Using a geocoded sampling frame to improve the quality of surveys

FAVRE-MARTINOZ Cyril, FONTAINE Maëlle, LE GLEUT Ronan, LOONIS Vincent
Insee

Thursday, June 28, 2018

Abstract

This article focuses on using geographical information for survey sampling. Such information can be used at different stages in the survey sampling process. In most surveys carried out face-to-face, multistage sampling designs are used, so as to lower data collection costs through geographically concentrated interviews. Using a geocoded sampling frame is therefore decisive for constructing the primary sampling units. This geographic information can also be used at selection stage to improve the statistical efficiency of the sample when the variables of interest are positively aurocorrelated.

Introduction

The French national statistical institute (Insee) has undertaken the writing of a Handbook of spatial statistic. It will soon undergo copyediting, proofreading, typesetting before it is published in its final form before the end of the this year. The handbook aims at promoting, developing and consolidating the various statistical methods that can be used by National Statistical Institutes (NSIs), only within the framework of a point based system. The specific objectives are :

- to propose a list and the description of these statistical methods which could be implemented in the scope of the point based system.
- to disseminate, promote and share the results among NSIs.
- to promote the application of spatial statistics into the statistical production chain, and feed into the work on the integration of statistics and geospatial information.

In order to provide a pedagogical answer to concrete issues faced by members of the statistical institutes, the pedagogy is specifically addressed to public institute statisticians: the application examples use public statistic's data and are written with R open source programmation language. The theoretical level is high enough to understand all the subtleties of the implementation with R, yet the Handbook doesn't present the latest theoretical developments. The Handbook emphasizes specific issues with which the NSIs are concerned and which do not appear in most of the handbooks : spatial sampling, spatial econometrics on survey data, confidentiality of spatial data, etc. Most of the chapters describe well known methods, yet the Handbook also addresses some innovative subjects or some promising methods that are not yet used in NSIs such as distance based measure, determinantal sampling or spatial econometrics on panel data. This article consists of the handbook's draft version of the chapter on spatial sampling.

Eurostat's Geostat 2 project (2015-2017) was intended to provide a reference framework within which geocoded statistical information could be produced efficiently and used easily. Regarding design-based surveys, the final project report *A Point-Based Foundation for Statistics*, identifies at least three steps of survey design that could benefit from a geocoded sampling frame. Firstly, upstream, when the collection method is face-to-face, precise knowledge of the location of all statistical units allows the creation of

primary sampling units[1] (PSU). Knowing the characteristics of these PSUs makes it easier to manage the interviewers' network while preserving the statistical qualities of the sampling. Secondly, whatever the data collection method, geographical information makes it possible, given certain conditions, to improve the accuracy of estimates by using spatial sampling methods. Thirdly, during the data collection phase, knowing the location of the statistical units sampled makes it easier to identify them when the quality of the addressing is not sufficient.

This chapter focuses exclusively on the first two points. In the first part, we briefly review the sampling theory framework. The second part then offers a method for constructing the smallest primary sampling units in terms of area while having a constant number of statistical units. The third section is dedicated to presenting different spatial sampling methods, while the last part empirically compares their properties, using simulation.

Among the rich literature on the subject, we rely on or direct the reader to Benedetti et al. (2015).

1 General

The purpose of the sampling theory is to estimate the value of a parameter θ measured on a population U of size[2] N. We can think of θ as a function of the values taken by one or more variables of interest associated to each statistical units. Let y_i be the value of the variable y for the statistical unit i in U. The survey statistician does not have access to y_i except for a sub-part of the population, referred to as the sample and expressed as s. He or she aggregates the values observed on the sample thanks to a function called the estimator, taking value $\hat{\theta}(s)$ for s. Estimating θ by $\hat{\theta}(s)$ is known as statistical inference. Properties of statistical inference are described only if s is chosen randomly.

A sampling design is a probability law across the set $\mathcal{P}(U)$ of parts (samples) of U. The classic notation of a random variable with values in $\mathcal{P}(U)$ is $\mathbb{S}$. A sampling design in which all samples with a size different from n ($n \in \mathbb{N}^*$) have zero probability of being selected is said to be of fixed size n. It is generally complex to manipulate a probability law on $\mathcal{P}(U)$. This is why the survey statistician works with summary versions of the law of $\mathbb{S}$, i.e first-order and second-order inclusion probabilities. They refer respectively to the probability of inclusion of unit i in the sample or the joint probability of inclusion of units i and j in the sample: $\pi_i = \mathbb{P}(i \in \mathbb{S})$ and $\pi_{ij} = \mathbb{P}((i,j) \in \mathbb{S})$.

Estimating θ by $\hat{\theta}$ is subject to multiple errors:

- coverage error: some statistical units in the population cannot be selected since they do not appear in the sampling frame;
- non-response error: some individuals have an unknown value of y_i even when they are selected;
- measurement error: collecting incorrect value y_i^* instead of y_i.

An estimator with an expected value different from θ is said to be biased, whereas the variability of values $\hat{\theta}(s)$ is assessed using the variance of $\hat{\theta}$. The objective is to make the bias and variance as small as possible, by paying special attention to the conditions in which information is collected and/or by judiciously choosing the sampling design.

Out of all parameters to be estimated, the most traditional is the population total of one variable of interest y: $\theta = t_y = \sum_{i \in U} y_i$. Out of all the different possible estimators of t_y, we will focus on the Narain-Horvitz-Thompson estimator : $\hat{t}_y = \sum_{i \in \mathbb{S}} y_i/\pi_i$. In the absence of coverage, non-response and measurement errors, this estimator is unbiased. Its variance for a fixed size design is:

$$V(\hat{t}_y) = -\frac{1}{2} \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (1)$$

[1] Primary sampling units are a sub-division of the population often based on geographical criteria. The first stage of selection of PSUs, and the second stage of selection of individuals in these PSUs, is such that collection can be concentrated and costs be reduced when the survey is conducted face-to-face.

[2] In this chapter, in contrast to the previous chapters, the notion of "size" with respect to a geographical area refers to the number of statistical units present inside, not to the surface.

By analysing Equation 1, one can glean indications as to the sampling designs to be used to achieve a more accurate estimate of t_y. If the π_is are proportional to the y_is, the variance equals zero. This solution being impossible in practice, an alternative involves using π_i proportional to x_i, where x is an auxiliary variable known for all statistical units and correlated with y.

This strategy is valid when the survey is mono-thematic (only one variable of interest y). The use of such probabilities for another variable of interest y' uncorrelated with x can indeed result in highly imprecise estimates. Therefore, when the survey is multi-thematic, statisticians often prefer choosing equal first-order inclusion probabilities. Equal first-order-inclusion probabilities make it possible to "reduce to a minimum the variances that would emerge in the most unfavourable configurations (referred to as the MINIMAX optic), [...] i.e. for the variables that are most likely to impair the accuracy of the estimates" (Ardilly (2006)).

When working with a set fixed of first-order-inclusion probabilities, the design should ascribe large π_{ij} when y_i/π_i is very different from y_j/π_j. In the case of spatialised variables and on the assumption that spatial autocorrelation decreases with distance, distant rather than close statistical units should be selected.

2 Constructing primary sampling units of small area and with a constant number of secondary statistical units

2.1 Rationale

Sometimes, organisational constraints imply that the face-to-face collection method is conducted on a low population density area. Then the two-stage sampling method is generally preferred. In order to reduce the costs arising from the interviewers' trips, the first-stage selection will include geographic entities (primary sampling units, PSUs) the geographical area of which must be as small as possible. To simplify, a PSU selected in this manner is then assigned to one interviewer only. Within each PSU, secondary sampling units (SSUs) are selected, each matching up with a statistical unit to be interviewed (individuals in their main dwelling, companies). In order to ensure sufficient workload for the interviewers for one or more surveys, each PSU must also include a minimum number of secondary sampling units.

For a network consisting of m interviewers and a final sample of n secondary sampling units, m PSUs are selected proportionally to their number of secondary sampling units: $\pi_i^{(1)} = m(N_i/N)$ for PSU i bringing together a total of N_i secondary sampling units. Assuming that m divides n, in each of these m PSUs, n/m SSUs are drawn based on an equal-probability design: $\pi_j^i(2) = n/(mN_i)$ for secondary sampling unit j in PSU i. The final inclusion probability is constant: $\pi_j^i = \pi_i^{(1)}\pi_j^{i^{(2)}} = n/N$.

The PSUs are constructed by combining the finest geographical meshes available in the sampling frame. When these meshes remain coarse, for example municipalities, the number of SSUs int final PSUs proves a difficult parameter to control. As a consequence, at first-sampling stage, the design does not benefit from the MINIMAX property referred to above, since probabilities are proportional to the numbers of SSUS. This property can be met if the PSUs are of equal size. Equal-size PSUs may also prove preferable for other reasons connected with coordinating the samples (selecting disjoint or nested samples). Note that:

- the complementary $\overline{\mathbb{S}} = U|\mathbb{S}$ to an equal-probability sample $\mathbb{S}$, is itself an equal-probability sample;
- a random sample $\mathbb{S}_2$, selected with equal probabilities in a sample $\mathbb{S}_1$ itself selected with equal probabilities, is itself a equal-probability sample.

The ideal solution is therefore to construct PSUs covering a small geographical area while having equal numbers of secondary sampling units.

2.2 Method

The problem of constructing equal-size primary sampling units having a small geographical area is a particular case of the more general problem consisting of constructing classification which is subject to size constraints. This topic has enjoyed renewed interest in the recent literature (Malinen and Fränti (2014), Ganganath et al. (2014), Tai and Wang (2017)). The aim is to subdivide the territory into classes within which the dispersion of geographical coordinates is as low as possible, while having an expected number of units per class. Here, we introduce a method initially developed to determine PSUs for the French Labour Force Survey (Loonis (2009)), and recently considered among other possibilities to establish PSUs for the French master sample for households surveys.

The general principle is as follows:

1. the statistical units are geocoded according to the most fine-grained geo-referencing possible. Due to the quality of geo-referencing or the nature of data, the number of statistical units n_{xy} located at the coordinate point $(x; y)$ may be greater than 1.

2. A path is drawn through all known locations. For this purpose, the methods discussed in Chapter 2: "Codifying the neighbourhood structure", are used. Insofar as there is no need for the path to return to its starting point, the Hamilton path has been chosen since it is the shortest (Hamilton path minimises the sum of the distances between two consecutive points without setting a starting or finishing point).

3. To construct M zones, we go through the whole path from the starting point, cumulating n_{xy} units along the way. When the total exceeds the threshold $c \simeq \frac{N}{M}$, the first PSU is constructed. The process is then repeated, from the first point not yet visited on the path.

4. Under ideal conditions where M divides N and $n_{xy} = 1$ for all pairs (x, y), the procedure results in geographically homogeneous primary sampling units of equal size. This heuristic approach does not, however, lead to a global optimum. As in any classification, a consolidation procedure needs to be provided to manage any atypical geographical situation and/or PSU size that is too remote from c. This type of situation can arise, for example, when the last point on the path integrated into a PSU has a very high n_{xy} value, or when dealing with the last PSU formed.

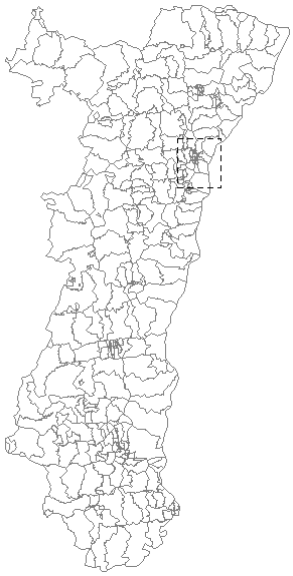
In the following section, we implement this procedure. We focus in particular on how to construct the path when the number of secondary sampling units is high.

2.3 Application

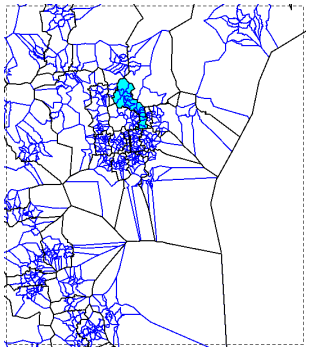
Figure 1 shows the results when the general strategy described above is applied to the Alsace region (former region, before the restructuring of the French regions in 2016). For the purposes of the French Labour Force Survey (LFS) survey, the main dwellings were gathered in PSUs of 2 600 main dwellings (figure 1b), divided into sectors with 120 main dwellings (figure 1c).

Due to the computation time needed to construct the PSUs, the approximately 616 000 main dwellings were initially grouped into 80 000 grid cells whose resolution is 100 meters (figure 1a), which therefore constitute the most fine-grained georeferencing of statistical units. To construct sectors within the PSUs, the main dwellings are by nature geocoded at building level. The original cells or buildings show high variability in size that implies variability in the n_{xy} as well. This partly explains the slight variability in the size of the PSUs and sectors (Table 1).
Constructing sectors of 120 main dwellings, from a large number of main dwellings may lead to performance issues. When using a Euclidean distance, the shortest Hamilton path can computed be exactly and easily if the number of points does not exceed a few hundred. When dealing with several thousand points, it is not reasonable nor useful to compute the exact optimal path. We therefore propose an approximate method aiming at constructing a path that fits for the purpose. The different steps in this approximate method are described below and illustrated in figure 2 for one given PSU. This PSU is composed of 2 600 main dwellings and 1 085 buildings.

(a) 616 000 main dwellings, in 80 000 cells 100 m per side...



(b) ... are gathered in homogeneous PSUs of 2 600 main dwellings...



(c) ... and divided into sections of 120 main dwellings each.

Figure 1: Construction of zones with a small geographical area and an equal number of main dwellings in Alsace

Order of Fractile	Size of cell of origin	Size of PSUs (figure 1b)	Size of sectors (figure 1c)
100 %	378	2776	139
99 %	59	2757	131
95 %	23	2685	130
90 %	15	2640	130
75 %	9	2606	124
50 %	5	2595	119
25 %	2	2591	118
10 %	1	2587	118
5 %	1	2502	117
1 %	1	2491	111
0 %	1	2479	99

Table 1: Quantiles of the number of main dwellings in the cells, PSUs and sections of figure 1

1. The 1 085 buildings and 2 600 main dwellings of the PSU in blue in Figure 1c, are gathered using the k-means method[3] into 20 different but geographically consistent classes. This classification is carried out with the geographical coordinates of the buildings. It should be noted that $20 \simeq \frac{2600}{120}$(figures 2a and 2b).
2. A **Hamilton** path is drawn through the barycentres of the 20 classes so that they can be ordered (figure 2c).
3. In a given class i, the **buildings** are sorted according to **two sub-classes** (figure 2d):
   (a) the first comprises the buildings in class i that are closer to G_{i-1} (barycentre of the previous class) than to G_{i+1} (barycentre of the next class) and increasingly sorted by distance to G_{i-1} ;
   (b) the second comprises the buildings in class i that are closer to G_{i+1} (barycentre of the next class) than to G_{i-1} (barycentre of the previous class) and decreasingly sorted by distance to G_{i+1} ;
4. By construction, the first buildings in class i are close to the last buildings of class $i-1$, and the last buildings in i are close to the first in class $i+1$. Following the path thus means running along buildings by class, by sub-class and finally by increasing or decreasing distance, depending on the case (figure 2e). If necessary, main dwellings inside a building can be sorted by floor.

3 How to draw a spatially balanced sample

The overall considerations have shown that the more the sampling design selects individuals geographically distant from one another, the more the estimatation will be precise for a spatially autocorrelated variable. Grafström and Tillé (2013), for example, have formalised these considerations more explicitly. In this section, we detail the methods for selecting spatially balanced samples. Existing methods can be grouped into two families.

Within the first family, the inclusion probabilites are updated locally in order to limit the selection of two neighbouring units. These methods include the spatially correlated Poisson sampling method (Grafström (2012)), the local pivotal method (Grafström et al. (2012)), and the local cube method (Grafström and Tillé (2013)). Within the second family, we turn the problem of proximity between units in several dimensions into a problem of order in $\mathbb{R}$. Then, the sampling is performed excluding two nearby units, basing on the sorted file. This family of methods includes the *General Randomized Tessellation Stratified* (GRTS, Stevens Jr and Olsen (2004)) method, the method based on a Peano curve (Lister and Scott (2009)), or on Traveling-Salesman Problem (TSP) algorithm (Dickson and Tillé (2016)).

3.1 The spatially correlated Poisson method (Grafström (2012))

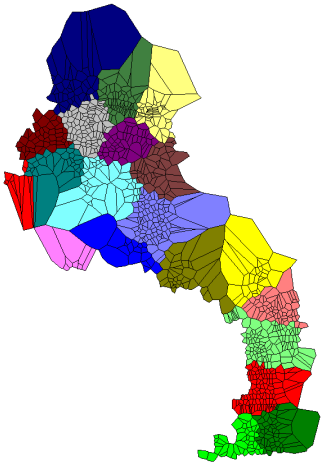
The spatially correlated Poisson sampling is an extension of the correlated Poisson sampling (*Correlated Poisson Sampling*, CPS) proposed by Bondesson and Thorburn (2008) to perform realtime sampling. The CPS method is based on sequential and orderly sampling of units. Units are ordered with indices ranging from 1 to N. The decision is first made as to unit 1, then unit 2, up to unit N. In the case of real time sampling, the order of the indices is a pre-established order of sampleable units. In the case of a spatial sampling, the order can be based on the proximity of the units, in accordance with an Euclidean distance function. At each stage, the inclusion probabilities are updated so as to create a positive or negative correlation between the unit selection indicators.

More specifically, the first unit is included in the sample with probability $\pi_1^0 = \pi_1$. If unit 1 was included, we set $I_1 = 1$. More generally speaking, at stage j, unit j is selected with probability π_j^{j-1} and the inclusion probabilities of units $i \geq j+1$ are updated as follows:

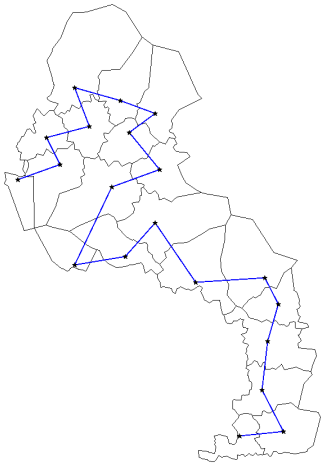
$$\pi_i^j = \pi_i^{j-1} - \left(I_j - \pi_j^{j-1}\right) w_j^i, \tag{2}$$

[3]The k-means method aims at creating homogeneous classes by maximising between-class variance and minimising within-class variance.

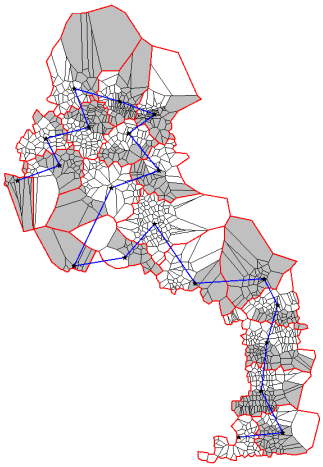
(a) The buildings and their related Voronoi polygons...



(b) ... are grouped, by k-means on the coordinates, into approximately 20 clusters of varying size.

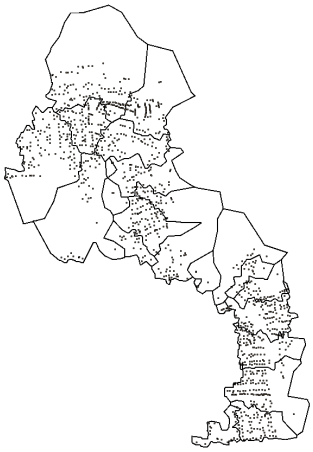


(c) A path through the cluster's barycentres...



(d) ... makes it possible for buildings to be classified according to whether they are closer to the barycentre of the previous cluster (white) or the next one (grey) ...

(e) ... and thus to create a path passing through all the buildings.



(f) Following this path, sectors of 123 to 128 main dwellings are built.

Figure 2: Main dwellings divided into 120-unit sections

where w_j^i is the weight given by unit j to units with indices $i \geq j+1$. The inclusion probabilities are updated stage by stage, with at most N stages until the selection indicator vector is obtained.

The choice of weights w_j^i is crucial as it helps determine whether a positive or negative correlation is introduced between the selection indicators. Bondesson and Thorburn (2008) give the expression of these weights for some conventional sampling designs, and a general expression for any sampling design. Consequently, this method is very general: any sampling design with fixed first-order inclusion probabilities can be implemented by the CPS method. Only the expression and conditions related to to weights [4] may vary according to the design. For example, for a fixed-size design, the sum of the weights $w_j^i, j < i$ must be equal to 1. In case of positive spatial auto-correlation (wherein nearby units are similar), the associated weights should be chosen positively, so as to introduce a negative correlation between the sampling selection indicators. It therefore seems appropriate to carry out a global spatial autocorrelation test to determine the sign of the weights to be used in this method.
Grafström (2012) suggests two versions for the weights w_j^i. Here, we show the version considering a Gaussian distribution. In this case, the weights are defined as:

$$w_j^i \propto \exp(-\left[d\left(i,j\right)/\sigma\right]^2),\ i = j+1, j+2, ..., N. \tag{3}$$

Since the sum of the weights must be equal to 1, the proportionality constant is set. These weights are all the larger as the units are close to the unit j. Thus, the closer is unit i (in the sense of distance $d(i,j)$) to unit j, the lower is probability π_i^j, and spatially balanced sampling can be carried out. Parameter σ makes it possible to manage the dispersion of these weights, and therefore distribute the update of inclusion probabilities in a wider or smaller neighbourhood, as needed.

This method is implemented in the R *BalancedSampling* package (Grafström and Lisic (2016)) using the function `scps()`.

3.2 The local pivotal method (Grafström et al. (2012))

3.2.1 Review of the local pivotal method

The local pivotal method is a sampling procedure thanks to which a sample with equal or unequal inclusion probabilities can be selected (Deville and Tille (1998)). At each stage of the algorithm, the inclusion probabilities of two units i and j in competition are updated and at least one of the two units is selected or rejected.
The inclusion probability vector of the two competing units (π_i, π_j) is updated according to the following rule (fight between units i and j):

- if $\pi_i + \pi_j < 1$, then:

$$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) \text{ with probability } \dfrac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) \text{ with probability } \dfrac{\pi_i}{\pi_i + \pi_j} \end{cases}$$

- if $\pi_i + \pi_j \geq 1$, then:

$$(\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1) \text{ with probability } \dfrac{(1-\pi_j)}{(2-\pi_i-\pi_j)} \\ (\pi_i + \pi_j - 1, 1) \text{ with probability } \dfrac{(1-\pi_i)}{(2-\pi_i-\pi_j)} \end{cases}$$

This procedure is repeated until an inclusion probability vector emerges containing $N-n$ times the number 0 and n times the number 1, which will completely determine the selected sample (steps at most N).

[4] the conditions imposed on weight are linked to the conditions imposed on the inclusion probabilities, i.e. the sampling design.

3.2.2 Extension to spatial sampling

The local pivotal method (Grafström et al. (2012)) is a spatial extension of the pivotal method. The idea of the method is still to iteratively update the inclusion probabilities vector $\boldsymbol{\pi}$, but this time, at each step, we select for the fight two neighbouring units, in terms of a certain distance (e.g. a Euclidean distance). Several various methods can be used to select these two neighbouring units:

- **LPM1**: two units as close as possible to one another are selected to participate in the fight, *i.e.* one unit i is randomly selected among N population units, then unit j closest to i is selected to participate if and only if i is also the closest unit to j (at best N^2 steps, at worst N^3 steps);
- **LPM2**: two neighbouring units are selected to participate in the fight, *i.e.* one unit i is randomly selected from among N units of the population, then unit j closest to i is selected to participate in the fight (N^2 steps);
- **LPM K-D TREE**: the two neighbouring units are selected using spatial partitioning k-d tree (Lisic (2015)) making it possible to search for closer neighbours quicker (complexity of the algorithm in $N\log(N)$).

These three local pivotal methods are implemented in C++ in the *BalancedSampling* package of the R software.

3.3 The cube method

3.3.1 General information about the cube method

Balanced sampling is a procedure aimed at providing a sample that complies with the following two constraints:

- the inclusion probabilities are respected;
- the sample is balanced on p auxiliary variables. In other words, the Narain-Horvitz-Thompson estimators of the totals of the auxiliary variables are equal to the totals of these auxiliary variables in the population:

$$\sum_{i\in\mathbb{S}} \frac{\boldsymbol{x}_i}{\pi_i} = \sum_{i\in U} \boldsymbol{x}_i \tag{4}$$

An algorithm for making such a sampling is called the cube algorithm. To describe the principle, it is appropriate to use the following geometric representation. A sample is one of the vertices of a hypercube with dimension N, expressed as C. All p constraints, recapitulated in equation (4), define a hyperplane with dimension $N-p$, expressed as Q. Using $K = Q \cap C$, we depict the intersection between the cube and the hyperplane. A graphical representation of the problem in dimension 3, derived from article Deville and Tillé (2004), is shown in Figure 3.

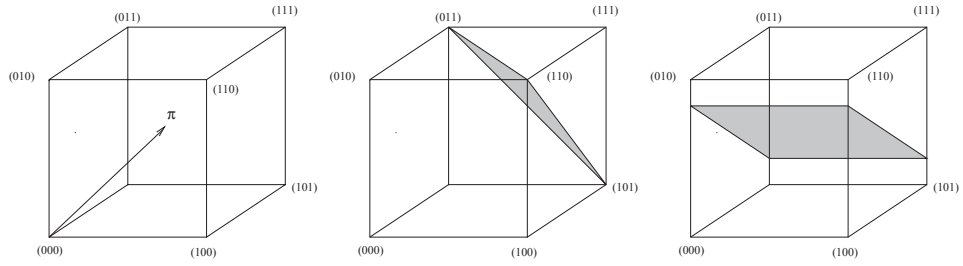


Figure 3: Graphical representation of the cube for $N = 3$ and different possible configurations of the space subject to constraint, with $p = 1$

The cube algorithm is divided into two phases. The first phase, referred to as the *flight phase* (figure 4), is a random walk from the inclusion probabilities vector and changes in K. For this, we start from $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$, then update the inclusion probabilities vector by choosing a vector $\boldsymbol{u}(0)$ such that $\boldsymbol{\pi} + \boldsymbol{u}(0)$ remains within the space of the constraints. By following the direction indicated by vector $\boldsymbol{u}(0)$, we

necessarily end up on one face of the cube. The way to update the inclusion probabilities vector is then provided by parameters $\lambda_1^*(0)$ and $\lambda_2^*(0)$, chosen so that updated vector $\boldsymbol{\pi}(1)$ reaches one face of the cube. The update is chosen randomly so that $E(\boldsymbol{\pi}(1)) = \boldsymbol{\pi}(0)$. The operation is then repeated by choosing a new vector $\boldsymbol{u}(1)$ for the direction and a new sense for updating the inclusion probabilities. This random walk stops when it reaches a vertex point $\boldsymbol{\pi}^*$ of K. At the end of this first phase, vertex $\boldsymbol{\pi}^*$ is not necessarily a vertex of the cube C. Let q be the number of non-integer components in vector $\boldsymbol{\pi}^*$ ($q \leq p$). If q is null, the sampling procedure is completed; otherwise a second step, referred to as the *landing phase*, needs to be initiated. It consists in relaxing the balancing constraints as little as possible, and re-initialising a flight phase with these new constraints until a sample is obtained. It is not possible to change the space of the constraints from the outset in a way that might mix the vertices of K with C, as this would amount testing all possible samples to first see whether one of them allows the constraints to be met. Changing the constrained space in a later phase (the landing phase) makes it possible to work on a population U^* of smaller size ($\dim(U^*) = q$). The problem can thus be solved because the number of samples to be considered is reasonable.

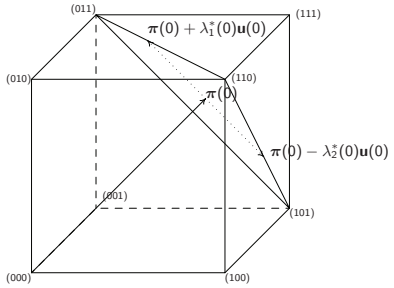


Figure 4: First step of the cube flight phase for $N = 3$ and a constraint ($p = 1$) of a fixed sample size $n = 2$

The implementation of this algorithm is available in SAS thanks to macro *FAST CUBE* or in the R *BalancedSampling* package.

3.3.2 The local cube method

The general idea of the spatially balanced sampling algorithm is to build a cluster of $p+1$ geographically close units, and then to apply the cube flight phase to this cluster. This leads to decide whether a unit is selected in this cluster or not, while respecting p local constraints within the cluster. Next, the probabilities are modified locally, ensuring that the inclusion probabilities of the nearby units are reduced if the unit on which a decision has been made is selected. This lowers the probability that one of its nearby units is selected in the following step of the algorithm. Then, the procedure is repeated: a unit is selected, followed by a cluster of $p + 1$ units around it, and we apply the cube flight phase with the inclusion probabilities updated in the previous step. The process is repeated as long as there are still more than $p + 1$ units. Finally, the traditional cube landing phase is applied.

The spatially-balanced sampling method described above is available in the R *BalancedSampling* package. This package, developed in C++, allows the algorithm to be applied very quickly.

3.3.3 Balancing on moments

The definition of a spatially balanced sample suggests a different use of the cube algorithm for spatial sampling. For Marker and Stevens Jr (2009), "a sample is spatially balanced if the spatial moments of the localised samples match the spatial moments of the population. Spatial moments are the centre of gravity and inertia." In the cube algorithm terminology, this definition may result in selecting a balanced sample on variables defined from the geographic coordinates: $x_i, y_i, x_i^2, y_i^2, x_i y_i$. In order to respect the first and second non-central moments:

- $T_x = \sum_{i \in U} x_i$,

- $T_y = \sum_{i \in U} y_i$,
- $T_{x^2} = \sum_{i \in U} x_i^2$,
- $T_{y^2} = \sum_{i \in U} y_i^2$,
- $T_{xy} = \sum_{i \in U} x_i y_i$.

3.4 Ordered spatial sampling methods

The methods in this family (Stevens Jr and Olsen (2004), Dickson and Tillé (2016), Lister and Scott (2009)) firstly rely on the creation of a path going through all statistical units. This path can be GRTS (*Generalized Random Tessellation Stratified*), Traveling-Salesman Problem (TSP) or a Peano curve. Given the order defined by this path, the aim is then to select a sample according to a method that excludes two nearby units, for example systematic sampling.
Other path-building methods exist (Hamilton paths, or curves filling space: Hilbert, Lebesgue). Similarly, other selection methods exclude nearby units with a given ordering, such as determinantal sampling designs (Loonis and Mary (2018)). We describe here the repulsiveness properties of the systematic and determinantal sampling designs.

3.4.1 The systematic sampling method

Systematic sampling is a sampling method that is simple to implement and makes it possible to carry out sampling with unequal probabilities while respecting those inclusion probabilities. This method was proposed by Madow (1949), then extended by Connor (1966), Brewer (1963), Pinciaro (1978), and Hidiroglou and Gray (1980). It is very often used in practice for telephone surveys, for sampling on continuous data flows, or in sampling housing units for INSEE household surveys.

To draw a fixed size sample n respecting the inclusion probability vector $\boldsymbol{\pi}$, we start by defining the cumulative sum of the inclusion probabilities by $V_i = \sum_{l=1}^{i} \pi_l, i \in U$, with $V_0 = 0$. For a fixed size sample, the result is $V_N = n$. The systematic sampling algorithm shown below is then used to decide on the units to be sampled.

Systematic sampling algorithm :

- Generate a random variable u uniformly distributed on the interval [0.1].
- For $i = 1, ..., N$,

$$I_i = \begin{cases} 1 & \text{if there is an integer } j \text{ so that } V_{i-1} \leq u + j - 1 < V_i, \\ 0 & \text{otherwise.} \end{cases}$$

Table 2 provides an example of the method for $n = 3$ and $N = 10$.

i	1	2	3	4	5	6	7	8	9	10
π_i	0.2	0.2	0.3	0.3	0.4	0.4	0.3	0.3	0.3	0.3
V_i	0.2	0.4	0.7	1	1.4	1.8	2.1	2.4	2.7	3(=n)

Table 2: An example of systematic sampling

For example, if the random number generated u is equal to 0.53, then units 2, 5 and 8 will be selected because they meet the constraints:

$$V_2 \leq u < V_3 V_5 \leq u + 1 < V_6 V_8 \leq u + 2 < V_9.$$

According to this method, units (i, j) respecting $|V_i - V_j| < 1$ have zero probability of being selected together. If the file is wisely sorted, this ensures the geographical spread of the sample.

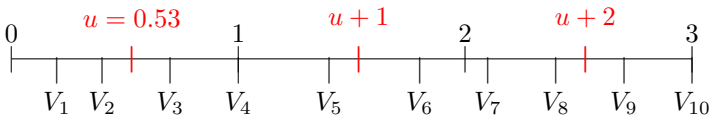


Figure 5: 3 units selected out of 10

Implementation of the GRTS method

The GRTS method is one of the most frequently-used methods in practice when it comes to systematic sampling on a geographically-ordered file. GRTS ordering is described in chapter 2 "Codifying geographical structure".
The R *gstat* package of software R has been implemented specifically to produce samples using this method. However, the GRTS method has some drawbacks, in particular the fact that the cutting algorithm and the sampling algorithm are not dissociated, nor is the method's computational time. Indeed, the method proposes by default to stop at 11 hierarchical levels in the decomposition process, as the time needed to execute the method may be too long if a more detailed cutting is requested. This makes it difficult for the GRTS algorithm to adapt to large populations. In order to overcome these computational limits, a new pivotal method using another tessellation algorithm (Chauvet and Le Gleut (2017)) has been developed in R. In this method, the tessellation algorithm (very similar to the GRTS one) is dissociated from the sampling algorithm. This method is based on binary decomposition, making it possible to carry out the decomposition directly on 31 levels. The computational time is therefore considerably improved. In addition, it is possible to use this method in more than two dimensions.

3.4.2 Determinantal sampling design

By definition, given a random variable $\mathbb{S}$ with values in 2^U, the probability law will be a determinantal sampling design if and only if there is a contracting hermitian matrix[5] K indexed by U, referred to as a kernel, as for all $s \in 2^U$,

$$p(s \subseteq \mathbb{S}) = \det(K_{|s}) \tag{5}$$

where $K_{|s}$ is under the matrix of K indicated by units of s. This definition directly gives rise to the calculation of inclusion probabilities (table 3).

π_i	=	$pr(i \in \mathbb{S})$	=	$\det(K_{\|\{i\}})$	=	K_{ii}
π_{ij}	=	$pr(i,j \in \mathbb{S})$	=	$\det\begin{pmatrix} K_{ii} & K_{ij} \\ \overline{K}_{ij} & K_{jj} \end{pmatrix}$	=	$K_{ii}K_{jj} - \mid K_{ij} \mid^2$

Table 3: Calculation of simple and joint inclusion probabilities in a determinantal sampling design DSD(K) ($|z|$ refers to the complex number module z.)

The diagonal entries of K are the simple inclusion probabilities. Another particularly important result of determinantal sampling designs is the following: a determinantal design is of a fixed size if and only if K is a projection matrix [6](Hough et al. (2006)).
Let us consider all projection matrices in which the diagonal is a vector Π of inclusion probabilities *a priori*. Among them, matrix K^Π (whose coefficients are provided in table 4) offers interesting properties in terms of spatial repulsion.

The repulsiveness of the determinantal sampling design associated to K^Π for close statistical units (given the order listed in the file) is illustrated by the following properties (Loonis and Mary (2018)):

1. the design will select at most one unit within a range of the form $]i_r + 1, i_{r+1} - 1[$;
2. if a unit is drawn in that range of indices, as well as the "close" unit i_{r+1}, then the design will not select an additional "close" unit, i.e. in $]i_{r+1} + 1, i_{r+2} - 1[$;

[5]A complex matrix K is hermitian if $K = \overline{K}^t$, where the coefficients of $\overline{K}$ are conjugates of those of K. A matrix is a contracting matrix if all its own values are between 0 and 1.
[6]A hermitian matrix is projection if its own values are 0 or 1.

	Values of j	
Values of i	$j = i_r$	$i_r < j < i_{r+1}$
$i_{r'} < i < i_{r'+1}$	$-\sqrt{\Pi_i}\sqrt{\frac{(1-\Pi_j)(\Pi_j-\alpha_l)}{1-(\Pi_j-\alpha_j)}}\gamma_r^{r'}$	$\sqrt{\Pi_i\Pi_j}\gamma_r^{r'}$
$i = i_{r'+1}$	$-\sqrt{\frac{(1-\Pi_i)\alpha_i}{1-\alpha_i}}\sqrt{\frac{(1-\Pi_j)(\Pi_j-\alpha_j)}{1-(\Pi_j-\alpha_j)}}\gamma_r^{r'}$	$\sqrt{\frac{(1-\Pi_i)\alpha_i}{1-\alpha_i}}\sqrt{\Pi_j}\gamma_r^{r'}$

where for every r such as $1 \leq r \leq n$:

- $1 < i_r \leq N$ is a integer such that $\sum_{i=1}^{i_r-1} \Pi_i < r$ et $\sum_{i=1}^{i_r} \Pi_i \geq r$; by convention, it is established that $i_0 = 0$
- $\alpha_{i_r} = r - \sum_{i=1}^{i_r-1} \Pi_i$. It should be noted that $\alpha_{i_r} = \Pi_{i_r}$ if $\sum_{i=1}^{i_r} \Pi_i = r$.
- $\gamma_r^{r'} = \sqrt{\prod_{k=r+1}^{r'} \frac{(\Pi_{i_k}-\alpha_{i_k})\alpha_{i_k}}{(1-\alpha_{i_k})(1-(\Pi_{i_k}-\alpha_{i_k}))}}$ for $r < r'$, $\gamma_r^{r'} = 1$ otherwise.

Table 4: Values of K_{ij}^{Π} with $i > j$

3. this design will always have at least one individual in an interval $[i_r + 1, i_{r+1} - 1]$;
4. if $|i - j|$ is large enough, then $\pi_{ij} \approx \Pi_i\Pi_j$ that is the joint inclusion probabilities of the Poisson design.

Applying the results of the determinantal sampling designs to the probabilities defined in table 2 results in quantities: $i_1 = 4, i_2 = 7, i_3 = 10$ and $\alpha_4 = 0.3 = \Pi_4, \alpha_7 = 0.2, \alpha_{10} = 0.3 = \Pi_{10}$. The joint inclusion probabilities are given in the matrix below.

$$\begin{pmatrix}
 & 0 & 0 & 0 & \frac{2}{25} & \frac{2}{25} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} \\
0 & & 0 & 0 & \frac{2}{25} & \frac{2}{25} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} \\
0 & 0 & & 0 & \frac{3}{25} & \frac{3}{25} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} \\
0 & 0 & 0 & & \frac{3}{25} & \frac{3}{25} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} \\
\frac{2}{25} & \frac{2}{25} & \frac{3}{25} & \frac{3}{25} & & 0 & \frac{1}{20} & \frac{7}{60} & \frac{7}{60} & \frac{7}{60} \\
\frac{2}{25} & \frac{2}{25} & \frac{3}{25} & \frac{3}{25} & 0 & & \frac{1}{20} & \frac{7}{60} & \frac{7}{60} & \frac{7}{60} \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{1}{20} & \frac{1}{20} & & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{7}{60} & \frac{1}{15} & & 0 & 0 \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{7}{60} & \frac{1}{15} & 0 & & 0 \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{7}{60} & \frac{1}{15} & 0 & 0 &
\end{pmatrix}.$$

The entries around the main diagonal tend to be null or close to 0, reflecting repulsiveness.

4 Comparing methods

Various sampling methods aimed at taking spatial information into account have been presented. This section compares their relative efficiency, using real data.

4.1 The principle

The tax data for 2015 is geocoded for all households, making it possible to split the territory of the Provence-Alpes-Côte d'Azur region (PACA) into 1 012 primary sampling units (PSU) of approximately 2 000 primary residences. Each of these PSUs is characterised by fifteen variables of interest describing its socio-economic or demographic situation. We focus here on the statistical properties of first-stage sampling, i.e. a sampling of m primary units from amongst the $M = 1012$ PSUs.

Only the geographical coordinates of the barycentres of the PSU at the time of sample selection are available. Two sets of inclusion probabilities are tested: the first with equal probabilities, the second

with probabilities proportional to the number of the unemployed. Both of these sets are tested for three different sample sizes: $m = 30, 60, 100$.

The aim is to assess the methods presented above by comparing their performance with those of a benchmark method. This benchmark is single random sampling (SRS) for equal probability designs, and systematic sampling on a randomly sorted file for unequal probabilities sampling designs. Each method is assessed through two types of indicators:

1. **Estimating the variance**
   For each method, the aim is to determine to what extent the variance of the total of a given variable is reduced in respect to to the variance found with the *benchmark* method. This is achieved by studying a set of variables of interests with different levels of spatial autocorrelation.
   For all methods apart from the determinantal sampling design, variances of totals are estimated using the Monte Carlo method, replicating 10,000 times each method for each set of inclusion probabilities and each sample size. Concerning determinantal sampling designs, variance can be computed exactly since the joint inclusion probabilities are known.
   The aim is to find out whether the gain in variance is greater when the variable is spatially autocorrelated. Therefore, the 15 variables of interest are ranked according to their level of spatial autocorrelation, measured by Moran's I dilated by inclusion probabilities, since $\frac{y_i}{\pi_i}$ determines the quality of the results according to Equation 1. When the design is an equal probability sampling design, it is similar to computing directly Moran's I for each variable (table 5).

2. **The Voronoï indicator**
   For each method, an empirical dispersion indicator (known as the Voronoï index) is also computed, by following Stevens Jr and Olsen (2004). The principle is as follows:

   - the Voronoï diagram is built only with the m selected PSUs;
   - for a given selected PSU i, PSUs located in the Voronoï polygon associated with i are identified from amongst the 1 012 original PSUs;
   - the sum δ_i of these PSUs' inclusion probabilities is computed. The average of the δ_i is equal to 1, since the sum of the inclusion probabilities over the 1 012 PSU is m and because the m polygons partition the territory. If the procedure has selected only few units around a given selected PSU i, δ_i will be greater than 1. If the procedure has selected a lot of other units close to i, δ_i will be less than 1 (see figure 6);
   - for a random sample $\mathbb{S}$, the Voronoï indicator is then defined by:

   $$\Delta_{\mathbb{S}} = \frac{1}{m-1} \sum_{i \in \mathbb{S}} (\delta_i - 1)^2.$$

   The more uniformly a procedure spreads the units, the lower the dispersion of δ_i measured by $\Delta_{\mathbb{S}}$ will be. The expected value of $\Delta_{\mathbb{S}}$ will be estimated by simulation (average over 10 000 replications, noted V).

   The Voronoï index can be computed in R using the function `sb()` of the *BalancedSampling* package or based on the R codes provided in Benedetti et al. (2015) (pp. 161-162).

4.2 Results

Ten spatial sampling methods are studied:

- 4 methods in the so-called "A" family in the following. Family A consists in updating inclusion: Poisson sampling, local pivotal, local cube[7] and balanced cube on spatial moments;

[7] The local pivotal and local cube are equivalent methods in this context.

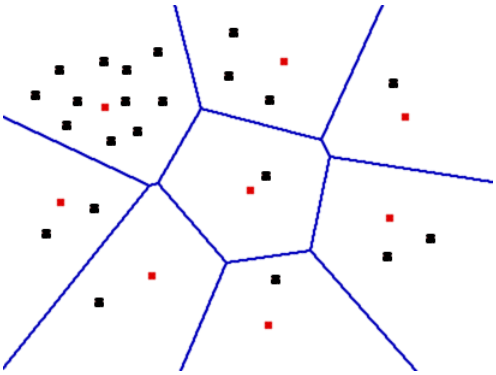
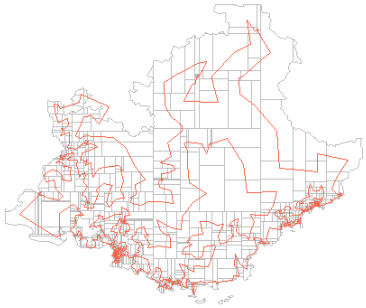


Figure 6: Calculation of the Voronoi index.
Note: Voronoi polygons are built around the selected units (red). δ_is are calculated on all units (red and black).

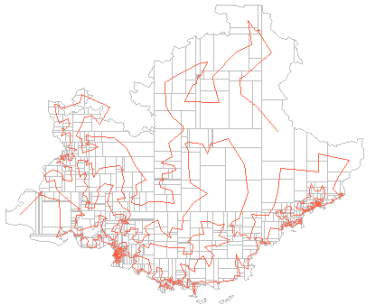
Variable	**Moran I** π constant	**Moran's I dilated** $(\frac{y_i}{\pi_i})$
Number of households earning agricultural income	0,68	0,66
Total wage income	0,62	0,55
Number of couples with child(ren)	0,61	0,54
number of those receiving minimum social benefits	0,60	0,61
Number of poor	0,58	0,58
Number of children	0,55	0,52
Number of people living in a neighbourhood targeted by City Policy	0,55	0,54
Number of households owning their homes	0,52	0,47
Total standard of living	0,46	0,46
Number of unemployed	0,45	0,42
Number of single-parent families	0,41	0,43
Number of individuals	0,40	0,34
Number of men	0,39	0,34
Number of women	0,24	0,34
Number of households	0,08	0,38

Table 5: Moran's indices for different variables computed at the PSU level of PACA.
Source : *INSEE, Fideli 2015.*

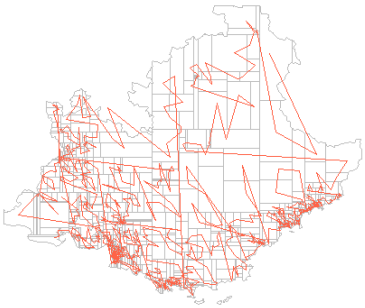
- 6 methods in a second so-called "B" family of methods. Family B methods are based on prior ordering of the file. In total, 3 paths are considered (figure 7): TSP (7a), Hamilton (7b), and GRTS (7c), and each followed by a systematic sampling or a determinantal sampling. These three paths are computed using an exact method. Then, all the sampling replications run on a single sorted file.



(a) Traveling-Salesman Problem (TSP)



(b) Hamilton



(c) GRTS

Figure 7: Paths connecting PSU centroids.
Source : *INSEE, Fideli 2015.*

Figure 8 provides $(V^q - V^{ref})/V^{ref}$, where V^q stands for the Voronoï index for method q and V^{ref} stands for the same indicator for benchmarking. A significantly negative value reveals a better spatial dispersion. The figure shows that for all methods and sample sizes, the Voronoï index is significantly improved: by -60 to -70 % compared to the benchmark. Only the balanced moments method is less efficient.

For a given method and sample size, Figures 9 and 11 represent, as they do for the Voronoï index, the decrease of variance of a variable of interest, in comparison to the benchmark. This decrease is related to the intensity of the spatial autocorrelation of the variable dilated by inclusion probabilities.

For the methods shown in figure 9, i.e most of the methods studied, the gain in terms of variance is all the greater as the variable is spatially autocorrelated. However, this result is clearer with equal probabilities (9a) than with unequal probabilities (9b). These methods are equivalent in terms of gain. Consequently, Poisson sampling, local pivotal, local cube and determinantal designs on ordered file (TSP or Hamilton ordering) almost halve the variance of the sample, for the most autocorrelated variables and for $m = 100$. Furthermore, for all methods represented in Figure 9, the relative gain in variance is all the greater as the sampling rate is higher. Figure 10 illustrates this result for determinantal plans with equal probabilities.

The four methods shown in red and blue in Figure 11 differ from the others because of their results:

- the cube method balanced on first and second-order spatial moments (x, y, $x*y$, x^2 and y^2, where x and y are spatial coordinates) is less effective in terms of gain in variance. Calibrating with the inertia of the total population finally reproduces in the sample the groupings and repulsions of units. That goes against the desired sample dispersion principle;

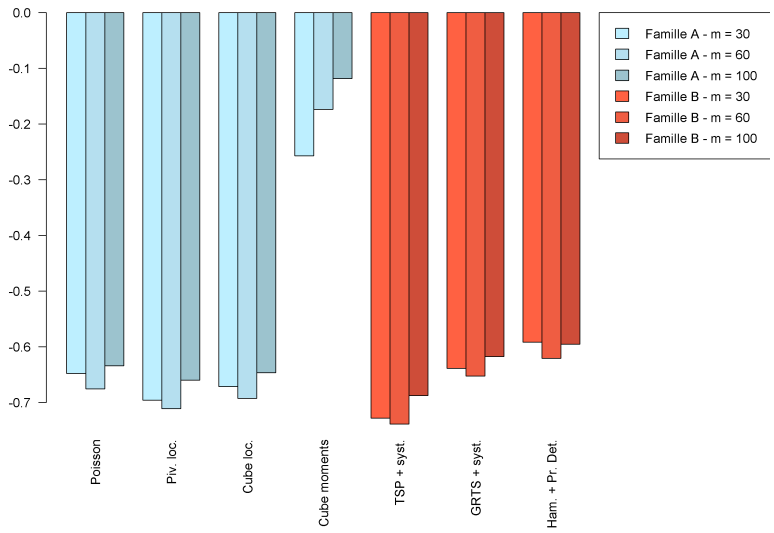
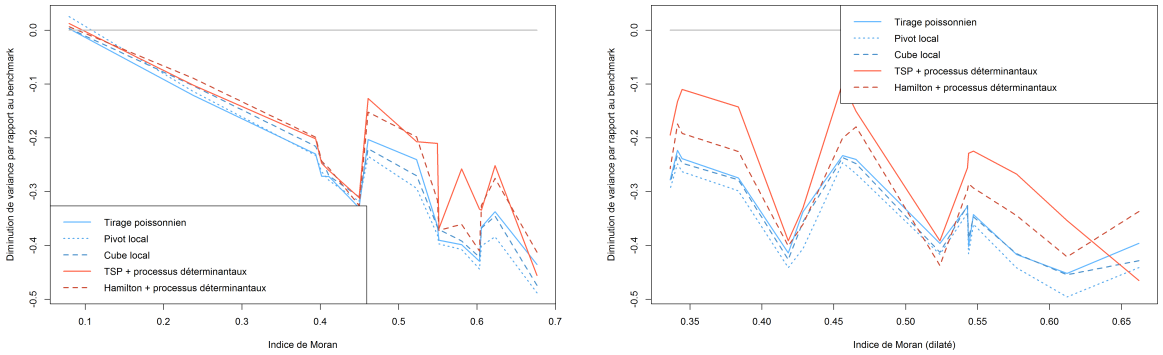


Figure 8: $(V^q - V^{ref})/V^{ref}$, where V^q is the Voronoï index for method q and V^{ref} for benchmark, for different values of m (example offered by equal probabilities).
Note: For a sampling with 30 PSUs using the Poisson sampling method with equal probabilities, the Voronoï indicator (averaged over 10 000 replications) is 65 % less than the single random sampling (benchmark).
Source :*INSEE, Fideli 2015.*



(a) Constant probabilities

(b) Unequal probabilities

Figure 9: Decrease in variance vs. benchmark for different methods, according to the spatial autocorrelation index of the variable (example with $m = 60$).
Note: Each curve stands for a spatial sampling method, and each point of the curve reflects 10 000 samples taken using the same method. The variation in variance of a given variable relative to a benchmark (in percentage) is shown, depending on the variable's level of spatial autocorrelation. For example, for an equal probabilities sample of 60 PSUs with the Poisson sampling method, the variance of the "number of women " variable (Moran's I =0.24) is 11 % less than with SRS.
Source :*INSEE, Fideli 2015.*

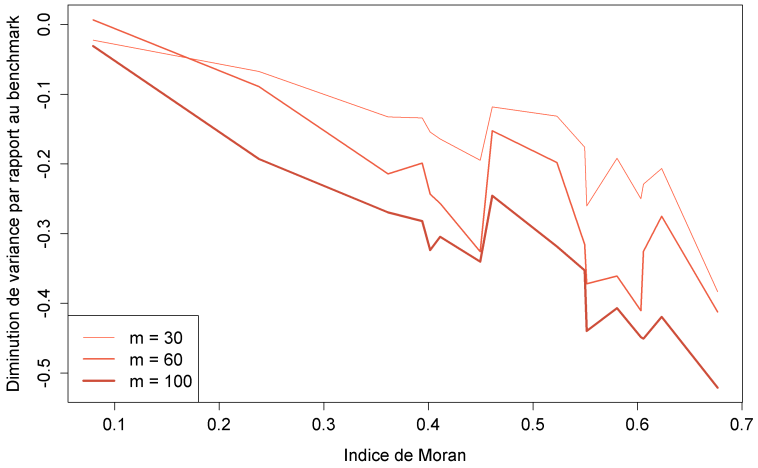


Figure 10: Reductions in variance vs. benchmark according to the spatial autocorrelation index of the variable, for different values of m (e.g., determinantal sampling with Hamilton ordering processes and equal probabilities).
Source : INSEE, Fideli 2015.
Note : When using determinantal sampling with equal probabilities, the variance of the variable "number of women " (Moran's I = 0.24) is decreased by 6.7 % for a sample of 30 PSUs, by 8.9 % for a sample of 60 PSUs, and by 19.3 % for a sample of 100 PSUs, compared to simple random sampling.

- file ordering (TSP, Hamilton or GRTS) followed by a systematic sampling, yields more erratic results than other methods in the same family. Entropy [8] of the systematic survey design is very weak, and this is even more the case on a uniquely sorted file. The number of potential samples with this method is M/m, explaining why curves in figure 11 look less smooth and why it is more difficult to draw conclusions. However, these methods still perform very well in terms of sample dispersion. In particular, the TSP ordering followed by systematic sampling is the one that reduces the Voronoï indicator the most (figure 8). It is also the one that most reduces the variance of the variables most spatially autocorrelated. GRTS ordering, meanwhile, is less efficient, due to lower ordering quality (the total length of the path obtained with GRTS is almost twice as long as the TSP or Hamilton path, see figure (7)).

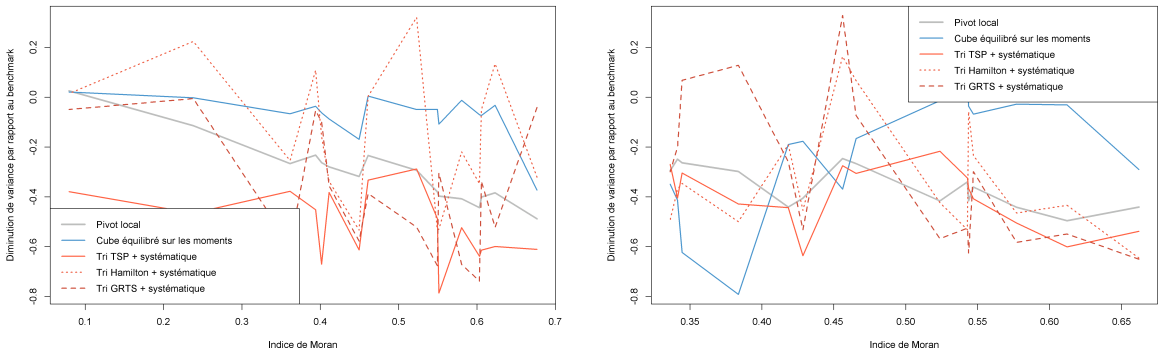
Conclusion

The creation of samples from a georeferenced sampling frame offers a possible new context for judicious mobilisation of geographical information. This chapter has presented various methods using this information at different stages of the sampling design process. We have carried out some tests based on real data, aiming at comparing these methods with traditional or original precision indicators, and testing different sets of parameters. The large majority of suggested methods prove to be effective in that they yield accurate estimates, even though the systematic sampling methods appear less effective. The statistical efficiency of a spatial sampling method increases with the level of spatial autocorrelation in the variable of interest to be estimated.

References

Ardilly, P. (2006). *Les techniques de sondage.* Editions Technip.

[8] Entropy is a measure of disorder. A high-entropy design enables a large number of samples to be selected and therefore leaves significant room for randomness



(a) Equal probabilities

(b) Unequal probabilities

Figure 11: Reductions in variance vs. *benchmark* for different methods, according to the spatial autocorrelation index of the variable (example with $m = 60$).
Note : The figure's local pivotal method 9 is represented in a grey line for comparison purposes.

Benedetti, R., Piersimoni, F., and Postiglione, P. (2015). *Sampling Spatial Units for Agricultural Surveys*. Springer.

Bondesson, L. and Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics*, 35(3):466–483.

Brewer, K. (1963). A model of systematic sampling with unequal probabilities. *Australian & New Zealand Journal of Statistics*, 5(1):5–13.

Chauvet, G. and Le Gleut, R. (2017). Asymptotic results for pivotal sampling with application to spatial sampling. *Work in progress*.

Connor, W. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61(314):384–390.

Deville, J.-C. and Tille, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85(1):89–101.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika*, 91(4):893–912.

Dickson, M. M. and Tillé, Y. (2016). Ordered spatial sampling by means of the traveling salesman problem. *Computational Statistics*, pages 1–14.

Ganganath, N., Cheng, C.-T., and Chi, K. T. (2014). Data clustering with cluster size constraints using a modified k-means algorithm. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2014 International Conference on*, pages 158–161. IEEE.

Grafström, A. (2012). Spatially correlated poisson sampling. *Journal of Statistical Planning and Inference*, 142(1):139–147.

Grafström, A. and Lisic, J. (2016). Balancedsampling: Balanced and spatially balanced sampling. *R package version*, 1(1).

Grafström, A., Lundström, N. L., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–520.

Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24(2):120–131.

Hidiroglou, M. and Gray, G. (1980). Algorithm as 146: Construction of joint probability of selection for systematic pps sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(1):107–112.

Hough, J. B., Krishnapur, M., Peres, Y., Virág, B., et al. (2006). Determinantal processes and independence. *Probab. Surv*, 3:206–229.

Lisic, J. (2015). *Parcel level agricultural land cover prediction*. PhD thesis, George Mason University.

Lister, A. J. and Scott, C. T. (2009). Use of space-filling curves to select sample locations in natural resource monitoring studies. *Environmental monitoring and assessment*, 149(1):71–80.

Loonis, V. (2009). La construction du nouvel échantillon de l'enquête emploi en continu à partir des fichiers de la taxe d'habitation. *JMS*, 2009:23.

Loonis, V. and Mary, X. (2018). Determinantal sampling designs. *Journal of Statistical Planning and Inference*.

Madow, W. G. (1949). On the theory of systematic sampling, ii. *The Annals of Mathematical Statistics*, pages 333–354.

Malinen, M. I. and Fränti, P. (2014). Balanced k-means for clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 32–41. Springer.

Marker, D. A. and Stevens Jr, D. L. (2009). Sampling and inference in environmental surveys. In *Handbook of Statistics*, volume 29, pages 487–512. Elsevier.

Pinciaro, S. J. (1978). An algorithm for calculating joint inclusion probabilities under pps systematic sampling. *of: ASA Proceedings of the Section on Survey Research Methods*, pages 740–740.

Stevens Jr, D. L. and Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465):262–278.

Tai, C.-L. and Wang, C.-S. (2017). Balanced k-means. In Nguyen, N. T., Tojo, S., Nguyen, L. M., and Trawiński, B., editors, *Intelligent Information and Database Systems*, pages 75–82, Cham. Springer International Publishing.