Received: 29 June 2022 | Revised: 2 October 2023 | Accepted: 22 October 2023

DOI: 10.1002/env.2832

RESEARCH ARTICLE

WILEY

Locally correlated Poisson sampling

Wilmer Prentius

Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden

Correspondence
Wilmer Prentius, Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183, Umeå, Sweden.
Email: wilmer.prentius@slu.se

Abstract

Designs that produces spatially balanced, or well-spread, samples are desirable as they increase the probability of obtaining a sample highly representative of the population. Spatially correlated Poisson sampling (SCPS) is a method for selecting well-spread samples. In the SCPS method, the sampling outcomes (inclusion or exclusion of units) are decided sequentially. After each decision, the inclusion probabilities of surrounding units are updated. A specific order for deciding the sampling outcomes is not enforced for SCPS, that is, the order can be chosen randomly or be fixed. A new modified method called locally correlated Poisson sampling (LCPS) is suggested. In this new method, the order of the decisions makes sure the inclusion probabilities are updated (more) locally. As a result, a stronger negative correlation between inclusion indicators of nearby units is achieved. Simulations on various data sets show that the resulting samples from LCPS, in general, are more spatially balanced and produce lower variance than samples from SCPS and the local pivotal method.

KEYWORDS

auxiliary variables, design-based sampling, environmental monitoring, local pivotal method, spatially correlated Poisson sampling, unequal probability sampling

1 | INTRODUCTION

For environmental surveys, there has been a long-standing interest in using spatial information to achieve some kind of spatial regularity in the sample. Commonly, variants of systematic or stratified designs have been used in order to get a sample which is more representative of the landscape that is surveyed.

The usefulness of sampling designs which can achieve representative samples is not limited to environmental surveys. In many other applications, where x and y coordinates are not applicable, stratification is used as a way to achieve a representative sample with respect to some, often just a few, categorical variables. In other cases, a systematic sample is used in order to capture the distribution of some single numerical, auxiliary variable.

In recent years, many sampling methods which produce well-spread samples over multiple auxiliary variables have been introduced. If these auxiliary variables have some explanatory power on the target variable, such designs will often lead to a decrease in the design-based variance for the target variable (Stevens & Olsen, 2004).

Stevens and Olsen (2004) introduced the general random-tessellation stratified design, which uses a function to map a two-dimensional space to an ordered list, selecting units using systematic πps-sampling, that is, a systematic without replacement design with inclusion probabilities proportional to size. While the mapping preserves some degree of the spatial structure, it cannot fully capture the structure of the population. The cube method, developed by Deville and Tillé (2004), uses auxiliary information in multi-dimensional space to select balanced samples, that is, samples where the Horvitz–Thompson estimate of a total of an auxiliary variable is approximately equal to the population total. Grafström and Tillé (2013) adapted the cube method in order to produce well-spread samples in addition to balanced samples. Balanced acceptance sampling (BAS) (Robertson et al., 2013) selects a sample from a continuous or finite population using quasi-random numbers, that is, pseudo-random numbers which are evenly distributed over an interval. BAS allows for importance sampling through acceptance/rejection sampling. Benedetti and Piersimoni (2017) introduced a sampling design which selects a sample with a probability proportional to the distance between sample units, however not allowing for prescribed inclusion probabilities. Jauslin et al. (2022) proposes a method which selects balanced samples from streamed or sequential populations.

Correlated Poisson sampling (CPS) was introduced by Bondesson and Thorburn (2008), as a πps method usable in real-time sampling situations. The method is list-sequential, that is, a decision is taken one unit at a time, and the conditional probabilities for the remaining undecided units are updated according to the outcome of the decision, using the splitting method (Deville & Tillé, 1998). From CPS, Grafström (2012) developed spatially correlated Poisson sampling (SCPS), where the outcome of a decision prioritized updating the probabilities for the units close in auxiliary variable space, introducing negative correlation for these units' inclusion indicators. The SCPS method produces well-spread samples respecting the prescribed inclusion probabilities (Grafström & Schelin, 2014).

The local pivotal method (LPM) (Grafström et al., 2012) operates similarly to SCPS through the splitting method, however only affecting two units at each iteration of the algorithm, whereas CPS/SCPS may affect multiple units. The LPM comes in two variants, LPM 1 and LPM 2. In LPM 1, for each iteration of the algorithm, two pairwise nearest neighbors are selected at random and compete against each other. Depending on the outcome, their probabilities are updated, moving probability mass in the direction of the winner. For the second variant, LPM 2, a unit is chosen at random, and its competitor is randomly selected among its closest neighbors. Generally, LPM 1 performs the better of the two, as both competitors are each other's nearest neighbors, creating a stronger negative correlation between inclusion indicators close in auxiliary variable space.

In this article, a modification of the SCPS is proposed, called locally correlated Poisson sampling (LCPS). Inspired by the difference between LPM 1 and LPM 2, the units which are affected at each iteration of the LCPS algorithm are selected in a way that guarantees that the updating is done for the smallest possible neighborhood of units. As such, each decision only affects units in a more local area, introducing a stronger negative correlation between the inclusion indicators of these units. Compared to LPM and SCPS, two of the top-performing methods for producing well-spread samples (Benedetti et al., 2015), the proposed modification makes LCPS more efficient than both, when evaluated against a variety of data sets.

In Section 2, the sampling algorithms for LPM and SCPS are presented. Then, in Section 3, the LCPS is introduced, and some properties of the design are presented. The methods are compared through simulation in Section 4, followed by a brief discussion in Section 5.

2 | SCPS AND LPM

Let U be a population of units labeled $1, 2, \dots, N$ with a prescribed inclusion probability vector $\boldsymbol{\pi}$. Furthermore, lets assume that there exists some fully known set of auxiliary variables, on which there exists a distance measure d. Let $\boldsymbol{\pi}^{(t)}$ be a conditional inclusion probability vector at step $t \geq 0$ such that $\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}$. Using the splitting method (Deville & Tillé, 1998), it is possible to split $\boldsymbol{\pi}^{(t-1)}$ into two parts, and selecting a new conditional inclusion probability vector

$$\boldsymbol{\pi}^{(t)} = \begin{cases} \boldsymbol{\pi}^{(t-1)} + (1 - \lambda^{(t)})\mathbf{u}^{(t)} & \text{with probability } \lambda^{(t)}, \\ \boldsymbol{\pi}^{(t-1)} - \lambda^{(t)}\mathbf{u}^{(t)} & \text{with probability } 1 - \lambda^{(t)}, \end{cases}$$

where $\mathbf{u}^{(t)}$ is the updating vector.

2.1 | Spatially correlated Poisson sampling

Let $i(t)$ be the step unit at step $t \geq 1$. For SCPS, $i(t)$ has some predetermined order, say $i(t) = t$, or can be considered randomly drawn from the set of undetermined units

$$U(t) = \left\{ j \in U : \pi_j^{(t-1)} \in (0, 1) \right\}. \tag{1}$$

For simplicity, i will denote $i(t)$, and $U_i(t) = U(t) \setminus i(t)$.

Using the maximal weight strategy (Grafström, 2012), at each step t with step unit i, $\lambda^{(t)} = \pi_i^{(t-1)}, u_i^{(t)} = 1$. The negative elements of $\mathbf{u}^{(t)}$ is decided by the distance

$$D_i^{(t)} = \min_{j \in U_i(t)} d(i,j) \text{ s.t.} \sum_{k \in U_i(t) : d(i,k) \leq d(i,j)} w_i^{(t)}\left(\pi_k^{(t-1)}\right) \geq 1, \tag{2}$$

where

$$w_i^{(t)}(x) = \begin{cases} x/\left(1-\pi_i^{(t-1)}\right), & x \in \left[0, 1-\pi_i^{(t-1)}\right], \\ (1-x)/\pi_i^{(t-1)}, & x \in \left(1-\pi_i^{(t-1)}, 1\right]. \end{cases} \tag{3}$$

For units $j \in R_i^{(t)} = \left\{k \in U_i(t) : d(i,k) < D_i^{(t)}\right\}$, the updating elements are $u_j^{(t)} = -w_i^{(t)}(\pi_j^{(t-1)})$, while the units $j \in \left\{k \in U_i(t) : d(i,k) = D_i^{(t)}\right\}$ on the border equally shares the remainder $-\left(1 - \sum_{j \in R_i^{(t)}} w_i^{(t)}(\pi_j^{(t-1)})\right)$, while ensuring $u_j^{(t)} \geq -w_i^{(t)}\left(\pi_j^{(t-1)}\right)$.

2.2 | The local pivotal method

In LPM 2, a unit i is randomly drawn from the set of unresolved units $U(t)$, defined as (1). Let $U_i(t) = U(t) \setminus i$. A single competitor j is randomly drawn from the set of nearest neighbors

$$\left\{ j \in U_i(t) : \min_{k \in U_i(t)} d(i,k) = d(i,j) \right\},$$

where $d(i,j)$ is the Euclidean distance between units i,j in auxiliary space. If $\pi_i^{(t-1)} + \pi_k^{(t-1)} \leq 1$, then and

$$\begin{aligned} u_i^{(t)} &= \pi_i^{(t-1)} + \pi_j^{(t-1)}, \\ u_j^{(t)} &= -u_i^{(t)}, \\ \lambda^{(t)} &= \frac{\pi_i^{(t-1)}}{\pi_i^{(t-1)} + \pi_j^{(t-1)}}, \end{aligned}$$

whereas if $\pi_i^{(t-1)} + \pi_j^{(t-1)} > 1$, then

$$\begin{aligned} u_i^{(t)} &= 2 - \left(\pi_i^{(t-1)} + \pi_j^{(t-1)}\right), \\ u_j^{(t)} &= -u_i^{(t)}, \\ \lambda^{(t)} &= \frac{1 - \pi_j^{(t-1)}}{2 - \left(\pi_i^{(t-1)} + \pi_j^{(t-1)}\right)}. \end{aligned}$$

For LPM 1, instead of drawing i from all unresolved units, a pair i,j is drawn from the set of pairwise nearest neighbors

$$\left\{ i,j \in U(t) : \min_{k \in U(t) \setminus i} d(i,k) = \min_{l \in U(t) \setminus j} d(j,l) = d(i,j) \right\}.$$

Thus, the average distance that probability is moved will be lower in LPM 1 compared to LPM 2.

3 | LOCALLY CORRELATED POISSON SAMPLING

The ideas behind LPM and SCPS are that it is possible to improve the spatial balance by introducing negative correlation in the inclusion indicators of units close in auxiliary space (Grafström, 2012; Grafström et al., 2012). The SCPS method

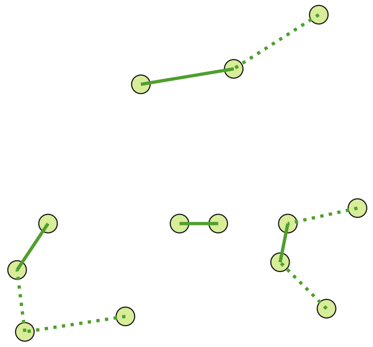

FIGURE 1 In LPM, probability mass is moved in the conditional probability vector between pairs of units close in auxiliary space, exemplified through Euclidean distance in $\mathbb{R}^2$. The set of possible pairs for LPM 1 is highlighted through solid lines, where possible pairs of possible competing units are highlighted through solid and dotted lines for LPM 2.

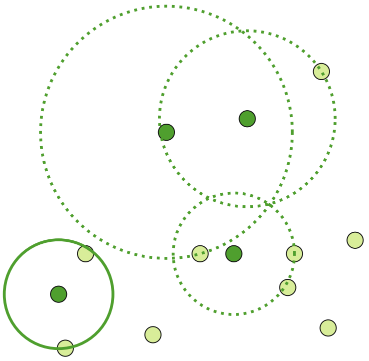

FIGURE 2 For SCPS, probability mass is moved in the conditional probability vector between sets of units close in auxiliary space, exemplified through Euclidean distance in $\mathbb{R}^2$. A step unit is decided, and probabilities will be moved within a radius of this unit. Four such radii are shown as dotted or solid lines, in a setting where each unit has (conditional) probability mass 1/3. For LCPS, the step unit is decided among the set of units with the smallest possible radius, highlighted through a solid line.

does this by sequentially updating the probability vector, moving probability mass to or from units close to a step unit, prioritizing those who are closest. LPM 2 operates similarly, by moving probability between a step unit and it's closest neighbor, whereas for LPM 1, decisions are only taken between units which are pairwise nearest neighbors. In Figure 1, it can be seen that the movement of probability mass is on average lower in LPM 1 compared to LPM 2, where the former generally also produces the most spatially balanced samples (Grafström et al., 2012).

By choosing the step unit for SCPS in a way that reduces the movement of probability mass

$$i(t) = \arg \min_{j \in U(t)} D_j^{(t)},$$

it is possible to increase the spatial balance of the samples that SCPS produces. This modification of SCPS is called LCPS. In Figure 2, the step unit with smallest distance is highlighted by the solid circle.

Grafström (2012) proved that if a population can be partitioned into distinct regions with integer probability mass, in which the maximum distance between units within a region is smaller than any distance to a unit outside the region, SCPS would produce fixed sized samples for each region. For LCPS, this property can be extended to hold for any single such distinct region.

Theorem 1. *For a population U, in which there exists a subset $U_m \subset U$ as an isolated region in auxiliary space such that for all units $i \in U_m$*

$$\max_{j \in U_m} d(i, j) < \min_{k \in U \setminus U_m} d(i, k), \tag{4}$$

where $n_m = \sum_{i \in U_m} \pi_i$ is integer, LCPS will produce a sample from U_m with a fixed size n_m.

Proof. Assume that there exists a partition U_m for which n_m is integer and (4) holds. For a unit $i \in U_m$, the weights that can be provided by other units can be described by the triangular function $w_i(x)$, defined as (3).

As long as weights $w_i(\pi_j), j \in U_m \setminus i$ can be found summing to (at least) one for any arbitrary unit $i \in U_m$, LCPS will decide this unit before deciding any unit outside of U_m which would move probability mass to or from U_m.

If there exists one other unit $j \in U_m \setminus i$, and U_m has integer probability mass, j must have probability $\pi_j = 1 - \pi_i$, and as such $w_i(\pi_j) = 1$. As $w_i(x)$ is linearly increasing/decreasing around $1 - \pi_i$, it is not possible to introduce more units without either moving π_j along the same side of $w_i(x)$, or having the sum of the weights be larger than 1, while keeping the probability mass integer. ■

Furthermore, LCPS provides the same bounds on partitions as LPM 2, for cases where the probability mass in a partition is not integer (Grafström et al., 2012).

Theorem 2. *Let $U_1, \dots U_M$ be a partitioning of U, such that for all partitions U_m*

$$\max_{i,j \in U_m} d(i,j) < \min_{i \in U_m, k \notin U_m} d(i,k), \tag{5}$$

and let $n_m = \sum_{i \in U_m} \pi_m$. If I_i is the inclusion indicator of a unit i, then the sum of inclusion indicators satisfies

$$\lfloor n \rfloor - \sum_{l \neq m} \lceil n_l \rceil \leq \sum_{i \in U_m} I_i \leq \lceil n \rceil - \sum_{l \neq m} \lfloor n_l \rfloor, \tag{6}$$

for all partitions U_m, where $\lfloor \cdot \rfloor, \lceil \cdot \rceil$ are the floor and ceiling functions respectively.

Proof. Let U_m be a partition of a population U satisfying (5). We consider first the case of the upper bound of (6). If the upper bound doesn't hold, it must be possible for a partition U_m to push more than $n_m - \lfloor n_m \rfloor$ into another partition. From Theorem 1, we know that if the probability mass n_m is integer, then no probability mass will leave U_m.

Assume that U_m has been resolved internally to the extent that no further decisions can be taken within U_m without affecting units outside of U_m. Let U_m^* and n_m^* be the remaining, undecided units, and their probability mass. In order to break the upper bound, it must be possible to remove strictly more than $n_m^* - \lfloor n_m^* \rfloor$ from U_m^*.

For an arbitrary unit $i \in U_m^*$, the amount which will be removed from U_m^* upon exclusion of i is $\pi_i W_i$, where

$$\begin{aligned} W_i &= 1 - \sum_{j \in U_m^* \setminus i} \min\left(\frac{\pi_j}{1 - \pi_i}, \frac{1 - \pi_j}{\pi_i} \right) \\ &= 1 - \sum_{j \in U_m^+} \frac{1 - \pi_j}{\pi_i} - \sum_{j \in U_m^-} \frac{\pi_j}{1 - \pi_i} \in (0, 1), \end{aligned} \tag{7}$$

and

$$\begin{aligned} U_m^+ &= \{ j \in U_m^* \setminus i \, : \, \pi_i + \pi_j > 1 \}, \\ U_m^- &= \{ j \in U_m^* \setminus i \, : \, \pi_i + \pi_j < 1 \}. \end{aligned}$$

Units in U_m^+ will have their probabilities updated to 1, whereas units in $j \in U_m^-$ will have their probabilities updated to $\pi_j / (1 - \pi_i)$. Using these components, we can rewrite n_m^* as

$$n_m^* = |U_m^+| + \sum_{j \in U_m^-} \frac{\pi_j}{1 - \pi_i} + \pi_i W_i, \tag{8}$$

where the cardinality $|U_m^+|$ is probability mass which will definitely be kept in in U_m, and the summation term is the remaining "free" probability mass in U_m^* after the exclusion of i.

As the sum of the two latter terms in (8) is strictly less than one,

$$n_m^* - \lfloor n_m^* \rfloor = \sum_{j \in U_m^-} \frac{\pi_j}{1-\pi_i} + \pi_i W_i,$$

which is the same quantity as the maximum amount of probability mass which can be removed from U_m^*, and thus the upper bound of (6) holds.

Similarly for the lower bound; in order to break the lower bound, it must be possible to add strictly more than $\lceil n_m^* \rceil - n_m^*$ from U_m^*. For an arbitrary unit $i \in U_m^*$, the amount which will be added to U_m^* upon inclusion of i is $(1-\pi_i)W_i$. Units in U_m^- will have their probabilities updated to 0, whereas units in $j \in U_m^+$ will have their probabilities update to $1-(1-\pi_j)/\pi_i$. Rewriting n_m^* as

$$n_m^* = 1 + |U_m^+| - \sum_{j \in U_m^+} \frac{1-\pi_j}{\pi_i} - (1-\pi_i)W_i,$$

where the summation term is the probability mass which is left in undecided units, it is obvious that

$$\lceil n_m^* \rceil - n_m^* = \sum_{j \in U_m^+} \frac{1-\pi_j}{\pi_i} + (1-\pi_i)W_i,$$

as these sum of these terms is strictly less than one. Since the maximum amount of probability which can be added to U_m^* after the inclusion of i is

$$\sum_{j \in U_m^+} \frac{1-\pi_j}{\pi_i},$$

it is not possible to add strictly more than $\lceil n_m^* \rceil - n_m^*$ and the lower bound of (6) holds. ■

4 | SIMULATION

The proposed method was evaluated through simulation, measuring the spatial balance of the produced samples through two methods, one using Voronoi polytopes (Grafström et al., 2012; Stevens & Olsen, 2004), and the other using a modified Moran's I index (Tillé et al., 2018). The spatial balance measure based on Voronoi polytopes does not have a fixed range, and the spatial balance of samples can only be interpreted relative to each other, where lower is better. The modified Moran's I index gives values on [−1, 1], where a value of −1 indicates a perfectly spatially balanced sample, and 1 a perfectly clustered sample.

The LCPS method was applied, together with LPM 1, LPM 2, SCPS, and simple random sampling without replacement (SRS), on seven artificial populations, as well as on three openly available real data sets. The LPM and SCPS implementations were provided by the R package `BalancedSampling`, which also contains a C++ implementation of LCPS (Grafström et al., 2022).

The seven artificial populations, shown in Figure 3, were constructed as follows:

a. Poisson cluster process: Three parent locations were randomly located on the unit square. Around each parent location, a random number of children are spawned according to a Poisson distribution with mean 40, and placed relative to the parent according to a normal distributions with variance 0.1. The number of observations were then reduced to 135. Any unit falling outside of the unit square were mirrored back onto the unit square.
b. Poisson cluster process: As (a), but with 20 parents, a Poisson distribution with mean of 20 children spread around the parents with variance 0.01.
c. Regular grid: A rectangular grid of 10 × 10 units.
d. Uniform: 120 units placed uniformly over the unit square.
e. Normal: 120 units placed according to a standard normal distribution along both axes.

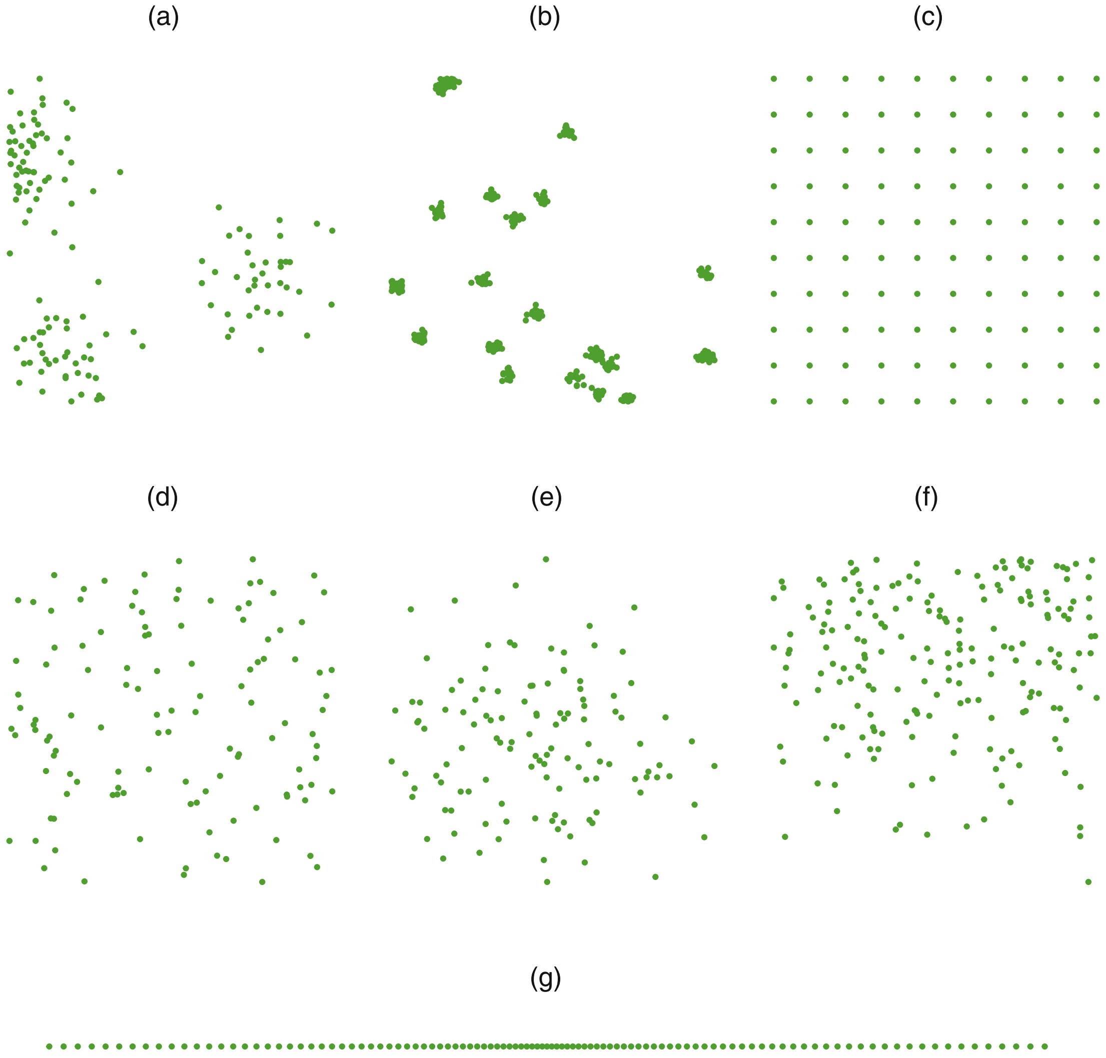

FIGURE 3 Artificial populations used in the simulation: (a) Poisson cluster process, (b) Poisson cluster process, (c) regular grid, (d) uniform, (e) normal, (f) triangular/uniform, (g) line.

f. Triangular/uniform: 200 units placed uniformly on one axis, and according to a (right-angled) triangular distribution on the other.
g. Line: 105 units placed along a straight line, with coordinates $x_i = (i - 53)(1 + \text{abs}(i - 53)/50)$, $i = 1, \ldots, 105$, that is, decreasing distance between units towards the center.

The three real data sets, with auxiliary variables provided in Table 1, were the following:

1. Baltimore: House sale price and characteristics from Baltimore, MD 1978, consisting of 211 observations, provided by the R package `spData` (Bivand et al., 2021).
2. Wheat: Wheat yield data from an agricultural field experiment by Mercer and Hall, consisting of 500 observations in a regular grid, provided by the R package `spData` (Bivand et al., 2021).
3. Meuse: Heavy metal concentrations along the flood plain of the river Meuse, consisting of 155 observations, provided by the R package `sp` (Bivand et al., 2013). Two observations were excluded, as they were missing data on organic matter.

From each population, 10,000 samples were taken with sample sizes 10, 20, and 40, using the previously mentioned methods. For each sample, mean spatial balance measures were calculated, and are presented in Tables 2 and 3. In Table 4, the relative mean squared errors (MSE) are presented for the Horvitz–Thompson estimators of the variable of interest in

TABLE 1 Variable descriptions for data sets.

Data set	Variable	Description
Baltimore	*PRICE*	Sales price of house in ($1000)
	X, Y	X- and Y-coordinates of the house
	NROOM	Number of rooms
	NBATH	Number of bathrooms
	NSTOR	Number of storeys
	GAR	Number of car spaces in garage
	AGE	Age of dwelling (years)
	LOTSZ	Lot size (100 sq. ft.)
	SQFT	Interior living space (100 sq. ft.)
Meuse	*cadmium*	Topsoil cadmium concentration (ppm)
	x, y	X- and Y-coordinates of plot location
	elev	Relative elevation above local river bed (m)
	dist	Distance to the Meuse (normalized to [0, 1])
	om	Organic matter, (pc)
Wheat	*yield*	Wheat yield
	lon, lat	X- and Y-coordinates of plot location

Note: Variable of interest marked by italics.

TABLE 2 Mean spatial balance using Voronoi polytopes for the five sampling methods and various populations.

n	Meth.	(a)	(b)	(c)	(d)	(e)	(f)	(g)	Bal.	Whe.	Meu.
10	SRS	0.503	0.464	0.267	0.311	0.331	0.330	0.407	0.481	0.295	0.336
	LPM1	0.223	0.137	0.082	0.112	0.134	0.119	0.073	0.318	0.087	0.173
	LPM2	0.220	0.144	0.084	0.114	0.138	0.121	0.078	0.329	0.088	0.172
	SCPS	0.215	0.149	0.070	0.106	0.133	0.111	0.070	0.325	0.074	0.164
	LCPS	0.213	0.133	0.070	0.102	0.123	0.107	0.062	0.298	0.072	0.164
20	SRS	0.479	0.560	0.245	0.352	0.354	0.345	0.390	0.434	0.307	0.383
	LPM1	0.173	0.139	0.071	0.127	0.151	0.114	0.068	0.257	0.076	0.190
	LPM2	0.183	0.148	0.073	0.132	0.156	0.120	0.073	0.272	0.078	0.193
	SCPS	0.183	0.156	0.060	0.128	0.153	0.113	0.069	0.267	0.064	0.188
	LCPS	0.164	0.138	0.059	0.124	0.144	0.106	0.056	0.243	0.061	0.184
40	SRS	0.385	0.604	0.173	0.361	0.364	0.350	0.318	0.414	0.298	0.394
	LPM1	0.158	0.201	0.055	0.163	0.147	0.120	0.070	0.228	0.072	0.188
	LPM2	0.168	0.202	0.057	0.170	0.164	0.128	0.086	0.246	0.075	0.202
	SCPS	0.168	0.205	0.045	0.173	0.168	0.127	0.080	0.244	0.062	0.201
	LCPS	0.154	0.200	0.043	0.165	0.150	0.118	0.066	0.222	0.057	0.186

Note: Lower values implies more spatial balance. n = sample size. Meth. = Sampling method.

TABLE 3 Mean spatial balance using a modified Moran's I index for the five sampling methods and various populations.

n	Meth.	(a)	(b)	(c)	(d)	(e)	(f)	(g)	Bal.	Whe.	Meu.
10	SRS	−0.042	−0.030	−0.049	−0.044	−0.042	−0.035	−0.049	−0.029	−0.024	−0.039
	LPM1	−0.294	−0.251	−0.356	−0.324	−0.312	−0.250	−0.402	−0.186	−0.166	−0.266
	LPM2	−0.275	−0.228	−0.333	−0.303	−0.291	−0.236	−0.382	−0.175	−0.157	−0.249
	SCPS	−0.302	−0.235	−0.407	−0.352	−0.339	−0.280	−0.436	−0.204	−0.200	−0.293
	LCPS	−0.350	−0.294	−0.439	−0.391	−0.376	−0.307	−0.479	−0.221	−0.214	−0.322
20	SRS	−0.024	−0.021	−0.023	−0.028	−0.026	−0.023	−0.030	−0.023	−0.016	−0.025
	LPM1	−0.375	−0.340	−0.397	−0.426	−0.410	−0.333	−0.502	−0.264	−0.222	−0.355
	LPM2	−0.346	−0.306	−0.377	−0.389	−0.373	−0.309	−0.469	−0.242	−0.209	−0.326
	SCPS	−0.380	−0.311	−0.459	−0.433	−0.409	−0.363	−0.522	−0.279	−0.270	−0.367
	LCPS	−0.439	−0.384	−0.480	−0.491	−0.467	−0.401	−0.584	−0.307	−0.288	−0.417
40	SRS	−0.014	−0.013	−0.016	−0.018	−0.014	−0.012	−0.014	−0.015	−0.010	−0.015
	LPM1	−0.500	−0.362	−0.443	−0.556	−0.555	−0.458	−0.561	−0.366	−0.300	−0.460
	LPM2	−0.451	−0.331	−0.432	−0.472	−0.481	−0.416	−0.505	−0.330	−0.283	−0.409
	SCPS	−0.478	−0.345	−0.546	−0.474	−0.483	−0.447	−0.573	−0.368	−0.360	−0.441
	LCPS	−0.555	−0.419	−0.563	−0.571	−0.573	−0.512	−0.626	−0.415	−0.387	−0.511

Note: Lower values implies more spatial balance. *n* = sample size. Meth. = Sampling method.

TABLE 4 Relative MSE's of the Horvitz–Thompson estimators of the variable of interest for the four sampling methods and each of the three data sets.

n	Meth.	Bal.	Whe.	Meu.
10	LPM1	0.574	0.831	0.377
	LPM2	0.575	0.850	0.370
	SCPS	0.576	0.846	0.329
	LCPS	0.510	0.815	0.351
20	LPM1	0.475	0.839	0.351
	LPM2	0.493	0.813	0.328
	SCPS	0.453	0.833	0.281
	LCPS	0.423	0.804	0.287
40	LPM1	0.409	0.776	0.273
	LPM2	0.430	0.780	0.276
	SCPS	0.413	0.768	0.252
	LCPS	0.370	0.769	0.246

Note: *n* = sample size. Meth. = Sampling method.

each of the real data sets. The relative MSE is defined relative to the MSE of the SRS, as

$$\text{Relative MSE}(*) = \text{MSE}(*)/\text{MSE}(\text{SRS}),$$

where $*$ is a placeholder for the sampling method. The variables of interest, marked by italics in Table 1, were the sales price for the Baltimore data set, the cadmium concentration for the Meuse data set, and the wheat yield for the wheat data set. The results show that LCPS produces the most spatially balanced samples for all populations. The spatial balance of the sample is also shown to have an effect on the MSE's, as more spatial balance produces lower MSE's.

5 | FINAL COMMENTS

Even though LPM generally performs better than SCPS, there are specific settings where SCPS is the better choice. However, as the results show, for every population, LCPS is better than any of the competing methods in creating well-spread samples.

The derived properties of SCPS and LCPS shows that it is possible to select fixed sized samples for single or multiple strata, if these have integer probability mass and are separated in the auxiliary variables. If the strata does not have integer probability mass, the sample size will at least have some known upper and lower bound.

As with other methods that produces second-order inclusion probabilities that are zero for some pairs of units, LCPS does not have an unbiased variance estimator. In order to get a rough estimate, a conservative variance estimator such as assuming an SRS sample, or a local variance estimator can be used (Grafström & Schelin, 2014).

One advantage of SCPS is the possibility to perform sample coordination, that is, to select two samples with some overlap, through the use of permanent random numbers (Zhao & Grafström, 2020). While it is possible to employ permanent random numbers for LCPS, it remains unclear how dependent this technique is on a fixed order.

Compared to SCPS and LPM, the LCPS algorithm is computationally more expensive to execute. Even though both LCPS and LPM 1 need to find sets of nearest neighbors, LPM 1 does not need the set with the smallest distance, only any such set. The time complexity of SCPS and LCPS would be $\mathcal{O}(N^2 \log N)$ and $\mathcal{O}(N^3 \log N)$ respectively. For the largest population and $n = 40$, the Wheat data set, the average running time on a modest laptop were 0.1 s for LCPS, and 0.5 ms for SCPS. However, this extra computational effort on the part of LCPS should not be of concern when drawing a single sample.

ACKNOWLEDGMENTS

The author is very grateful to the associate editor and the anonymous reviewer for their many useful comments.

DATA AVAILABILITY STATEMENT

Data is available as R Packages spData and sp in CRAN. Simulated populations available upon request to the author.

ORCID

Wilmer Prentius https://orcid.org/0000-0002-3561-290X

REFERENCES

Benedetti, R., & Piersimoni, F. (2017). A spatially balanced design with probability function proportional to the within sample distance. *Biometrical Journal*, *59*(5), 1067–1084. https://doi.org/10.1002/bimj.201600194

Benedetti, R., Piersimoni, F., & Postiglione, P. (2015). *Sampling spatial units for agricultural surveys*. Springer.

Bivand, R., Nowosad, J., Lovelace, R., Monmonier, M., & Snow, G. (2021). *spData: Datasets for spatial analysis* (R package version 2.0.1). https://cran.r-project.org/package=spData

Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R* (2nd ed.). Springer https://asdar-book.org/

Bondesson, L., & Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics*, *35*(3), 466–483. https://doi.org/10.1111/j.1467-9469.2008.00596.x

Deville, J. C., & Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, *85*(1), 89–101. https://doi.org/10.1093/biomet/85.1.89

Deville, J. C., & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, *91*(4), 893–912. https://doi.org/10.1093/biomet/91.4.893

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, *142*(1), 139–147. https://doi.org/10.1016/j.jspi.2011.07.003

Grafström, A., Lisic, J., & Prentius, W. (2022). *Balanced sampling: Balanced and spatially balanced sampling* (R package version 1.6.1). https://cran.r-project.org/package=BalancedSampling

Grafström, A., Lundström, N. L., & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, *68*(2), 514–520. https://doi.org/10.1111/j.1541-0420.2011.01699.x

Grafström, A., & Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, *41*(2), 277–290. https://doi.org/10.1111/sjos.12016

Grafström, A., & Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, *24*(2), 120–131. https://doi.org/10.1002/env.2194

Jauslin, R., Panahbehagh, B., & Tillé, Y. (2022). Sequential spatially balanced sampling. *Environmetrics*, *33*(8), e2776. https://doi.org/10.1002/env.2776

Robertson, B. L., Brown, J. A., McDonald, T., & Jaksons, P. (2013). BAS: Balanced acceptance sampling of natural resources. *Biometrics*, *69*(3), 776–784. https://doi.org/10.1111/biom.12059

Stevens, D. L., Jr., & Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, *99*(465), 262–278. https://doi.org/10.1198/016214504000000250

Tillé, Y., Dickson, M. M., Espa, G., & Giuliani, D. (2018). Measuring the spatial balance of a sample: A new measure based on Moran's I index. *Spatial Statistics*, *23*, 182–192. https://doi.org/10.1016/j.spasta.2018.02.001

Zhao, X., & Grafström, A. (2020). A sample coordination method to monitor totals of environmental variables. *Environmetrics*, *31*(6), e2625. https://doi.org/10.1002/env.2625

How to cite this article: Prentius, W. (2024). Locally correlated Poisson sampling. *Environmetrics*, *35*(2), e2832. https://doi.org/10.1002/env.2832