DEM Working Papers

Department of Economics and Management

Incomplete geocoding and spatial sampling: the effects of locational errors on population total estimation

Maria Michela Dickson, Giuseppe Espa,

Diego Giuliani

N. 2016/04

UNIVERSITÀ DEGLI STUDI
DI TRENTO

Università degli Studi di Trento

Department of Economics and Management, University of Trento, Italy.

Guidelines for authors

Papers may be written in Italian or in English. Faculty members of the Department must submit to one of the editors in pdf format. Management papers should be submitted to R. Gabriele. Economics Papers should be submitted to L. Andreozzi. External members should indicate an internal faculty member that acts as a referee of the paper.

Typesetting rules:

1. papers must contain a first page with title, authors with emails and affiliations, abstract, keywords and codes. Page numbering starts from the first page;
2. a template is available upon request from the managing editors.

Incomplete geocoding and spatial sampling: the effects of locational errors on population total estimation

Maria Michela Dickson[*][♥], Giuseppe Espa[♣] and Diego Giuliani[♠]

[♥♣♠]*Department of Economics and Management, University of Trento, Via Inama 5, 38122,Trento, Italy.*

[*] Corresponding author. E-mail: mariamichela.dickson@unitn.it

Abstract

Due to the increasing availability of georeferenced microdata in several fields of research, surveys can benefit greatly from the use of the most recent spatial sampling methods. These methods allow to select spatially balanced samples, which lead to particularly efficient estimates, by incorporating the distances among the exact locations of statistical units into the design. Unfortunately, since locations of units are rarely exact in practice due to imperfections in the geocoding processes, the implementation of spatial sampling designs is actually often limited. This paper aims at demonstrating that spatial sampling designs can be implemented even when spatial information is not completely accurate. In particular, by means of a Montecarlo sampling simulation study about the estimation of water pollution, it is proved that the use of spatial sampling methods still lead to more spatially balanced samples, and more efficient estimates, also when the geocoding of population is not exact.

Keywords: GPS uncodified; Locational Accuracy; Spatial Sampling Methods; Estimation.

1. Introduction

Thanks to the development of technology for data collection and transmission, the availability of georeferenced microdata has increased rapidly over the last decades. Careful analysis of this kind of data has led, among other things, to a better understanding of spatial interactions essential, for example, to design policies, to study climate and environmental phenomena or to predict the diffusion of diseases (see, among others, Burrough, 2001; Boulos, 2004; Rushton et al., 2006; Dale and Fortin, 2014).

More recently, georeferenced microdata have also proved to be very useful in designing sample surveys. The spatial information contained in the georeferenced microdata, which is the exact point-level geographical coordinates of all subjects in the study population, can indeed be exploited to select samples that are spread in the space and/or spatially balanced. It has been shown (see e.g. Grafström and Tillé, 2013) that spatially balanced samples lead to particularly efficient estimates, especially when dealing with environmental and health data. In light of this, a recent stream of literature (Grafström et al., 2011; Grafström, 2012; Grafström and Tillé, 2013) proposed spatial sampling procedures that incorporate the distances among the exact locations of subjects into the design and assure that the selected samples are spatially balanced and spread in the geographic space.

Unfortunately, however, locations of units are rarely exact in practice, due to two kinds of problem. First, especially in studies on populations, it is necessary to protect the privacy of individuals (VanWey et al., 2005). Second, the geo-referencing process is not lacking of errors due to incomplete geo-coding. The first aspect is particularly significant in health studies, in which protecting form the so-called disclosure risk is a relevant issue (Hunderpool et al., 2010). For example, it is important to hide the identity of an individual affected by an infectious disease in order to avoid his/her exclusion from the society. In these cases, errors in location of units are imposed, while trying to preserve the spatial distribution of variables but minimizing in the same time the possibility to identify a unit. Several approaches are used to geo-mask positions, such as, among others, truncating, swapping and displacing of coordinates (VanWey et al., 2005; Curtis et al., 2006; Allshouse et al., 2010).

The second above-mentioned problem of geo-referenced data, on whose the present paper is focused, concerns locational errors induced by imperfection in data capture procedures. It happens especially in zones not perfectly covered by satellite connections to GPS, such as isolated or mountainous lands, for which spatial data are missing or they almost invariably contain locational errors (Zimmerman, 2008; Zimmerman and Li, 2010). Geocoding techniques commonly used are essentially based on the comparison between the observed addresses and referenced addresses databases of the area under observation, which require the availability of correct reference data and a robust model for the verification of addresses (Yang et al., 2004; Zandbergen, 2008; Cozzi and Filipponi, 2012; Jacquez, 2012). Sometimes, for various technical reasons, the geocoding process of all records of a dataset cannot be exact. Usually, for a share of subjects, that in some cases is not negligible, it is not possible to identify the exact spatial coordinates. This should results in georeferenced microdata with missing spatial data. However, since normally there is always some spatial information coarser than the point-level geographical coordinates observed for the subject that fail to be geocoded (e.g. an areal-level location such as the Zip code), the most frequent situation in practice is that unprecise georeferenced microdata are collapsed in a single point on the map, e.g. centroid of a region (Zimmerman, 2008).

In principle, spatial sampling designs that make use of the distances among the exact locations of subjects cannot be implemented if the population georeferenced microdata are characterized by unprecise locations. In such a case, indeed, spatial information are uncertain and cannot be used properly within the sample selection procedure and hence a non-spatial sampling design should be implemented instead. In spite of this, the present paper aims at verifying whether spatial sampling designs may still be implemented, using coarser areal-level information when the exact location is missing, assessing if they still lead to more efficient estimates than the non-spatial sampling designs. In particular, as a case study, we investigate the effects of locational errors on total population estimation of water pollutants by means of a simulation experiment based on data known in literature.

The paper is structured as follow. Section 2 contains a detailed explanation of spatial sampling methods for georeferenced microdata. Section 3 presents a simulation study based on data about the pollution of river water to assess the effects of locational errors on the efficiency of spatial sampling designs. Finally, conclusions and suggestions for the use of spatial sampling in practice are given in Section 4.

2. Spatial sampling designs

In many environmental and health survey studies, samples characterized by some form of spatial autocorrelation in the target variable are not desirable. It is indeed renowned that dependence among the subjects can be a nuisance, especially if the focus of the analysis is the estimation and inference of a population total. The state-of-the-art of spatial sampling methods (such as the *Local Pivotal Methods* by Grafström *et al.*, 2011 and the *Spatially Correlated Poisson Sampling* by Grafström, 2012) overcomes this problem by selecting samples in which the subjects are well spread in space, so that it is likely that the sample distribution of the target variable is not spatially autocorrelated (Grafström and Tillé, 2013). Well spread, or spatially balanced, samples have proven to lead to relatively more efficient estimates (Grafström, 2012).

Sampling on spatial populations has been implemented for many years by using classic sampling schemes including random, clustered and systematic sampling. These techniques, which are applicable at multiple levels in a designed spatial hierarchy (e.g. cities, urban areas, neighborhoods, regions), do not consider the information about the locations of units in the sample selection and hence do not guarantee that samples are spatially balanced.

In recent years, due to both the availability of georeferenced microdata and the progresses in computational statistics, new sampling methods have been proposed in the literature that exploit distance between population units. Some examples are traceable in a massive literature based on the use of contiguous units (see, among others, Hedayat et al., 1988; Wright and Stufken, 2008; Mandal et al., 2009) and particularly in some complex methods, described below.

The DUST (*Dependent Areal Units Sequential Technique,* Arbia, 1993) is the first sequential technique that incorporates spatial correlation in the sample selection. The hypothesis behind DUST is to have a first observable variable X in not-overlapped population units N, which must be estimated through sampling experiments and by using auxiliary information about second-order properties of a second variable Y. The two variables could be linked with different kinds of relationships, such as Y could be X in a previous period or Y can be a proxy of a known variable X, without sampling errors. By using this procedure, sampling units have the same probability to be drawn that increases when their distance from the sampled areas increases (Arbia, 1993).

Another sampling method is the GRTS (*Generalized Random Tesselation Stratified,* Stevens and Olsen, 2004), which uses a function that maps the two-dimensional space into one-dimensional space, preserving the spatial order of units. The area under observation is divided into cells and then a mapping function runs to assign an order to the units. The sample is then selected in one dimension, using systematic πps sampling and then mapped back in two dimensions. The method could be implemented for point, linear and areal frames and it has been the main spatial method used in environmental studies for many years (Stevens and Olsen, 2004).

More recently, some new methodologies that use explicitly the distances between the point-level locations of units in the selection procedure have been proposed in the statistical literature, such as the Local Pivotal Methods (Grafström *et al.,* 2011) and the Spatially Correlated Poisson Sampling (Grafström, 2012). These methodologies are described in detail below and they will be used in the simulation study presented in Section 4.

3.1 Local Pivotal Methods

Local Pivotal Methods (LPMs, Grafström *et al*., 2011) are an extension of the Pivotal method (Deville and Tillé, 1998) to case of georeferenced microdata. The Pivotal method is a random sampling method in N steps. Its working can be briefly described as follows. At each step, the inclusion probabilities are updated for two units and the sampling outcome is decided for at least one of the two. When the updated inclusion probabilities π_i' are equals to 0 or 1, the unit i is *finished* and it may not be chosen again. The updating procedure is repeated with these updated inclusion probabilities until all units are finished (Deville and Tillé, 1998). Grafström *et al*. (2011) generalize the pivotal method to include the information about the point-level locations of units. LPMs update the inclusion probabilities according to the updating rule of Deville and Tillé (1998) but for two nearby units at each step. To choose the two nearby units i and j, the authors proposed two different methods, the LPM1, which is more balanced, and the LPM2, which is computationally simpler and fast. The LPM1 randomly chooses the first unit i and then the nearest neighbor unit j (if two or more units have the same distance to i, the method randomly choses between them). If j is not the nearest neighbor of i, the method restarts from the beginning. When all units have been visited, the method stops. To select a sample, LPM1 has an expected number of computations at most proportional to N^3. LPM2 works similarly to LPM1, but the inclusion probabilities are directly updated with the pivotal method updating rule. The expected number of computations needed to select a sample is, in this case, proportional to N^2, reducing consistently the computation time (Grafström *et al*., 2011). The Local Pivotal Methods are two algorithms easy to implement. They both produce samples spread in a space and spatially balanced.

3.2 Spatially Correlated Poisson Sampling

Grafström (2012) proposed a spatial extension of Correlated Poisson Sampling (Bondesson and Thorburn, 2008) to select spatially balanced samples. The aim of Spatially Correlated Poisson Sampling (SCPS) is to select equal or unequal inclusion probabilities samples, that are spread over a spatial population, by using a distance function $d(i,j)$ between the point-level locations of population units. The method visits all units (one by one and all at once) and decides whether one should be sampled, with the intent to create negative correlation between the inclusion indicators, so that units close in distance rarely appear simultaneously in a sample.

Within the Correlated Poisson Sampling, the sampling outcome is first decided for unit 1, then for unit 2, and so on up to unit j. If unit 1 is included with probability $\pi_1^{(0)} = \pi_1$, the method set $\mathrm{I}_1 = 1$ and otherwise $\mathrm{I}_1 = 0$. After each step, the inclusion probabilities for the remaining units in the list are updated, according to a specific rule, described as, for $\pi_j^{(0)} = \pi_j$, with $j \geq 1$, $\pi_j^{(i)} = \pi_j^{(i-1)} - \left(I_i - \pi_i^{(i-1)}\right) w_{j-1}^{(i)}$, with $j \geq i+1$, and $i = 1, 2, \ldots, n$, where $w_{j-i}^{(i)}$ are weights, depending on $I_{1,}I_2, \ldots, I_{i-1}$, but not on I_i (Bondesson and Thorburn, 2008). Gradually, the inclusion probability vector is updated in N steps, until it becomes the vector of inclusion indicators. Weights $w_j^{(i)}$depend on the previous units sampling outcomes, but not on the future outcomes. Positive weights give negative correlations between the inclusion indicators and negative weights give positive correlation. The described method uses a probability function (Grafström, 2012), which can be written as

$$\mathrm{Pr}(\mathrm{I} = \mathrm{x}) = \prod_{\mathrm{i}=1}^{\mathrm{N}} \left(\pi_{\mathrm{i}}^{(\mathrm{i}-1)}\right)^{\mathrm{x_i}} \left(1 - \pi_{\mathrm{i}}^{(\mathrm{i}-1)}\right)^{1-\mathrm{x_i}}, \quad \mathrm{x} \in \{0,1\}^{\mathrm{N}}.$$

The SCPS incorporates a known distance between units and it can be definable as a set of strategies to choose weights for Correlated Poisson Sampling. Two strategies have been suggested to choose weights: maximal weights and Gaussian preliminary weights (for details, see Grafström, 2012).

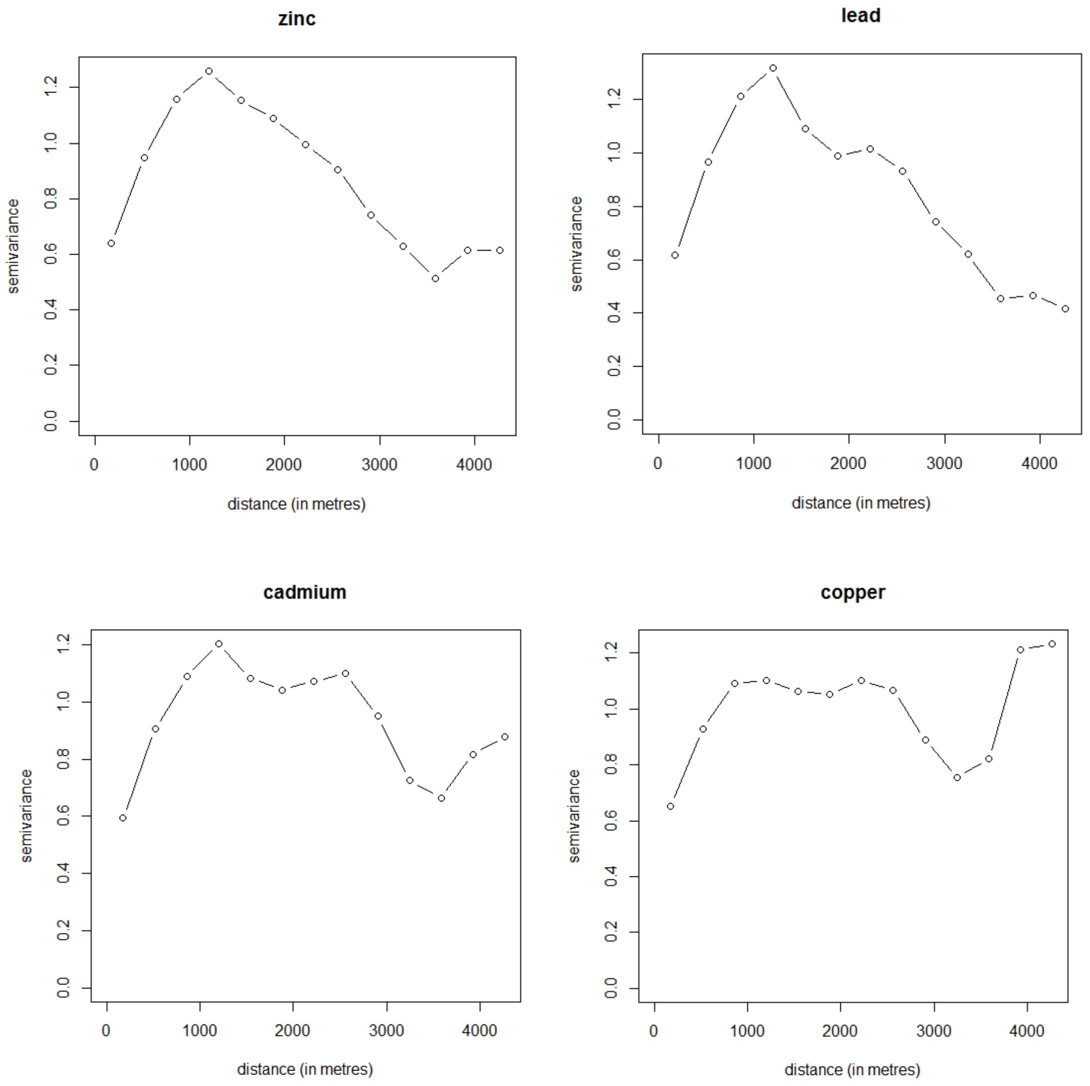
Samples obtained with SCPS are spread in the space, independently by the order in which the units appear, because it considers only the distance between them.

3. A simulation study on water quality

The ‘Meuse’ dataset is available in the R package ‘gstat’. The dataset contains information about top soil heavy metals concentrations and others soil and landscape variables. Each observation is also provided with geographical information about the precise locations of the collected sample in a flood plain area of approximately 15m x 15m of the river Meuse, near the Stein village (Pebesma, 2011). The dataset is composed by 164 units collected in the area. The goal of our study is to estimate the total values of zinc, lead, cadmium and copper in order to evaluate the pollution of water. The Meuse is one of the biggest rivers in Europe, crossing some industrialized areas of three countries and its water is used both for fishing and drinking. As it is known, water polluted by heavy metals can cause many problems for human and animal health (Alberign *et al.*, 1999; Schilderman *et al.*, 1999). Here we consider the 164 samples as the reference population to conduct our study, as seen in Grafström and Tillé (2013).

First of all, it has been verified whether the variables of interest are characterized by autocorrelation by means of the empirical semivariogram. For a target variable y, the empirical semivariogram shows the quantity $v(d_{ij}) = \frac{1}{2}(y_i - y_j)^2$ on the y-axis, where d_{ij} denotes the spatial distance that separates the units i and j. A variable y is characterized by spatial autocorrelation if $v(d_{ij})$ shows some evident peaks. Figure 1 shows the empirical semivariograms for the four top soil heavy metals concentrations obtained by averaging the $v(d_{ij})$ values within distance bands.

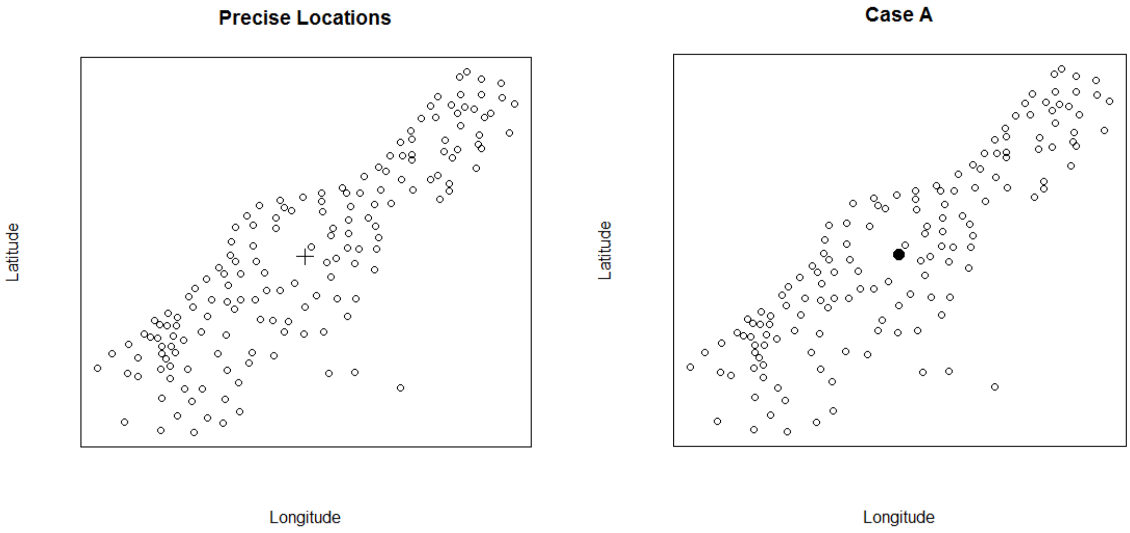
Figure 1. Empirical semivariograms of interesting variables (zinc, lead, cadmium and copper).



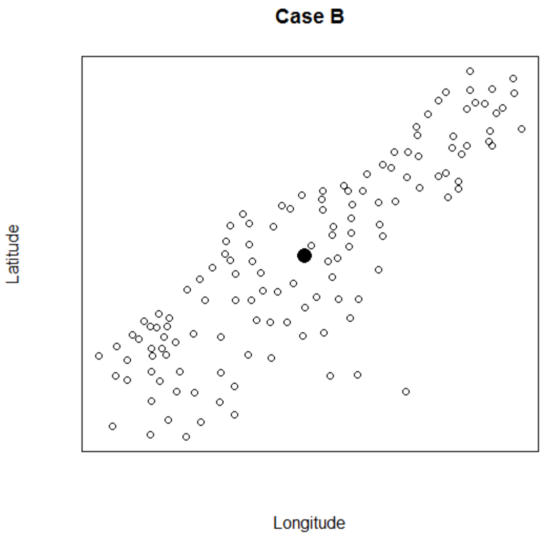
As can be noted, $v(d_{ij})$ has relevant picks for all four considered variables between 1000 and 2000 metres. This evidence indicates that exists a small-scale spatial trend amongst units for the heavy metals under study. Therefore, in this context, spatial sampling designs may have a powerful role in the estimation of population totals because they avoid the selection of close units with similar information content by spreading samples in the space.

The present study focuses on the consequences on the estimation of the total value of a variable when the population is a georeferenced microdata with coarsened locations. First, we use the dataset provided with geographical locations of units, without any perturbation in their positions. Then, after computing the centroid of the region under study, as identified with a cross in Figure 2-a, we proceed in perturbing three proportions of locations, such as 10% (Case A, Figure 2-b), 20% (Case B, Figure 2-c) and 30% (Case C, Figure 2-d). The proportions of units incorrectly positioned on the territory have been selected randomly over the population and they have been positioned on the centroid of the area under study, in order to reproduce real situations in which is not possible to locate all units correctly.

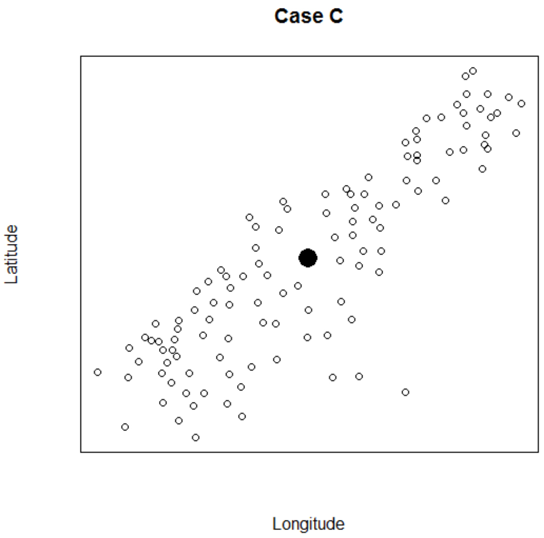
Figure 2. Population of units for the Meuse dataset and three proportions of incorrect locations.



(a) **(b)**



(c)



(d)

The simulation study consists on selecting samples of 50 units, with equal inclusion probabilities, by using simple random sampling without replacement (SRSWOR) and LPM1, LPM2 and SCPS. We compute the Horvitz-Thompson estimator for the total (Horvitz and Thompson, 1952) by means of 10,000 Monte Carlo simulations (Robert and Casella, 2004) to estimate the population totals of heavy metals concentrations. The Horvitz-Thompson estimator has the form

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

where y_i is the value of a target variable in the population (in the present case topsoil metals concentrations) and π_i is the value of the inclusion probability of sample unit i.

The results of the simulation study are shown in Table 1. We compute, as a measure of efficiency, the relative Root Mean Square Error (rRMSE) for all sampling methods and for all the different scenarios. The rRMSE has been estimated as

$$rRMSE = \frac{\sqrt{\sum_{nsim}\left(\hat{Y}_i - Y_i\right)^2 / nsim}}{Y_i}$$

where Y_i represents the total of the target variable in the population and $nsim$ represents the number of Monte Carlo simulations.

Table 1. Relative RMSE results. Sample size equal to 50 units.

		Relative Root Mean Square Error			
Design		**zinc**	**lead**	**cadmium**	**copper**
SRSWOR	No Locations	0.0958	0.0884	0.1043	0.0710
LPM1	Precise Locations	0.0747	0.0655	0.0851	0.0523
LPM2	Precise Locations	0.0720	0.0633	0.0857	0.0518
SCPS	Precise Locations	0.0680	0.0608	0.0795	0.0487
LPM1	Case A	0.0745	0.0658	0.0863	0.0524
LPM2	Case A	0.0740	0.0643	0.0870	0.0521
SCPS	Case A	0.0699	0.0614	0.0816	0.0504
LPM1	Case B	0.0751	0.0676	0.0857	0.0539
LPM2	Case B	0.0768	0.0686	0.0879	0.0559
SCPS	Case B	0.0726	0.0660	0.0833	0.0530
LPM1	Case C	0.0875	0.0777	0.0921	0.0644
LPM2	Case C	0.0855	0.0754	0.0916	0.0627
SCPS	Case C	0.0832	0.0743	0.0884	0.0616

As can be noted, the values of rRMSE gradually increase when the proportion of coarsened locations increases, but still remain lower than the values for SRSWOR. Due to these considerations, we have made a comparison between spatial designs and SRSWOR, firstly to evaluate if and how much it is convenient the use of spatial sampling in case of accurate locations, and then to assess if it remains convenient also with increasing proportions of inaccurate locations.

The gain in efficiency could be expressed as

$$\left(1-\left(\frac{rRMSE_{spatial\ design}}{rRMSE_{SRSWOR}}\right)\right)100(\%)$$

which represents the percentage improvement in terms of efficiency, with respect to simple random sampling without replacement. Results are shown in Table 2.

Table 2. Gain in efficiency of spatial sampling designs respect to SRSWOR.

Design		**zinc**	**lead**	**cadmium**	**copper**
LPM1	Precise Locations	22.05%	25.88%	18.40%	26.22%
LPM2	Precise Locations	24.85%	28.33%	17.85%	27.06%
SCPS	Precise Locations	29.05%	31.18%	23.78%	31.39%
LPM1	Case A	22.23%	25.57%	17.32%	26.17%
LPM2	Case A	22.75%	27.19%	16.63%	26.60%
SCPS	Case A	27.06%	30.49%	21.74%	29.01%
LPM1	Case B	21.62%	23.47%	17.84%	24.08%
LPM2	Case B	19.88%	22.35%	15.75%	21.22%
SCPS	Case B	24.22%	25.33%	20.11%	25.26%
LPM1	Case C	8.68%	12.11%	11.68%	9.23%
LPM2	Case C	10.76%	14.72%	12.20%	11.58%
SCPS	Case C	13.11%	15.88%	15.25%	13.20%

The results showed in Table 2 highlight that in all cases the use of spatial sampling designs is more efficient than simple random sampling. This evidence confirms the notion that when information about the point-level spatial locations of subjects is available, it is always convenient and efficient to consider it the in sampling selection procedure (Dickson *et al.*, 2014). SCPS algorithm gives the best results for these data, both when locations are exact and coarsened. As expected, when the proportion of incorrect locations increases, the efficiency in total estimation decreases. However, all results are highly efficient with respect to SRSWOR.

Moreover, in case of equal inclusion probability, it is possible to evaluate the representativeness index of a design, by computing the spatial balance of samples drawn (Grafström and Schelin, 2013).

The spatial balance could be computed following the Voronoi polygons approach (Stevens and Olsen, 2004). If i is a unit of a sample s, a Voronoi polygon contains all units close in distance to i. A unit can be included only in one polygon, so that if it has equal distance to more units of a sample, it could be located in more than one polygon and its inclusion probability will be divided between two polygons. Let v_i the sum of the inclusion probabilities of units located in the $i-th$ Voronoi polygon, then $E(v_i) = 1$ and $\sum_{i \in s} v_i = \sum_{j \in U} \pi_j = n$. A spatially balanced sample is obtained when all $v_i s$ are close to 1. Then it is possible to use the variance

$$VarSB = \frac{1}{n} \sum_{i \in s} (v_i - 1)^2$$

as a spatial balance measure. For samples drawn in the context of precise locations and for SRSWOR, the spatial balance index has been computed in order to assess, over the 10,000 simulations, the representativeness index of the sampling methods used. The results are shown in Table 3.

Table 3. Spatial balance of the drawn samples. Lower is the value, higher is the spatial balance.

		Spatial balance index			
Design		**zinc**	**lead**	**cadmium**	**copper**
SRSWOR	No Locations	0.3287366	0.3277499	0.3278469	0.3277499
LPM1	Precise Locations	0.1308928	0.1308611	0.1310522	0.1308611
LPM2	Precise Locations	0.1355972	0.1349901	0.1353536	0.1349901
SCPS	Precise Locations	0.140915	0.1410384	0.1410912	0.1410384

As expected, the SRSWOR leads to a poor spatial balance because it does not consider the geographic locations of subjects. Instead, all the others considered designs have a spatial balance index that is very low, indicating a good level of spatial balance. The trend of values is the same for the four analyzed heavy metals and they do not show significant differences between them. Following Grafström and Schelin (2013), with a high level of spatial balance, it is possible to establish the existence of a high level of sample representativeness. So that, from the results obtained in the present study it is possible to argue that spatial sampling is more efficient and more representative with respect to non-spatial sampling methodologies.

4. Conclusion

On one hand, the increasing availability of georeferenced microdata has opened the opportunity to conduct environmental and health survey studies using spatial sampling designs that exploit the exact point-level geographical coordinates of all subjects in order to select samples that are spatially balanced and hence lead to more efficient estimates. On the other hand, the state-of-the-art of geocoding procedures and technologies have limited this opportunity because the locations of subjects are rarely exact in practice and the georeferenced microdata researchers have to typically deal with have a non-negligible proportion of coarsened locations.

By means of a simulation study, based on real environmental data, this article has shown that the use of spatial sampling designs is still preferable to the use of non-spatial methods even though the georeferenced microdata are characterized by a high proportion of coarsened locations. The evaluation has been conducted by a comparison in terms of relative RMSE and spatial balance. Samples drawn with spatial designs showed an important reduction of rRMSE and an increase in spatial balance, leading to the conclusion that spatial sampling is more efficient than classic methods, also when not all the locations of population units are exact.

References

Albering, H.J., van Leusen, S.M., Moonen, E.J., Hoogewerff, J.A., and Kleinjans, J.C. (1999) Human health risk assessment: A case study involving heavy metal soil contamination after the flooding of the river Meuse during the winter of 1993-1994, *Environmental Health Perspective*, **107**, 37-43.

Allshouse, W.B., Fitch, M.K., Hampton, K.H., Gesink, D.C., Doherty, I.A., Leone, P.A., Serre, M.L. and Miller, W.C. (2010) Geomasking sensitive health data and privacy protection: An evaluation using an E911 database, *Geocarto International*, **25**, 443-452.

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*, New York: Springer.

Anselin, L. (2010) Thirty years of spatial econometrics, *Papers in Regional Science*, **89**, 3-25.

Arbia, G. (1993) The use of GIS in spatial statistical surveys, *International Statistical review*, **61**, 339-359.

Bondesson, L. and Thorburn, D. (2008) A List Sequential Sampling Method Suitable for Real-Time Sampling, *Scandinavian Journal of Statistics*, **35**, 466-483.

Boulos, M. N. K. (2004) Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom, *International Journal of Health Geographics*, 3(1):1.

Burrough, P.A. (2001) GIS and geostatistics: essential partners for spatial analysis, *Environmental and Ecological Statistics*, **8**, 361-377.

Cozzi, S. and Filipponi, D. (2012) *The new geospatial Business Register of Local Units: potentiality and application areas*, 23rd Meeting of the Wiesbaden Group on Business Registers - International Roundtable on Business Survey Frames, Washington, D.C.: 17 – 20 September 2012.

Curtis, A.J., Mills, J.W. and Leitnern, M. (2006) Spatial confidentiality and GIS: re-engineering mortality locations from published maps about hurricane Katrina, *International Journal of Health Geographics*, **5**, 44-56.

Dale M.R.T. and Fortin, M.-J. (2014) Spatial Analysis: a guide for ecologists, 2nd Edition. Cambridge: Cambridge University Press.

Deville, J.-C. and Tillé, Y. (1998) Unequal Probability Sampling Without Replacement Through a Splitting Method, *Biometrika*, **85**, 89-101.

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

Dickson, M.M., Benedetti, R., Giuliani, D. and Espa, G. (2014) The use of Spatial Sampling Designs in Business Surveys, *Open Journal of Statistics*, **4**, 345-354.

Grafström, A., Lundström, N.L.P. and Schelin, L. (2011) Spatially Balanced Sampling through the Pivothal Method, *Biometrics*, **68**, 514-520.

Grafström, A. (2012) Spatially correlated Poisson sampling, *Journal of Statistical Planning and Inference*, **142**, 139-147.

Grafström, A., and Schelin, L. (2013) How to select representative samples, *Scandinavian Journal of Statistics*, **41**, 277-290.

Grafstöm, A. and Tillé, Y. (2013) Doubly spatial sampling with spreading and restitution of auxiliary totals, *Environmetics*, **24**, 120-131.

Hedayat, A.S., Rao, C.R. and Stufken, J., (1988) Sampling plans excluding contiguous units, *Journal of Statistical Planning and Inference*, **19**, 159-170.

Horvitz, D.G. and Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association,* **47**, 663-685.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordholt, E.S., Seri, G. and De Wolf, P.-P. (2010) *Handbook on statistical disclosure control*, The Hague: ESSNet SDC, A Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control, Netherlands.

Jacquez, M.G. (2012) A research agenda: Does geocoding positional error matter in health GIS studies?, *Spatial and Spatio-temporal Epidemiology*, **3**, 7-16.

Mandal, B.N., Parsad, R., Gupta, V.K. and Sud, U.C., (2009) A family of distance balanced sampling plans, *Journal of Statistical Planning and Inference*, **139**, 860 – 874.

McRoberts, R.E., Holden, G.R., Nelson, M.D., Liknes, G.C., Moser, W.K. and Lister, A.J. (2005) Estimating and circumventing the effects of perturbing and swapping inventory plot locations, *Journal of Forestry*, **103**, 275-279.

Pebesma E. (2011) Reference manual for R-package ‘gstat’. Available from: http://cran.r-project.org/web/packages/gstat/gstat.pdf [Accessed on 26/12/2015].

Robert, C. and Casella, G. (2004) *Monte Carlo Statistical Methods*, New York: Springer.

Rushton, G., Armstrong, M., Gittler, J., Greene, B., Pavlik, C., West, M. and Zimmerman, D.L. (2006) Geocoding in cancer research—a review, *American Journal of Preventive Medicine*, **30**, S16-S24.

Schilderman, P.A.E.L., Moonen, E.J.C., Maas, L.M., Welle, I. and Kleinjans, J.C.S. (1999) Use of Crayfish in Biomonitoring Studies of Environmental Pollution of the River

Meuse, *Ecotoxicology and Environmental Safety*, **44**, 241-252.

Stevens, D.L.Jr. and Olsen, A.R. (2004) Spatially Balanced sampling of Natural Resources, *Journal of the American Statistical Association*, **99**, 262-278.

VanWey, L.K., Rindfuss, R.R., Gutmann, M.P., Entwisle, B. and Balk., D.L. (2005) Confidentiality and spatially explicit data: Concerns and challenges, *Proceedings of the National Academy of Science of the United States of America*, **102**, 15337-15342.

Wright, J.H. and Stufken, J. (2008) New balanced sampling plans excluding adjacent units, *Journal of Statistical Planning and Inference*, **138**, 3326 – 3335.

Yang D., Bilaver, L.M., Hayes, O. and Goerge, R. (2004) Improving geocoding practices: evaluation of geocoding tools, *Journal of Medical Systems*, **28**, 361-370.

Zandbergen P. (2008) A comparison of address point, parcel and street geocoding techniques, *Computers, Environment and Urban Systems*, **32**, 214–232.

Zimmerman, D.L. (2008) Estimating the Intensity of a Spatial Point Process from Locations Coarsened by Incomplete Geocoding, *Biometrics*, **64**, 262-270.

Zimmerman, D.L. and Li, J. (2010) The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies, *International Journal of Health Geographics*, 9:10.