METRON (2017) 75:249–264
DOI 10.1007/s40300-017-0123-1

Alternative and complementary approaches to spatially balanced samples

R. Benedetti[1] · F. Piersimoni[2] · P. Postiglione[1]

Received: 1 June 2017 / Accepted: 11 September 2017 / Published online: 18 September 2017

Abstract The spatial distribution of a population represents an important tool in sampling designs that use the geographical coordinates of the units in the frame as auxiliary information. These data may represent a source of auxiliaries that can be helpful to design effective sampling strategies, which, assuming that the observed phenomenon is related with the spatial features of the population, could gather a considerable gain in their efficiency by a proper use of this particular information. We present and compare various methods to select *spatially balanced* samples. These selection algorithms are compared with the intuitive principle of partitioning the space into n strata and selecting only one unit per stratum. The fundamental interest is not only to evaluate the effectiveness of such different approaches, but also to understand if it is possible to combine them to obtain more efficient sampling designs. The performances of the *spatially balanced* designs are compared in terms of their root mean squared error using the simple random sampling without replacement as benchmark. An important result is that these complex designs provide better results than the simple principle of stratifying the study area. It also does not help so much to improve efficiencies even if it is combined with balancing on known totals of some auxiliary variables, such as the geographic coordinates.

Keywords Spatial stratification · Generalized random tessellation stratified design · Spatial dependence · Traveling salesman problem · Spatial indexing

✉ R. Benedetti
benedett@unich.it

F. Piersimoni
piersimo@istat.it

P. Postiglione
postigli@unich.it

[1] Department of Economic Studies (DEc), “G. d’Annunzio” University, Viale Pindaro 42, 65127 Pescara, Italy

[2] Istat, Directorate for Methodology and Statistical Process Design, Via Cesare Balbo 16, 00184 Rome, Italy

Mathematics Subject Classification 62D05 · 62H11

1 Introduction

Maps have become a very important tool on which a large amount of auxiliary variables are available. The importance of selecting samples of statistical units taking into account their geographical position is now more than ever recognized in the measuring process of several phenomena [7,26]. Nowadays, it is a common practice that national statistical offices geo-reference their sampling frames of physical or administrative bodies, used for social and economic surveys, not only according to the codes of a geographical nomenclature, but also adding information regarding the exact, or estimated, position of each record. From 2000 to the present, there has been a steady increase in the use of area frame surveys [7], which often supported, or even replaced, the traditional methods based on list frames mainly due to the absence of coverage errors and the lower probability of non-response that they guarantee.

Moreover, the increase in the statistical use of administrative data has led to the need of estimating the coverage of administrative lists, which is often carried out through area frame surveys. Spatial surveys also present some drawbacks with regard to the reduced information that can be collected in a survey involving only a direct observation and not a questionnaire to be filled. This problem has been a key point in the increase of the use of a dual-frame approach when conducting area frame surveys [25,28]. Within this context, [4] point out that area frame surveys will play a very important role in the future development of agricultural surveys in particular in developing countries.

The aim of this paper is to discuss the performances of different *spatially balanced* designs. These are defined as probability samples that are well spread in every dimension [30]. These designs are also compared with the simple criterion of stratifying the study area. Finally, we introduce a complementary framework that produces a design defined through the combination of the two previous approaches: spatial balancing and stratification.

A first attempt to define *spatially balanced* samples was introduced by [23] as a criterion to increase the amount of information on the population by avoiding the selection of pairs of contiguous units. Since then, *spatially balanced* samples have experienced a noticeable development and several designs for finite and continuous populations have been proposed [6,8,14,15,18,20,21,30]. In recent years, several works that focus on designs that explicitly consider the geographical distances of the units to better spread the sample in space have arisen (see [8] for a review). Often spatial units are also artificially defined and made available over a domain partitioned into a number of predetermined regularly or irregularly shaped sets of spatial objects. This may happen, for example, when the original data generating process lies over a continuous spatial domain and, to simplify the problem, the researcher chooses to observe it only in a selection, possibly made at random, of fixed points or averaged over a selection of predefined polygons. Even if in this paper we will not analyze situations of this kind, they cover an important part of the possible sampling problems. There is a huge list of phenomena which can be observed in any site of a linear object, such as, for example, a river, or of a surface as it is for meteorological data. In these cases the resulting sample is a set of points or polygons whose possible positions are not predefined but chosen from an infinite set of possible sites.

In this paper, we only focus on finite populations. The reason of this choice is motivated by the fact that the spatial distribution of the frame is a strong constraint which we suspect that could have a considerable impact on the performance of a random sampling method. For example, the traditional solution of extending the systematic sampling to multidimensional data by overlaying a grid of points to a spatial domain could not be feasible if the population is far to be considered distributed on a regular grid as it is clustered or it shows to have different intensities of the units across the domain. The units to be observed should be randomly selected using the spatial distribution of these finite populations that represents an important information in the sampling strategy.

Spatially balanced samples proved to be so efficient that several selection algorithms of this kind were suggested by researchers and survey practitioners. Given the particular nature of the spatial information, its efficient use in sampling often requires methods that can not be adapted from those used when dealing with classical auxiliary variables. To use some covariates in sample design, we always assume that there is, at least approximately, a certain degree of dependence between a survey variable y and the set of auxiliaries $\mathbf{X}$. When these covariates consist of a set of geographic coordinates, i.e. when $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}\}$ where $\mathbf{x_1}$ and $\mathbf{x_2}$ are respectively the east–west and north–south coordinates, the distance matrix might be used to evaluate the assumed similarity between adjacent units, and, therefore, to emphasize the importance of the spread of the sample over the study region. The similarity of information provided by nearby units is shortly expressed in Tobler's [31] first law of geography, according to which "everything is related to everything else, but near things are more related than distant things". As according to a design-based approach, that is the framework used in this paper, y is considered deterministic and not as a realization of a stochastic process. To introduce the concept of dependence within this framework, we necessarily resort to the use of the anticipated variance (AV) [21]. In the design phase we assume that, for each unit $i = 1, \ldots, N$ of a finite population U, a linear model holds for each y_i, given the known auxiliaries $\mathbf{x}_i$:

$$\begin{cases} y_i = \mathbf{x}_i^t \beta + \epsilon_i \\ E_\xi\left(\epsilon_i\right) = 0 \\ Var_\xi\left(\epsilon_i\right) = \sigma_i^2 \\ Cov_\xi\left(\epsilon_i \epsilon_j\right) = \sigma_i \sigma_j \rho_{ij} \end{cases}, \tag{1}$$

where E_ξ, Var_ξ, and Cov_ξ denote respectively expectation, variance and covariance with respect to the model ξ, β is a vector of regression coefficients, ϵ_i is a random variable with variance σ_i^2 and ρ_{ij} is its autocorrelation coefficient. The anticipated variance of the Horvitz–Thompson (HT) estimator of the total of y, under the working model (1), is:

$$AV\left(\hat{t}_{y,HT} - t\right) = E_s\left[\left(\sum_{i \in s} \frac{\mathbf{x}_i}{\pi_i} - \sum_{i \in U} \mathbf{x}_i\right)^T \beta\right] + \sum_{i \in U}\sum_{j \in U} \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}, \tag{2}$$

where π_i and π_{ij} are, respectively, the first and second order inclusion probabilities, and E_s denote expectation with respect to the sample design. From (2), it is clear that uncertainty on estimates can be divided into two components: the error on the auxiliary variables and the dependence of observed units. The first can be reduced, if not eliminated, by constraining the units selected with respect to the average value of the population's coordinates, while the second, assuming that ρ_{ij} decreases as the distance d_{ij} between the selected units i and j increases, leads us to select units as far apart as possible. Under a model-based framework, no randomization assumption is required. Thus, the concern is in finding an optimal sample configuration, using some combinatorial optimization algorithms, which is the representative of the whole population [5]. However within a design-based framework, where the randomization assumption is crucial, an optimal sample selected with certainty is of course not acceptable, and we have to make use of alternative solutions that allow us to select more distant units with higher probability. These spatial aspects have to be included in the design according to an efficient strategy that should help also to understand how they can be exploited to increase the efficiency of the estimators. A stratification obtained clustering the coordinates is a simple and widely used solution suggested for practical reasons that, in our opinion, makes a partial use of these features.

The concept of *spatial balance* is indeed an interesting and promising framework to be followed for setting up a design to select units from a spatial population. Its flexibility is very important to combine the spatial distribution of the sample with another framework derived from the introduction of constraints on the known totals of some auxiliary variables, as geographical coordinates. The number of *spatially balanced* designs thus increases for the introduction of the whole set of possible combinations of *balancing* in space together with the respect of a defined stratification and with the assumption of linear or non–linear relationships with the covariates. The performance of these possible solutions is related to several population characteristics. The aim is to better understand, in order to increase the design efficiency, which features are really important and in which practical situation.

This article proceeds as follows. Section 2 introduces some sample selection algorithms that, spreading the sample over the study region, seek to exploit the spatial characteristics of the population. Section 3 examines the performance of the suggested design when compared with other sampling designs that are evaluated in terms of the root mean squared errors (RMSE) of the estimates by using the simple random sampling (SRS) as benchmark. Finally, Sect. 4 contains some concluding remarks that focus on outstanding research issues that are associated with spatial sampling designs.

2 A taxonomy of spatially balanced samples designs

It can be seen from (2) that a gain in the efficiency of the HT estimator can be realized both by respecting, if known, the average values of some covariates, as the geographic coordinates in the case of the presence of a linear spatial trend, and/or by defining a design in which the second order inclusion probabilities are higher for any couple i, j that have a high distance $d_{i,j}$, expecting that the correlation coefficient decreases as the distance increases. These two aspects allow us to introduce a first classification of methods to select sampling units: those dedicated to respecting the known totals of a set of auxiliary variables and those that tend to spread units over space. The first category only comprises *balanced sampling*, i.e. the CUBE algorithm, proposed as a general (not spatial) solution [11,13] to the restriction of the support S of samples which can be selected by imposing a set of linear constraints on the covariates. These restrictions represent the intuitive requirement that the sample estimates of the total, or of the average, of a covariate should be equal to the known parameter of the population. In a spatial context, this plan could be applied by imposing that, for any selected sample, the first p moments of each coordinate should coincide with the first p moments of the population, assuming implicitly that the survey variable y follows a polynomial spatial trend of order p. This logic was subsequently extended to approximate any nonlinear trends through penalized splines with particular reference to the space [10].

The second category is instead more articulated and composed by algorithms of different nature. To better understand the characteristics of this group, it is appropriate the use of Voronoi polygons that are employed to define an index that will be the basis for the definition of *spatially balanced* samples. Let us denote by S the set of all the possible random samples of fixed size n which can be selected from U, where its generic element is $s = \{s_1, \dots, s_i, \dots, s_N\}$ and s_i is equal to 1 if the unit with label i is in the sample, and 0 otherwise. For a generic s, the Voronoi polygon associated to the sample unit s_i=1 includes all population units closer to i than to any other sample unit j. If we let ν_i be the sum of the inclusion probabilities of all units in the ith Voronoi polygon, for any sample unit we have $E(\nu_i) = 1$. Thus the *spatial balance index* (SBI):

$$SBI = Var\,(\nu_i) = \frac{\sum_{i=1}^{n} (\nu_i - 1)^2}{n} \tag{3}$$

can be used as a measure of difference from the state of perfect *spatial balance*, as for a *spatially balanced* sample all the ν_i should be close to 1. Note that, despite the obvious assonance of the two terms, this notion is quite far from that of *balanced sampling* introduced to define the first category through the CUBE method [8]. A selection strategy conceived with the clear goal of minimizing (3) should use the concept of distance that is a basic tool to describe the spatial distribution of the sample units and leads to the intuitive criterion that units that are close should seldom appear simultaneously in the sample. This condition can be considered as reasonable under the assumption that, increasing the distance between two units i and j, always increases the difference $\left|y_i - y_j\right|$ between the values of the survey variable. In such a situation, it is clear that the variance of HT estimator will necessarily decrease if we set high joint inclusion probabilities to couples with very different y values as they are far from each other. To understand when and how it could be an efficient strategy to spread in some way the selected units over the population, we need to suppose that the distance matrix summarizes all the features of the spatial distribution of the population and, as a consequence, of the sample. This general hypothesis within a model based perspective is equivalent to assuming that the data generating process is stationary and isotropic, i.e. its distribution does not change if we shift or rotate the space of the coordinates. Focusing on the set of coordinates without using any other information coming from other covariates, this assumption implies that the problem of selecting *spatially balanced* samples is to define a design $p(S)$ with probability proportional to some synthetic index of the within sample distance matrix when it is observed within each possible sample s (e.g. see [6]).

Following an approach based on distances, inspired by purely model-based assumptions on the dependence of the stochastic process generating the data, [1] suggested a draw-by-draw scheme: the dependent areal units sequential technique (DUST). Starting with a unit selected at random, say i, in any step $t < n$, the selection probabilities are updated according to a multiplicative rule depending on a tuning parameter useful to control the distribution of the sample over the study region. This algorithm, or at least the design that it implies, can be easily interpreted and analyzed in a design–based perspective in particular referring to a careful estimation and analysis of its first and second order inclusion probabilities.

Another solution that does not specify the probability of the entire sample but that is based on a more classic list sequential algorithm was suggested by [18]. Introduced as a variant of the correlated Poisson sampling, in the SCPS (spatially correlated Poisson sampling) for each unit, in any step, it updates the inclusion probabilities according to a rule in such a way that the required inclusion probabilities are respected. The suggested *maximal weights* criterion, used to update the inclusion probabilities in each step, provides as much weight as possible to the closest unit, then to the second closest unit and so on. A procedure to select samples with fixed first order inclusion probabilities and correlated second order inclusion probabilities was derived in [20] as an extension of the pivotal method initially introduced to select πps samples [12]. It is essentially based on an updating rule of the probabilities that at each step should locally keep the sum of the updated probabilities as constant as possible and differ from each other in a way to choose the two closest units. This method is referred to as the local pivotal methods (LPMs).

The most dated criterion to produce samples that are well spread over the population is based on the intuitive idea, widely used by practitioners [2,17], to stratify the units of the population on the basis of their location. Within this approach, an ultimate scheme to guarantee that the sample is well spread over the population could be that to define a maximal stratification, i.e. a partition of the study region in as many strata as possible and selecting one unit per stratum. The problems arising by adopting this strategy lie also in the evidence that it does not have a direct and substantial impact on the second order inclusion probabilities, surely not within a given stratum, and that frequently it is not clear how to obtain a good partition of the study area. However, this simple and quick scheme to guarantee that the sample is well spread over the population, is somewhat arbitrary because it highly depends on the stratification criterion which should be general and efficient.

A geographical stratification of a finite geo-coded population usually requires that the size of each stratum should be approximately the same to ensure that inclusion probabilities are constant for every unit in the population and to facilitate fieldwork. Providing a predetermined number of n clusters of fixed size equal to $n_h = N/n$ units in each stratum that need to be held constant within each cluster is a problem that is not treated in the classical solutions proposed by cluster analysis. The goal here is thus to obtain clusters of the same size while reducing the total spatial distance from the center of the cluster. A solution can be derived by introducing some changes to the well known *k-means* algorithm. A schematic description of the method could follow the following steps [16,34,35]:

1. set equal cluster size $n_h = N/n, \forall h$, or assign more generally a set of fixed sizes;
2. randomly assign each unit of the population to the n groups;
3. calculate the center of each cluster;
4. select the first observation and assign it to the closest cluster;
5. since the two groups now have different sizes, in particular $n_h + 1$ and $n_h - 1$, thus we have to define [10] an exchange strategy to match the sizes. The closest observation is moved from the cluster with size $n_h + 1$ to the cluster with size $n_h - 1$;
6. this process is applied to every unit i of the population;
7. the sum of the distance from each observation to its assigned centroid is calculated;
8. if in the next iteration the distance does not decrease (within a tolerance threshold) then *stop*;
9. continue the process from *step 3* until the maximum number of iterations is reached.

A third category of selection procedures is borrowed from the spatial database environment. The basic idea is to map the two-dimensional population into a one dimension index, while trying to preserve some multidimensional features of the population, and then use this induced ordering to systematically select the sample units. The basic principle is to extend the use of systematic sampling to two or more dimensions even when the population does not lie on a grid. This idea is behind the generalized random tessellation stratified (GRTS) design [30] that maps the two dimensional population into one dimension using a grid structure. The spatial index is built using a tree hierarchical structure, which means that it represents the units in the order of a tree. Therefore the index creation process decomposes the space into a k-level grid hierarchy. Each successive level further decomposes the level above it, in such a way that each upper-level cell contains a complete grid at the next level. On a given level, all the grids have the same number of cells along both axes (in GRTS it is a 2×2, but this process can also be applied using for example, 4×4 or 8×8).

Another solution belonging to this framework is based on the computation of the minimum distance between points, under the travel salesman problem (TSP) approach. Starting from any unit, this procedure forces a TSP algorithm to visit every other unit of the population by ordering the units in this way, searching for the shortest path [15]. The link that this solution has with distance based methods is that, having to minimize the sum of the distances between adjacent units in the itinerary, it is natural to expect them to be as close as possible. It is worth noticing that an ordering can also be considered as a partitioning criterion of the population into strata, rather it is easy to argue that it provides more information also allowing for the use of systematic sampling in one dimension. In Sect. 3, the TSP has been used as a partitioning criterion of the population units into strata of the same size by cutting the ordered list into parts of size $n_h = N/n$. New prospectives for the development of these sampling designs were opened when [21] realized that the criteria behind each category of algorithms did not necessarily have to be used exclusively but also in combination within multiple targeting methods. In particular, they suggested the doubly balanced spatial sampling (DBSS) which, at least in its original version, attempts to integrate the CUBE with *spatially balanced* samples to simultaneously reduce both components of (2). If it is true, however, that a stratified design can be seen as a special case of a balanced sample by the introduction, as additional constraints, of a set $\left\{\delta_{i,h}; i = 1, \ldots, N; h = 1, \ldots, n\right\}$ of indicator variables equal to 1 if the unit i belongs to the stratum h and 0 otherwise, then we could try to integrate in the same environment also the maximal stratification in n strata. The hypothesis to be verified is that if each of these aspects is based on a logic that leads to greater efficiencies then perhaps their combined use should further reduce the root mean squared errors (RMSE) of the HT estimates. Finally, the estimation of the sample variance of these complex designs is still an open problem, the interested reader may refer to [3,29].

3 Sampling designs comparison on artificial and real populations

To empirically compare the relative strengths and weaknesses of the considered designs via Monte Carlo experiments, several simulations of the suggested designs have been carried out on simulated and real populations by using the free software environment for statistical computings R [27]. In particular, we used the following R packages: `sampling` [32], `BalancedSampling` [19], `spsurvey` [24] and `TSP` [22].

Concerning the simulated example, we have considered three frames of size $N = 1000$ generated through point processes with three different levels of clustering of the units to control the distribution of the coordinates $\{x_1, x_2\}$ and with different spatial features of the response variable y. The bi-dimensional coordinates x_1 and x_2 are generated in the unit square according to a Neyman–Scott process with Cauchy cluster kernel [33], where the intensity of the cluster centers of the Poisson process was set equal to 10. The expected number of units per cluster is 100 with three different cluster kernels equal to 0.005, 0.01 and 0.03, representing respectively a highly clustered, a clustered and a sparse population of spatial units (see Fig. 1). For each frame, six possible outcomes y have been generated according to a Gaussian stochastic process with or without a spatial linear trend $x_1 + x_2 + \epsilon$, that explain approximately the 80% of the variance of the generated population variable y. For the errors ϵ, three intensities of a spatial dependence parameter $\rho = \{0.001, 0.01, 0.1\}$ have been used, representing respectively: low, medium and high dependence between units. Finally, to avoid the possible effects due to different variability, each population was finally standardized to the same mean $\mu_y = 5$ and standard deviation $\sigma_y = 1$.

To verify if the sampling rate has any effect on the efficiency, from each of the 18 simulated y populations (i.e., three spatial distributions × two trend levels × three dependence levels), 10,000 samples of size $n = \{10, 50\}$ have been selected. The comparison between different

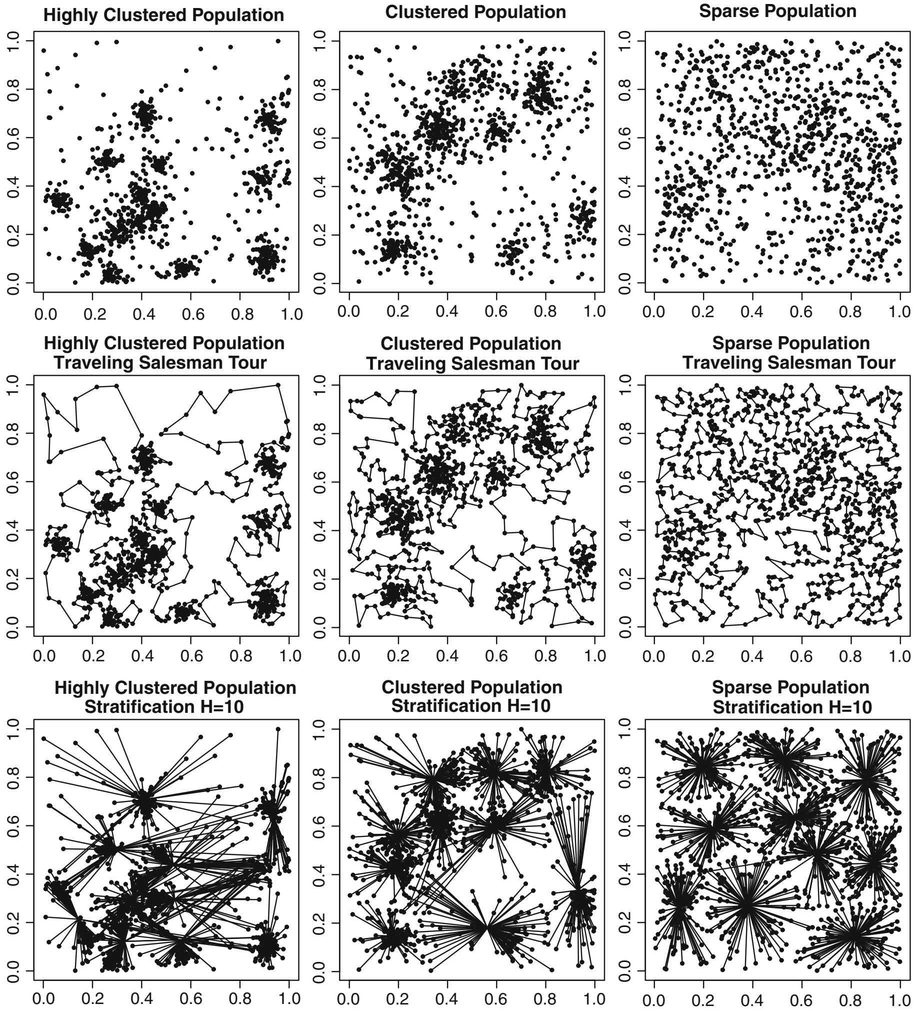


Fig. 1 Spatial distribution of the three simulated populations: highly clustered, clustered and sparse. For each population are reported the TSP tour (2nd row) and the stratification (3rd row) in $H = 10$ strata obtained by the constrained *k-means* algorithm

designs was performed by using the (RMSE) of the 10,000 HT estimates of the population mean as relative to the RMSE obtained when using a SRS design that is used, thus, as a scale factor to remove the known effects of the sizes of the population N and of the sample n on the sampling errors. It is worth noticing that in every simulation performed, as the HT is unbiased, the RMSEs were always very close to the standard error of each design as the bias can be considered negligible. As possible alternatives, we considered the GRTS [30], the balanced sampling [11,13] constrained to the average of the two coordinates (CUBE), the SCPS [18] and the LPMs (in particular, in this experiment we used the algorithm LPM1 that, from now on, will be simply denoted as LPM) [20]. The partition of each population is made according to the *k-means* (KM *spatially balanced*–KMSB), modified to generate clusters of approximately the same size [16,34,35], in which for all populations convergence is achieved in less than

20 iterations, and dividing the orders proposed by the traveling salesman problem (TSP *spatially balanced*–TSPSB) [15] in n groups. As combinations of these designs, resorting to the doubly balanced spatial sampling (DBSS) [21], we first considered the possibility of balancing on the geographical coordinates, as in the CUBE, and simultaneously spatially distribute the sample as in the LPM. Then we evaluated the hypothesis of constraining the LPM algorithm to having one unit per stratum so as to obtain *spatially balanced* and stratified samples, both according to KM and TSP (KM one per stratum, KMOS, and TSP one per stratum, TSPOS, respectively). Finally, we added to these designs the additional constraint to respect the average of the spatial coordinates to obtain stratified and doubly balanced designs (KMDBSS and TSPDBSS respectively). To ensure comparability, all these alternative designs have been used setting the first order inclusion probabilities constant and equal to n/N. It is appropriate to add a note on the use of $n = 10$, this sample size is usually not considered to be of practical interest in real surveys, but it is very useful in understanding the effect of n on the accuracy of the HT estimates deriving from the different designs involved in the experiment. Moreover, a n so low implies for stratified designs to find a limited number of groups of the same size that should facilitate the task of efficiently partitioning the population than finding a much larger number of groups.

The first empirical evidence that can be derived from Tables 1, 2 and 3 is that all the proposed algorithms in the absence of any spatial effect behave in a way comparable to the SRS, thus showing a certain robustness in the absence of the hypotheses that justify their use. On the other hand, when only one of the spatial components is present, though slightly mild, then RMSE declines very sensibly (especially in the case of the SCPS and the LPM). If the presence of a linear trend is well captured by the CUBE, it clearly fails to exploit the presence of a dependence even when it is high. Conversely, the results of the SCPS and the LPM greatly benefit from the dependence, but they also succeed in taking advantage of the trend. The spatial balance index seems to be an excellent tool for evaluating each design, assessing very clearly and comprehensively the spread of units on the study region, which implies a strong correlation of this index with the efficacy of the design. The clustering of the population has a negative effect on the efficiency of the various designs that is not reduced even by those *spatially balanced*. It represents a strong additional constraint whose effects are difficult, if not impossible, to reduce. The higher the sampling rate, the more these effects are evident, as if having more sampling units to effectively distribute on the region would help *spatially balanced* designs to achieve even greater efficiencies up to the limit of $n = N/2$ that we assume to be an upper bound beyond which this trend necessarily reverses. The GRTS seems to be a good solution, but its results are never comparable to those provided by the SCPS and the LPM, demonstrating that a simple projection of a multidimensional space into a one dimensional index is not enough to represent the relative positions of the population units which are instead better summarized in a distance matrix. The inclusion of additional constraints to respect the average values of the coordinates known in the population does not help: at least for low sampling rates, it implies an unexpected increase in the spatial balance index and therefore in the RMSE of the DBSS which is, on the other hand, the design but only when a sufficient number of sampling units are available to comply with additional constraints. As regards stratified sample designs, the TSP sample selection algorithms (TSPs) seem to ensure efficient results at least when many sample units are available and the population is not already well distributed in space, while the KM sample selection algorithms (KMs) for higher sample sizes often fail to find a good partition and consequently generates non comparable RMSEs with those provided by *spatially balanced* designs. Somehow the TSPs, though based on a one dimensional unit coordinate projection,

Table 1 Relative efficiency of the sample mean $\left(\frac{RMSE}{RMSE_{SRS}}\right)$ and average SBI for each design estimated in 10,000 replicated samples in the highly clustered population for different sample sizes, trend and dependence

Design	n	No trend			Linear trend			Average SBI
		Dependence			Dependence			
		Low	Medium	High	Low	Medium	High	
GRTS	10	1.005	0.976	0.800	0.676	0.662	0.576	0.206
CUBE	10	1.004	0.980	0.993	0.557	0.549	0.483	0.328
LPM	10	1.005	0.983	0.771	0.594	0.587	0.486	0.152
SCPS	10	1.009	0.967	0.784	0.615	0.599	0.497	0.170
DBSS	10	1.003	1.005	0.993	0.567	0.546	0.480	0.326
KMSB	10	1.012	0.987	0.836	0.656	0.651	0.598	0.208
TSPSB	10	1.016	0.972	0.747	0.593	0.567	0.451	0.191
KMOS	10	1.003	0.981	0.824	0.624	0.626	0.552	0.193
TSPOS	10	1.011	0.972	0.745	0.601	0.564	0.447	0.186
KMDBSS	10	1.002	0.985	0.825	0.612	0.610	0.542	0.191
TSPDBSS	10	1.018	0.969	0.736	0.565	0.531	0.426	0.186
GRTS	50	0.993	0.963	0.612	0.544	0.519	0.349	0.308
CUBE	50	0.997	1.007	0.990	0.508	0.502	0.426	0.555
LPM	50	1.013	0.935	0.591	0.532	0.498	0.322	0.265
SCPS	50	0.983	0.935	0.592	0.535	0.498	0.336	0.279
DBSS	50	0.989	0.938	0.550	0.513	0.476	0.299	0.250
KMSB	50	0.999	0.966	0.842	0.701	0.701	0.642	0.424
TSPSB	50	0.997	0.956	0.617	0.518	0.507	0.313	0.284
KMOS	50	0.983	0.967	0.806	0.650	0.640	0.558	0.399
TSPOS	50	0.990	0.972	0.625	0.548	0.527	0.350	0.292
KMDBSS	50	0.989	0.960	0.834	0.669	0.658	0.594	0.411
TSPDBSS	50	0.992	0.967	0.603	0.511	0.494	0.300	0.277

seems to be more effective than the GRTS, which also implies a certain degree of robustness to the sampling rate and to the population structure.

The algorithm used to determine stratification plays an essential role in selecting more or less well distributed samples and thus has a decisive impact on the efficacy of the estimates. It is worth noting that for $n = 10$, the number of partitions to be found is exactly the same as those actually present in the simulated population, an extremely favorable situation for the KMs which, being intentionally designed to find circular structures, should adequately respect the distribution of the population. Against all odds it is in clustered populations that the KMs cannot produce equally efficient samples as those of the TSPs, although major differences are seen for sample size $n = 50$.

Moreover, two well known case studies, largely debated and analyzed in the fields of spatial statistics and spatial econometrics, have been used: the Mercer–Hall and the Baltimore datasets. The first consists of 500 observations on a 20 × 25 regular grid concerning the uniformity trial of wheat in 1910, with the grain yield in pounds as the main variable (see Fig. 2). The second entails 211 observations on 17 variables regarding, among other characteristics, the sales price of the house and its coordinates in Baltimore (Maryland, USA, see Fig. 2). Both datasets are known to show a spatial trend, even if it is much more evident in Baltimore

Table 2 Relative efficiency of the sample mean $\left(\frac{RMSE}{RMSE_{SRS}}\right)$ and average SBI for each design estimated in 10,000 replicated samples in the clustered population for different sample sizes, trend and dependence

Design	n	No trend			Linear trend			Average SBI
		Dependence			Dependence			
		Low	Medium	High	Low	Medium	High	
GRTS	10	1.004	0.975	0.803	0.641	0.657	0.553	0.224
CUBE	10	1.008	0.979	0.936	0.543	0.542	0.454	0.310
LPM	10	1.005	0.990	0.750	0.579	0.587	0.470	0.178
SCPS	10	1.010	0.998	0.769	0.576	0.590	0.477	0.183
DBSS	10	1.003	0.994	0.917	0.540	0.546	0.446	0.308
KMSB	10	0.990	0.995	0.830	0.557	0.559	0.437	0.166
TSPSB	10	1.003	0.988	1.008	0.997	0.995	0.993	0.450
KMOS	10	1.005	0.985	0.819	0.561	0.569	0.447	0.169
TSPOS	10	1.006	0.978	1.000	0.993	0.994	0.985	0.448
KMDBSS	10	0.990	0.981	0.837	0.533	0.533	0.401	0.158
TSPDBSS	10	1.011	0.985	0.997	0.994	0.997	0.992	0.450
GRTS	50	0.981	0.993	0.638	0.519	0.524	0.349	0.216
CUBE	50	1.005	0.990	0.917	0.495	0.496	0.385	0.444
LPM	50	0.996	0.977	0.572	0.513	0.506	0.321	0.177
SCPS	50	0.991	0.975	0.583	0.504	0.505	0.320	0.180
DBSS	50	0.993	0.983	0.543	0.501	0.495	0.296	0.164
KMSB	50	1.003	0.983	0.711	0.554	0.558	0.421	0.247
TSPSB	50	0.998	0.969	0.601	0.501	0.506	0.305	0.204
KMOS	50	0.999	0.988	0.694	0.575	0.588	0.456	0.247
TSPOS	50	0.986	0.974	0.607	0.527	0.534	0.355	0.209
KMDBSS	50	1.000	0.994	0.700	0.534	0.546	0.402	0.239
TSPDBSS	50	0.998	0.970	0.601	0.495	0.498	0.298	0.198

data than in Mercer–Hall data. They are also available in the R packages `agridat` [36] and `spdep` [9]. In this two real populations the same set of selection methods has been used in 10,000 replicated sample of size $n = 10$ and 50.

The idea of adding the balance on the coordinates and the LPM criterion to these stratifications seems to be productive as it reduces the RMSEs even though it still fails to reach the peaks obtained by the non-stratified LPM and DBSS. This empirical evidence can be justified only by the hypothesis that a stratification in n strata, as it may seem intuitive, is not a solution to the problem of distributing the units in space, but rather represents a complex constraint that may compromise the capabilities of the *spatially* or *doubly balanced* designs.

From results of Table 4, we can conclude that the last consideration is true even when we are in the optimal situation for stratification, i.e. when dealing with a regular grid. In the Mercer–Hall dataset, the KMs have been replaced by a systematic partition of the coordinates so as to always form regular rectangles of population units in each stratum, thus avoiding the subjectivity of the partition. The results are obviously highly influenced by this particular situation, but they are still slightly less efficient than at least one design between the SCPS, the LPM or the DBSS. It is clear that for the stratified to be better than the *spatially balanced*

Table 3 Relative efficiency of the sample mean $\left(\frac{RMSE}{RMSE_{SRS}}\right)$ and average SBI for each design estimated in 10,000 replicated samples in the sparse population for different sample sizes, trend and dependence

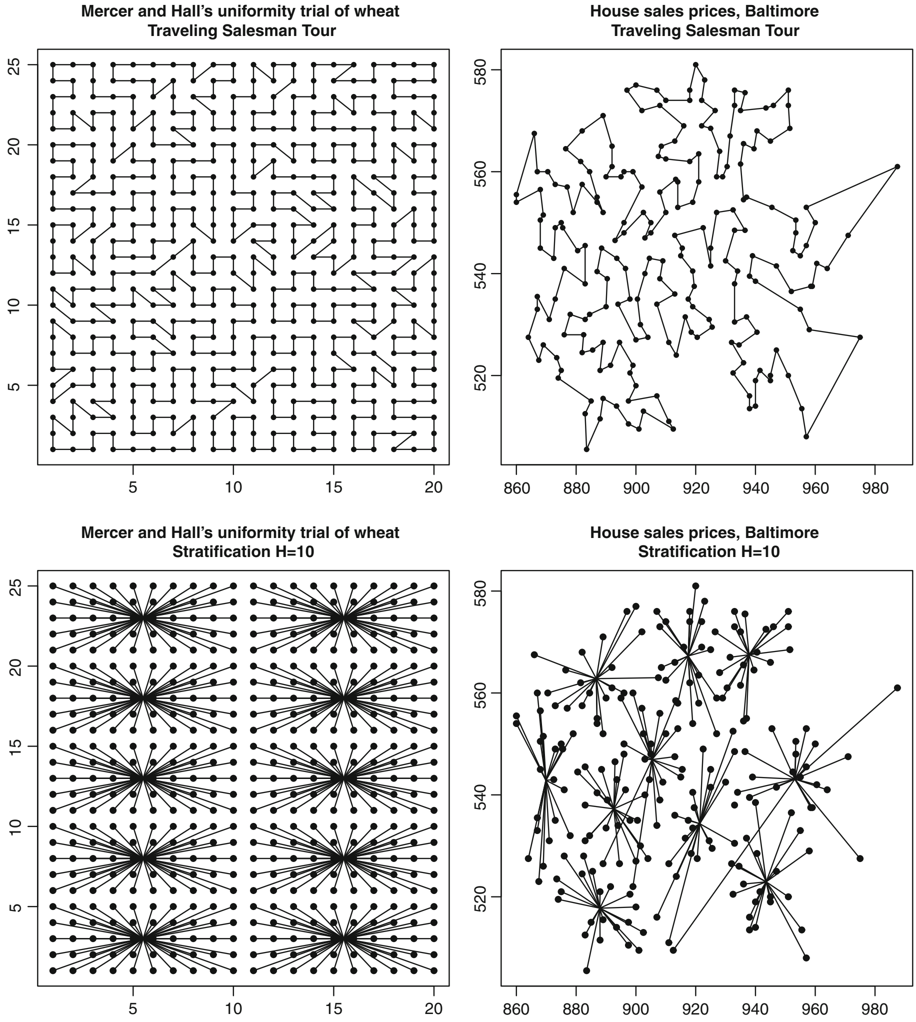
Design	n	No trend			Linear trend			Average SBI
		Dependence			Dependence			
		Low	Medium	High	Low	Medium	High	
GRTS	10	1.005	1.000	0.913	0.661	0.674	0.704	0.167
CUBE	10	1.003	0.987	0.976	0.569	0.561	0.633	0.229
LPM	10	0.995	0.985	0.893	0.614	0.624	0.629	0.121
SCPS	10	1.013	0.986	0.871	0.632	0.619	0.626	0.122
DBSS	10	1.012	0.988	0.889	0.592	0.579	0.587	0.110
KMSB	10	0.998	0.988	0.928	0.581	0.594	0.561	0.113
TSPSB	10	0.989	0.979	1.004	0.991	0.994	1.005	0.361
KMOS	10	1.005	0.988	0.921	0.598	0.615	0.580	0.118
TSPOS	10	1.007	0.999	0.999	1.006	0.998	0.999	0.364
KMDBSS	10	1.010	0.984	0.950	0.546	0.553	0.551	0.104
TSPDBSS	10	1.016	0.986	1.002	0.999	0.999	1.005	0.366
GRTS	50	0.996	0.977	0.782	0.560	0.549	0.519	0.132
CUBE	50	1.006	1.004	0.981	0.515	0.518	0.592	0.309
LPM	50	1.011	0.978	0.703	0.527	0.521	0.448	0.094
SCPS	50	1.005	0.986	0.691	0.531	0.515	0.437	0.093
DBSS	50	1.010	0.983	0.691	0.517	0.503	0.419	0.086
KMSB	50	1.004	0.992	0.764	0.543	0.531	0.476	0.119
TSPSB	50	1.005	0.987	0.746	0.525	0.514	0.465	0.115
KMOS	50	0.993	0.982	0.764	0.586	0.589	0.529	0.132
TSPOS	50	1.010	0.980	0.749	0.558	0.548	0.496	0.125
KMDBSS	50	1.010	0.976	0.758	0.535	0.516	0.462	0.115
TSPDBSS	50	1.012	0.994	0.743	0.508	0.505	0.448	0.111

samples, it is not enough for the population to lie on a regular grid, but it is also necessary that the sample size is such that it is possible to build strata with a shape that is approximately a square. In the Baltimore dataset, the strong trend favors the use of the DBSS, but again the non–stratified version seems to be preferable.

4 Conclusions

Many populations in environmental, agricultural, and forestry studies are distributed over space but it is almost clear nowadays that spatial units cannot be sampled as if they were generated within the classical urn model. This is mainly due to the impact on the sample design that a set of inherent structures which characterizes spatial data have: clustering of the coordinates, dependence, spatial trends and local stationarity. It is clear that these features may have a strong impact on the efficiency of a sample design.

The main strength of selecting samples according to some *spatially balanced* criterion lies in the ability to produce samples that are well spread over the population and that



Mercer and Hall's uniformity trial of wheat
Traveling Salesman Tour

House sales prices, Baltimore
Traveling Salesman Tour

Mercer and Hall's uniformity trial of wheat
Stratification H=10

House sales prices, Baltimore
Stratification H=10

Fig. 2 Spatial distribution of the Mercer and Hall's uniformity trial of wheat and of the house prices in Baltimore. For each population are reported the TSP tour and the stratification in $H = 10$ strata obtained by the constrained *k-means* algorithm

take advantage of the presence of any of the peculiar spatial structures that can be met in the analysis of geo-coded populations. From the results of the simulations carried out, it is clear that when we have enough ground to assume that any of these characteristic features of spatial data exists, there could be a drastic reduction of the sampling error if the suggested method is carefully employed. The question is how to incorporate these spatial aspects into the design following an efficient approach and to understand to which limits these aspects can be exploited to reduce the variance of the estimators. The common and widely used methods of spatial stratified sampling employ these features only partially. For this reason, in this article, we treat a framework for sampling from a spatial population that is based on the concept of spatial balance which could give some explanation to these questions and therefore propose a broad and flexible class of sampling designs which may

Table 4 Relative efficiency of the sample mean $\left(\frac{RMSE}{RMSE_{SRS}}\right)$ and average SBI for each design estimated in 10,000 replicated samples in the Mercer and Hall (variable *wheat*) and in the Baltimore dataset (variable *house prices*) for different sample sizes

Design	n	Mercer–Hall		Baltimore	
		Wheat	Average SBI	House prices	Average SBI
GRTS	10	0.927	0.142	0.860	0.173
CUBE	10	0.963	0.199	0.886	0.221
LPM	10	0.919	0.101	0.823	0.126
SCPS	10	0.923	0.086	0.817	0.112
DBSS	10	0.919	0.088	0.779	0.114
KMSB	10	0.910	0.093	0.833	0.119
TSPSB	10	0.942	0.111	0.852	0.117
KMOS	10	0.906	0.092	0.829	0.118
TSPOS	10	0.921	0.108	0.832	0.118
KMDBSS	10	0.914	0.082	0.807	0.108
TSPDBSS	10	0.926	0.094	0.793	0.107
GRTS	50	0.875	0.111	0.763	0.179
CUBE	50	0.955	0.278	0.862	0.313
LPM	50	0.862	0.085	0.706	0.148
SCPS	50	0.846	0.074	0.656	0.151
DBSS	50	0.847	0.073	0.679	0.145
KMSB	50	0.885	0.096	0.696	0.186
TSPSB	50	0.878	0.103	0.672	0.155
KMOS	50	0.874	0.109	0.751	0.197
TSPOS	50	0.879	0.112	0.712	0.170
KMDBSS	50	0.876	0.091	0.674	0.179
TSPDBSS	50	0.888	0.097	0.652	0.151

differ from each other by combining the spatial balance with balancing on the known totals of some auxiliary variables, as the geographic coordinates, or with the stratification itself. Particularly, a linear relationship between the coordinates used as covariates and the study variable y proved to be a valuable attribute to be exploited by *spatially balanced* designs even if significant RMSE reduction have been also found in presence of dependence of the data with closer units and clustering of the coordinates. Although they seem to be very sensible to the occurrence of any of these properties, they are also quite robust to their absence in the data as, at most, they have a variance similar to that obtained by using the SRS. The performance of several *spatially balanced* designs are stunning under a variety of population characteristics. Our main findings show that it is important, to increase efficiency, to include into the design auxiliary information, not only arising from some covariates but also from the spatial distribution of the units. The gain in efficiency depends on the strength of the relationship of the auxiliary variables with the variable of interest and from the dependencies among closer units. In addition, there are significant differences in the results depending on the method used to spread the units over the population. In fact, the simple strategy of stratifying the population into n strata and selecting one unit per stratum, strictly depends on the algorithm used to partition the population. When the units are not evenly distributed over the study region so as to be considered very far from a regular grid or the sample size is not small, this criterion may produce unsatisfactory and highly variable results depending on the criterion used to stratify the population units. The combined use of this design with balancing on covariates and/or spatial balancing significantly reduces this problem, but rarely produces better results than just spatial balancing. Thus, our findings appear to be in contrast with those of a practical nature derived by [17], which states: "stratification has performance similar to the more complex spatial schemes but, contrary to these schemes, it straightforwardly provides *spatially balanced* samples and can be well understood and readily planned even by non-statisticians". Further in-depth analyses are therefore necessary to understand how and whether it is worth defining more sophisticated algorithms to achieve greater efficiencies.

Other issues remain open for future research basically related to the possibility of finding a better partition of the study region through the use of flexible and robust algorithms that should assure an appreciable gain in the efficiency of the estimates regardless of the distribution of the observed population. From a practical point of view, it should be emphasized that the use of *spatially balanced* designs, apparently more complex, involves fewer subjective decisions, calculations and approximations than deriving an optimal stratification of the population with a fixed number of strata of the same size and, moreover, possibly consisting of contiguous units.

Acknowledgements The authors would like to thank an anonymous referee and the editors of this issue for their valuable comments which helped to improve the quality of the manuscript.

References

1. Arbia, G.: The use of GIS in spatial statistical surveys. International Statistical Review **61**(2), 339–359 (1993)
2. Barabesi, L., Franceschi, S.: Sampling properties of spatial total estimators under tessellation stratified designs. Environmetrics **22**, 271–278 (2011)
3. Benedetti, R., Espa, G., Taufer, E.: Model-based variance estimation in non-measurable spatial designs. J. Stat. Plan. Inference **181**, 52–61 (2017)
4. Benedetti, R., Bee, M., Espa, G., Piersimoni, F. (eds.): Agricultural Survey Methods. Wiley, Chicester (2010)
5. Benedetti, R., Palma, D.: Optimal sampling designs for dependent spatial units. Environmetrics **6**, 101–114 (1995)
6. Benedetti, R., Piersimoni, F.: A spatially balanced design with probability function proportional to the within sample distance. Biom. J. (2017). doi:10.1111/insr.12216
7. Benedetti, R., Piersimoni, F., Postiglione, P.: Sampling spatial units for agricultural surveys. Advances in Spatial Science Series. Springer, Berlin (2015)
8. Benedetti, R., Piersimoni, F., Postiglione, P.: Spatially balanced sampling: a review and a reappraisal. Int. Stat. Rev. (2017). doi:10.1111/insr.12216
9. Bivand, R., Piras, G.: Comparing implementations of estimation methods for spatial econometrics. J. Stat. Softw. **63**, 1–36 (2015)
10. Breidt, F.J., Chauvet, G.: Penalized balanced sampling. Biometrika **99**, 945–958 (2012)
11. Chauvet, G., Tillè, Y.: A fast algorithm of balanced sampling. Comput. Stat. **21**, 53–62 (2006)
12. Deville, J.C., Tillè, Y.: Unequal probability sampling without replacement through a splitting method. Biometrika **85**, 89–101 (1998)
13. Deville, J.C., Tillè, Y.: Efficient balanced sampling: the cube method. Biometrika **91**, 893–912 (2004)
14. Dickson, M.M., Benedetti, R., Giuliani, D., Espa, G.: The use of spatial sampling designs in business surveys. Open J. Stat. **4**, 345–354 (2014)
15. Dickson, M.M., Tillè, Y.: Ordered spatial sampling by means of the traveling salesman problem. Comput. Stat. **31**, 1359–1372 (2016)
16. Elliott, M.R.: A simple method to generate equal-sized homogenous strata or clusters for population-based sampling. Ann. Epidemiol. **21**(4), 290–296 (2011)

17. Fattorini, L., Corona, P., Chirici, G., Pagliarella, M.C.: Design-based strategies for sampling spatial units from regular grids with applications to forest surveys, land use, and land cover estimation. Environmetrics **26**, 216–228 (2015)
18. Grafström, A.: Spatially correlated Poisson sampling. J. Stat. Plan. Inference **142**, 139–147 (2012)
19. Grafström, A., Lisic, J.: BalancedSampling: balanced and spatially balanced sampling. R package version 1.5.2. https://CRAN.R-project.org/package=BalancedSampling (2016). Accessed 10 Mar 2017
20. Grafström, A., Lundström, N.L.P., Schelin, L.: Spatially balanced sampling through the pivotal method. Biometrics **68**, 514–520 (2012)
21. Grafström, A., Tillè, Y.: Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics **24**, 120–131 (2013)
22. Hahsler, M., Hornik, K.: TSP Infrastructure for the Traveling Salesperson Problem. R package version 1.1-4. https://CRAN.R-project.org/package=TSP (2017). Accessed 10 Mar 2017
23. Hedayat, A., Rao, C.R., Stufken, J.: Sampling designs excluding contiguous units. J. Stat. Plan. Inference **19**, 159–170 (1988)
24. Kincaid, T.M., Olsen, A.R.: spsurvey: spatial survey design and analysis. R package version 3.3 (2016)
25. Lohr, S., Rao, J.N.K.: Inference in dual frame surveys. J. Am. Stat. Assoc. **95**(449), 271–280 (2000)
26. Müller, W.G.: Collecting spatial data: optimum design of experiments for random fields, 3rd edn. Springer, Berlin (2007)
27. R Core Team.: R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/ (2016). Accessed 10 Mar 2017
28. Ranalli, M.G., Arcos, A., Rueda, M., Teodoro, A.: Calibration estimation in dual-frame surveys. Stat. Methods Appl. **25**(3), 321–349 (2016)
29. Stevens Jr., D.L., Olsen, A.R.: Variance estimation for spatially balanced samples of environmental resources. Environmetrics **14**, 593–610 (2003)
30. Stevens Jr., D.L., Olsen, A.R.: Spatially balanced sampling of natural resources. J. Am. Stat. Assoc. **99**, 262–278 (2004)
31. Tobler, W.R.: A computer movie simulating urban growth in Detroit region. Econ. Geogr. (Supplement) **46**, 234–240 (1970)
32. Tillè, Y., Matei, A.: sampling: survey sampling. R package version 2.7. https://CRAN.R-project.org/packageDsampling (2015)
33. Waagepetersen, R.: An estimating function approach to inference for inhomogeneous Neyman–Scott processes. Biometrics **63**, 252–258 (2007)
34. Walvoort, D.J.J., Brus, D.J., de Gruijter, J.J.: An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Comput. Geosci. **36**, 1261–1267 (2010)
35. Wesley.: Spatial Clustering with Equal Sizes. https://www.r-bloggers.com/spatial-clustering-with-equal-sizes/ (2013). Accessed 1 May 2017
36. Wright, K.: agridat: agricultural datasets. R package version 1.12. https://CRAN.R-project.org/packageDagridat (2015)