of a sampled unit to each unsampled unit inside the Voronoi cell constructed around the sampled unit.
[4] determines the consistency condition of NNI from a design-based point of view. In particular, consistency holds if the adopted sampling scheme provides spatial balance. Owing to consistency, for population and sample sizes sufficiently large, NNPPs are likely to provide precise representation of the real spatial population.

3 Simulation study

The purpose of this study is to empirically evaluate the performance of three described criteria for constructing PPs from which resampling is performed to approximate the distribution of the HT estimators of total in spatial sampling. Comparison is performed with respect to several spatial populations showing different spatial patterns, several spatially balanced sampling schemes and several population sizes whose increase mimics the sequence of nested populations theoretically supposed throughout the paper.
To generate finite and nested spatial populations, an artificial surface on the unit square was considered, where for any point $\boldsymbol{p} = [p_1, p_2]$ the surface was defined by

$$y(\boldsymbol{p}) = C sin(3p_1) sin^2(3p_2) \tag{3}$$

where the constant C ensured a maximum Y value of 10. The surface (3) was chosen to represent the major characteristics of spatial populations. It was continuous, in such a way that the Y values in neighbouring locations tended to be similar, thus entailing a spatial autocorrelation in the resulting populations. Moreover, it varied relevantly throughout the unit square showing an increasing trend toward the centre of the square, thus entailing a spatial stratification with different values of the survey variable in different parts of the square.
From (3), three nested populations of $N = 250,\ 500, 1000$ points were located in the unit square in accordance with four spatial patterns referred to as regular, random, trended, and clustered patterns.
For any spatial population arising from the combination of spatial patterns and population sizes, $R = 10000$ samples of fixed size $n = 0.1N$ were independently selected by means of LPM and DBSS. Then for each sample S_i selected at the i-th simulation run ($i = 1, \ldots, R$), the HT estimate of the population total T_y was computed by means of

$$\hat{T}_i = \frac{T_x}{n} \sum_{j \in S_i} \frac{y_j}{x_j}$$

where T_x was the total of size variable in the population. Moreover, from the sample S_i, a PP ($\hat{y}_{i,j},\ \hat{x}_{i,j},\ j = 1, \ldots, N$) was created in accordance with each of the three criteria considered in section 2, i.e. MPP, HDPP, and NNPP. From each PP, $B = 1000$ bootstrap samples

$$S^*_{i,1}, \ldots, S^*_{i,B}$$

were selected adopting the same sampling scheme adopted to select the original sample S_i, and for each bootstrap sample the HT estimate of the population total was computed by means of

$$\hat{T}^*_{i,b} = \frac{\hat{T}_{i,x}}{n} \sum_{j \in S_{i,b}} \frac{\hat{y}_{i,j}}{\hat{x}_{i,j}} \quad , \quad b = 1, \ldots, B$$

where $\hat{T}_{i,x}$ was the total of the size variable in the i-th PP.
For each combination of spatial patterns and population sizes, the Monte Carlo distributions of the HT estimators of total $\hat{T}_1, \ldots, \hat{T}_R$ were adopted to empirically determine the actual distribution of the estimator and the relative standard error. At this point, the ability of PPB to mimic the actual distribution of the HT estimator was determined for each of the three PP criteria by means of their worst fitting (WF) performance quantified by the two-sample Kolmogorov-Smirnov statistic.
Moreover, the mimic ability of PPs was quantified by the capacity of the 95% bootstrap confidence intervals to approach the nominal level of 95% that was determined by means of their empirical coverage associated with their average length.
Finally, the capacity of the bootstrap distributions to reproduce the actual precision of the HT estimators was determined comparing the empirical expectations of the bootstrap relative standard error estimators to the actual relative standard errors. Because the actual relative standard error and their bootstrap estimates were likely to approach 0 as population and sample sizes increased, their ratio (RAT) was adopted.

3.1 Simulation results

Table 1 reports the results of the simulation study for the random population under DBSS, but similar results also apply for the other spatial patterns under both sampling schemes. As for the relative standard errors (RSE), they quickly decrease as the population sizes increase, showing the presumable consistency of the HT estimator of population totals under DBSS. These findings agree with [3] that theoretically proved the consistency of the HT estimation in spatial populations under very simple schemes such as simple random sampling without replacement (SRSWOR), but without proving the consistency under

more complex spatially balanced schemes such as LPM and DBSS owing to the lack of analytical expressions of the second-order inclusion probabilities. However, stated the superiority of these schemes in providing spatial balance with respect to SRSWOR, they concluded that "consistency presumably holds also for these schemes". The fitting index (FIT) was evaluated as the Monte Carlo distributions of the root of average squared errors. The simulation results show that the FIT of the NNPPs quickly improve as the population sizes increase, confirming the presumable consistency of the NNI under these spatial schemes, whereas the FIT shows that consistency does not hold for the HDPPs and it is even worse for MPPs. The RAT values achieved under NNPPs are always greater than 1, but invariably smaller than those achieved under HDPPs and MPPs, and quickly approach 1 as the population sizes increase, showing a tendency to be moderately conservative. On the other hand, RAT values achieved by HDPPs and MPPs show a tendency to a large overestimation that unsuitability masks the actual precision of the two spatial strategies and that does not decrease as the population sizes increase. The superiority of NNPPs was also demonstrated by the performance of bootstrap confidence intervals that for all the PP criteria show coverages similar to or greater than the nominal level, but with average lengths that in the case of NNPPs are much smaller, in some cases even two-three times smaller, than those achieved by HDPPs and MPPs. The same conclusion hold for maximum values of the two-sample Kolmogorov-Smirnov statistic.

SP	N	RSE	PP	FIT	RAT	C_{95B} (L_{95B})	WF
Random	250	5.19	MPP	3.96	2.16	94.61(362.74)	1.00
			HDPP	2.64	2.33	96.93(366.47)	0.99
			NNPP	1.40	2.01	94.29(300.56)	0.97
	500	3.02	MPP	4.04	2.63	95.33(514.21)	1.00
			HDPP	2.65	2.74	97.86(523.63)	0.99
			NNPP	0.95	1.68	97.92(316.92)	0.91
	1000	1.7	MPP	3.98	3.42	96.68(747.25)	1.00
			HDPP	2.67	3.46	98.39(748.44)	1.00
			NNPP	0.66	1.66	99.70(338.19)	0.97

Table 1: Values of RSE of the HT estimator of totals, FIT, RAT, coverages of the 0.95 bootstrap confidence intervals (C_{95B}) and expectations of their lengths (L_{95B}, in parentheses) and WF. Reported values refer to the results for the random spatial pattern under DBSS.

4 Final remarks

Spatial surveys, and especially environmental surveys, have been traditionally approached from a design-based perspective, bypassing the complex task of modelling spatial phenomena, viewing these phenomena as fixed and attributing uncertainty only to sampling (e.g., [9]). In the last years, even the mapping of ecological resources, traditionally approached in the realm of the model-based geostatistical procedures (e.g. [2]), has been approached in a design-based perspective by [4] that derive the design-based properties of the NNI.

Because in a design-based framework properties of any estimator are completely determined by the sampling design, the design choice is then crucial in this context. Regarding the sampling schemes usually adopted in spatial surveys, in this paper we have emphasized the importance of spatial balance, i.e. the capacity of the sampling schemes to evenly spread locations over the study region in such a way that no portion of the region is over- or under-represented. At the same time, we have also outlined the drawbacks involved by the use of spatially balanced schemes, i.e. a) the impossibility of using finite

population central limit theorems for confidence intervals and b) the presence of some second-order inclusion probabilities equal to 0 that precludes the unbiased estimation of variance.
Because the use of PPB seems to be a viable solution to overcome both the issues, the focus of the paper has switched to the choice of PPs capable of providing good representations of the spatial populations from which balanced samples are selected. At this step, the results by [4] have been crucial. The authors proved the design-based consistency of the NNI under mild conditions regarding the spatial populations and the sampling schemes. Conditions about populations simply require the smoothness of the Y values in neighbouring locations, that well approaches the theoretical condition of local continuity, while conditions on the sampling scheme simply require an asymptotical spatial balance that is satisfied even by SRSWOR. Therefore, when using spatial scheme explicitly tailored for achieving spatial balance, the consistency of NNI should hold *a fortiori*. For these reasons, the NNI of real populations has been proposed as a criterion for constructing PPs in spatial surveys, referred to as NNPPs. The obvious intuition behind this proposal is that if the NNPPs converge to the true populations, bootstrap distributions arising from these maps should converge to the actual distributions of the estimators.

Acknowledgments. The authors acknowledge the funding by PRIN 2020 (cod 2020E52THS) - Research Projects of National Relevance funded by the Italian Ministry of University and Research entitled: "Multi-scale observations to predict Forest response to pollution and climate change" (MULTIFOR, project number 2020E52THS). The authors also acknowledge the support of NBFC to University of Siena, funded by the Italian Ministry of University and Research PNRR, Missione 4 Componente 2, "Dalla ricerca all'impresa", Investimento 1.4, Project CN00000033.

References

[1] Conti, P.; Marella, D.; Mecatti, F.; Andreis, F. (2020). A unified principled framework for resampling based on pseudo-populations: Asymptotic theory. *Bernoulli* **26**, 1044–1069.

[2] Cressie, N.A.C. (1993). *Statistics for Spatial Data*, Revised Edition. Wiley. New York.

[3] Fattorini, L.; Marcheselli, M.; Pisani, C.; Pratelli, L. (2020). Design-based consistency of the Horvitz-Thompson estimator under spatial sampling with applications to environmental surveys. *Spat. Stat.* **35**, 100404.

[4] Fattorini, L.; Marcheselli, M.; Pisani, C.; Pratelli, L. (2021). Design-based properties of the nearest neighbour spatial interpolator and its bootstrap mean squared error estimator. *Biometrics* **78**, 1454–1463.

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

[6] Grafström, A.; Lundström, N.L.P.; Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics* **68**, 514–520.

[7] Isaki, C.T.; Fuller, W.A. (1982). Survey design under the regression superpopulation model. *J. Am. Stat. Assoc.* **77**, 89–96.

[8] Sverchkov, M.; Pfefferman, D. (2004). Prediction of finite population totals based on the sample distribution. *Surv. Methodol.* **30**, 79–92.

[9] Thompson, S.K. (2002) *Sampling*, 2nd Edition. Wiley. New York.

[10] Tobler, W.R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *J. Econ. Geogr.* **46**, 234–240.