Properties of pivotal sampling with application to spatial sampling

Guillaume Chauvet (Ensai/Irmar)
Joint work with Ronan Le Gleut (Insee)

ITalian COnference on Survey Methodology (ITACOSM),
Florence
06/06/2019

Summary

There exists a large number of sampling algorithms, among which systematic sampling is probably the most famous (Madow, 1949 ; Tillé, 2006). It has found applications in a variety of fields.

Systematic sampling enjoys good practical properties, but suffers from a lack of randomness. Some common statistical properties are unlikely to hold, unless explicitly making strong model assumptions (which we try to avoid).

Pivotal sampling appears as a good alternative. While possessing also good practical properties, it introduces more randomness in the sample selection $\Rightarrow$ better statistical properties.

We consider an application for spatial sampling.

Some (short) reminders on sampling

Properties of pivotal sampling

Spatial sampling

Some (short) reminders on sampling

Notations

We are interested in a finite population of statistical units

$$U = \{1, \ldots, k, \ldots, N\}.$$

Denote by y a variable of interest taking the value y_k for some unit k, and t_y the total.

We note $\pi_k = Pr(k \in S) > 0$ the selection probability of some unit k. The sum $\sum_{k \in U} \pi_k \equiv n$ gives the average sample size.

By using a sampling design matching these inclusion probabilities, the total t_y is unbiasedly estimated by the Horvitz-Thompson (HT) estimator

$$\hat{t}_{y\pi} \quad = \quad \sum_{k \in S} \frac{y_k}{\pi_k}. \tag{1}$$

Algorithms of sampling

Large number of sampling algorithms matching a prescribed set of inclusion probabilities (see Tillé, 2006). We consider two of them : systematic sampling and pivotal sampling.

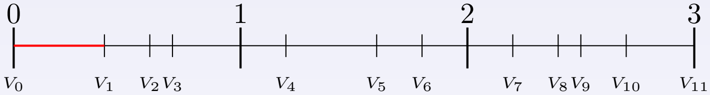
Systematic sampling (Madow, 1949) consists in randomly selecting a first unit, and then performing deterministic jumps to select the remaining units.

Pivotal sampling (Deville and Tillé, 1998 ; Srinivasan, 2001) is based on a principle of duels between units : the units fight, until one of them cumulates a sufficient probability so that a new selection is possible.

Systematic sampling on an example

Population U of size $N = 11$, with $n = 3$ and

$$\pi = (0.4 \;\; 0.2 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

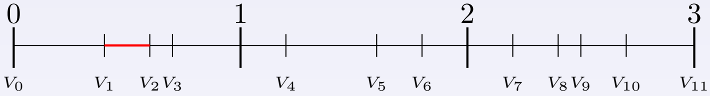


We represent the cumulated inclusion probabilities on a segment of length n. Each sub-segment represents one unit.

Systematic sampling on an example

Population U of size $N = 11$, with $n = 3$ and

$$\pi = (0.4 \;\; \color{red}{0.2} \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

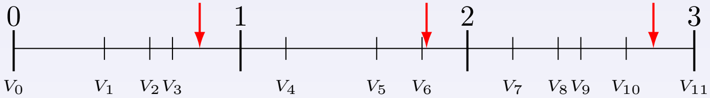


We represent the cumulated inclusion probabilities on a segment of length n. Each sub-segment represents one unit.

Systematic sampling on an example

Population U of size $N = 11$, with $n = 3$ and

$$\pi = (0.4 \;\; 0.2 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

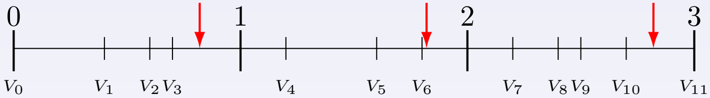


We represent the cumulated inclusion probabilities on a segment of length n. Each sub-segment represents one unit.
The sample is obtained through a random start $u \sim U[0, 1]$, followed by jumps of length 1.

Systematic sampling on an example

Population U of size $N = 11$, with $n = 3$ and

$$\pi = (0.4 \quad 0.2 \quad 0.1 \quad 0.5 \quad 0.4 \quad 0.2 \quad 0.4 \quad 0.2 \quad 0.1 \quad 0.2 \quad 0.3)^\top .$$



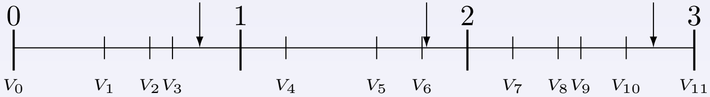
We represent the cumulated inclusion probabilities on a segment of length n. Each sub-segment represents one unit.
The sample is obtained through a random start $u \sim U[0,1]$, followed by jumps of length 1.

$$\begin{aligned}
u = 0.82 \in [V_3, V_4] &\quad \Rightarrow \quad \text{unit } 4 \text{ selected,}\\
1 + u = 1.82 \in [V_6, V_7] &\quad \Rightarrow \quad \text{unit } 7 \text{ selected,}\\
2 + u = 2.82 \in [V_{10}, V_{11}] &\quad \Rightarrow \quad \text{unit } 11 \text{ selected.}
\end{aligned}$$

Systematic sampling on an example (2)

Population U of size $N = 11$, with $n = 3$ and

$$\pi = (0.4 \;\; 0.2 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



Very simple method, sequential, matching exactly the π_k's. Extensively used in surveys and in spatial sampling (Thompson, 2002 ; Ripley, 2004).

One unit selected per **microstratum** $\Rightarrow$ stratification effect.

Avoids the selection of neighbouring units $\Rightarrow$ well-spread sample.

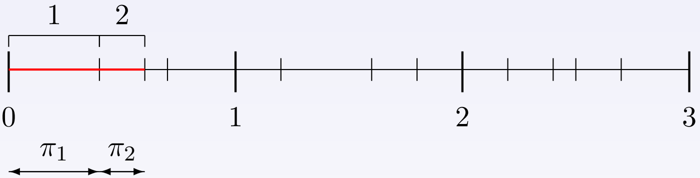
Drawbacks :

- unefficient if the variable of interest exhibits some periodicity,
- very few randomness $\Rightarrow$ limited statistical properties.

Pivotal sampling on an example

Population U of size $N = 11$, with $n = 3$ and

$$\pi = (0.4 \;\; 0.2 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

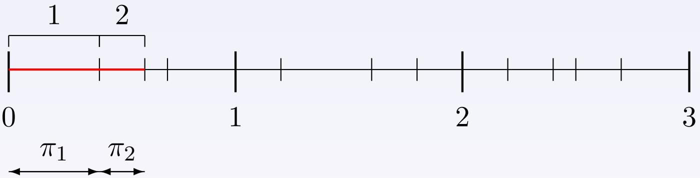


$$(\pi_1, \pi_2) = (0.4, 0.2) = \begin{cases} (0.6, 0) & \text{with proba } 0.4/0.6, \\ (0, 0.6) & \text{with proba } 0.2/0.6 \end{cases}$$

Pivotal sampling on an example

Population U of size $N = 11$, with $n = 3$ and

$$\pi = (\color{red}{0.4} \;\; \color{red}{0.2} \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



$$(\pi_1, \pi_2) = (0.4, 0.2) = \begin{cases} (0.6, 0) & \text{with proba } 0.4/0.6, \\ (0, 0.6) & \text{with proba } 0.2/0.6 \end{cases}$$

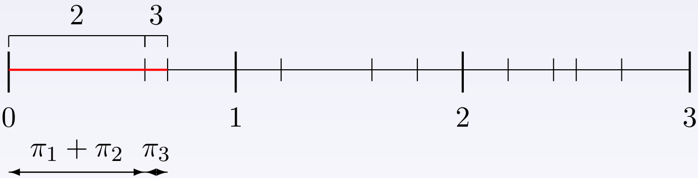
If unit 2 survives, we get

$$\pi^{(1)} = (\color{red}{0} \;\; \color{red}{0.6} \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

Pivotal sampling on an example (2)

Population U of size $N = 11$, with $n = 3$.

$$\pi^{(1)} = (0 \;\; 0.6 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

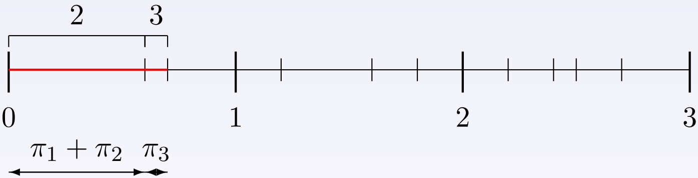


$$(\pi_2^{(1)}, \pi_3^{(1)}) = (0.6, 0.1) = \begin{cases} (0.7, 0) & \text{with proba } 0.6/0.7, \\ (0, 0.7) & \text{with proba } 0.1/0.7 \end{cases}$$

Pivotal sampling on an example (2)

Population U of size $N = 11$, with $n = 3$.

$$\pi^{(1)} \;=\; (0 \;\; \textcolor{red}{0.6} \;\; \textcolor{red}{0.1} \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



$$(\pi_2^{(1)}, \pi_3^{(1)}) = (0.6, 0.1) = \begin{cases} (0.7, 0) & \text{with proba } 0.6/0.7, \\ (0, 0.7) & \text{with proba } 0.1/0.7 \end{cases}$$

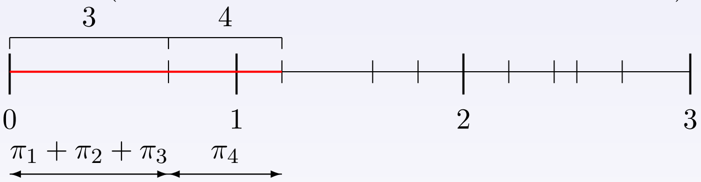
If unit 3 survives, we get

$$\pi^{(2)} \;=\; (0 \;\; \textcolor{red}{0} \;\; \textcolor{red}{0.7} \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

Pivotal sampling on an example (3)

Population U of size $N = 11$, with $n = 3$ and

$$\pi^{(3)} = (0 \;\; 0 \;\; \color{red}{0.7} \;\; \color{red}{0.5} \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

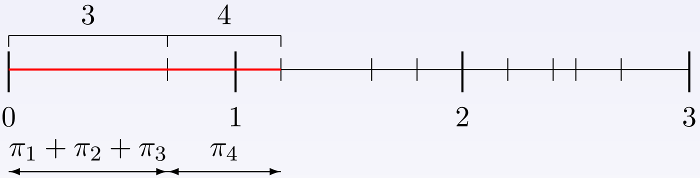


$$(\pi_3^{(2)}, \pi_4^{(2)}) = (0.7, 0.5) = \begin{cases} (1, 0.2) & \text{with proba } 0.5/(2-1.2), \\ (0.2, 1) & \text{with proba } 0.3/(2-1.2) \end{cases}$$

Pivotal sampling on an example (3)

Population U of size $N = 11$, with $n = 3$ and

$$\pi^{(3)} = (0 \;\; 0 \;\; \color{red}{0.7} \;\; \color{red}{0.5} \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



$$(\pi_3^{(2)}, \pi_4^{(2)}) = (0.7, 0.5) = \begin{cases} (1, 0.2) & \text{with proba } 0.5/(2-1.2), \\ (0.2, 1) & \text{with proba } 0.3/(2-1.2) \end{cases}$$

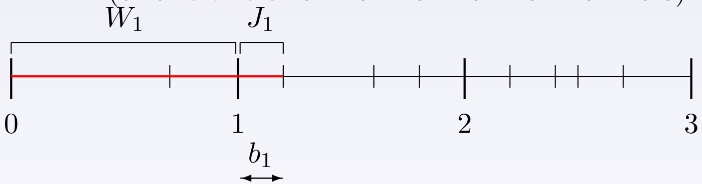
If unit 3 wins, we get

$$\pi^{(3)} = (0 \;\; 0 \;\; \color{red}{1} \;\; \color{red}{0.2} \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top , \ldots$$

Pivotal sampling on an example (4)

Population U of size $N = 11$, with $n = 3$ and

$$\pi^{(3)} = (0 \;\; 0 \;\; 0.7 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



$$(\pi_3^{(2)}, \pi_4^{(2)}) = (0.7, 0.5) = \begin{cases} (1, 0.2) & \text{with proba } 0.5/(2-1.2), \\ (0.2, 1) & \text{with proba } 0.3/(2-1.2) \end{cases}$$

If unit 3 wins, we get

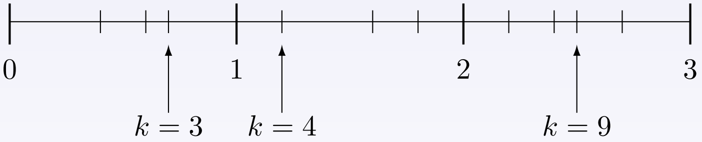
$$\pi^{(3)} = (0 \;\; 0 \;\; 1 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top , \ldots$$

Unit 3 is the first winner (W_1). Unit 4 is the first jumper (J_1).

Pivotal sampling on an example (5)

Population U of size $N = 11$, with $n = 3$ and

$$\pi = (0.4 \;\; 0.2 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



Simple method, sequential, matching exactly the π_k's.
One unit selected per **microstratum** $\Rightarrow$ stratification effect.
Avoids the selection of neighbouring units $\Rightarrow$ well-spread sample.
More randomness $\Rightarrow$ good statistical properties.
Particular case of the cube method (Deville and Tillé, 2004).

Properties of pivotal sampling

Asymptotic set-up and assumptions

Asymptotic set-up of Fuller (2011) : U belongs to a nested sequence of populations of size $N \to \infty$.

H1 : Non-degenerate : There exists some $0 < f_0 \leq f_1 \leq 1$ s.t.

$$f_0 \frac{n}{N} \leq \pi_k \leq f_1 \text{ for any } k \in U.$$

H2 : Finite moment of order 4 : $\exists C_1$ s.t.

$$\sum_{k \in U} \pi_k \left(\frac{y_k}{\pi_k} - \frac{t_y}{n} \right)^4 \leq C_1 \frac{N^4}{n^3}$$

$$\left[\Leftrightarrow \frac{1}{N} \sum_{k \in U} \left(y_k - \frac{t_y}{N} \right)^4 \leq C_1 \text{ if all } \pi_k' s = \frac{n}{N}. \right]$$

H3 : Non-vanishing variance within microstrata : $\exists C_2 > 0$ s.t.

$$\sum_{i=1}^{n} \sum_{k \in U_i} \alpha_{ik} \left(\frac{y_k}{\pi_k} - \sum_{l \in U_i} \alpha_{il} \frac{y_l}{\pi_l} \right)^2 \geq C_2 \frac{N^2}{n}.$$

Properties of pivotal sampling

The HT-estimator is mean-square consistent for the total (Chauvet, 2017) :

$$E_p\left[\left\{N^{-1}\left(\hat{t}_{y\pi}-t_y\right)\right\}^2\right] \;=\; O(n^{-1}).$$

The estimator $\hat{t}_{y\pi}$ is asymptotically normally distributed (Chauvet and Le Gleut, 2019) :

$$\frac{\hat{t}_{y\pi}-t_y}{\sqrt{V_p(\hat{t}_{y\pi})}} \;\rightarrow_{\mathcal{L}}\; \mathcal{N}(0,1).$$

Problem : design-unbiased variance estimation is not possible, but we can produce a conservative variance estimator (CLG, 2019)

$$\begin{aligned} E_p\{v_{DIFF}(\hat{t}_{y\pi})\} &\geq V_p(\hat{t}_{y\pi}), \\ \text{with } v_{DIFF}(\hat{t}_{y\pi}) &\simeq \sum_{i=1}^{n/2}\left(\frac{y_{W_{2i}}}{\pi_{W_{2i}}}-\frac{y_{W_{2i-1}}}{\pi_{W_{2i-1}}}\right)^2. \end{aligned}$$

Spatial sampling

Working model

In a context of spatial sampling, first law of geography of Tobler : "Everything is related to everything else, but near things are more related than distant things".

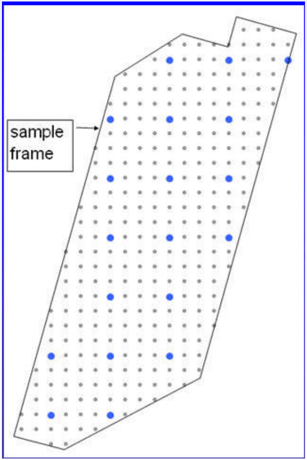
Working model of type (see Grafström and Tillé, 2013) :

$$\begin{array}{rcl} y_k & = & \beta\pi_k + \epsilon_k, \\ E_m(\epsilon_k) = 0 & \text{et} & Cov_m(\epsilon_k, \epsilon_l) = \sigma_k \sigma_l \rho^{d(k,l)}. \end{array}$$

$\Rightarrow$ better to avoid selecting neighbouring units, which carry a similar information.
$\Rightarrow$ better to spread well the sample over space.

More auxiliary information may be available, resulting in more efficient sampling strategies (Grafström and Tillé, 2013).

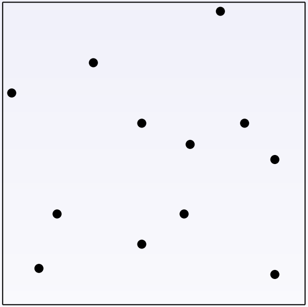
Systematic sampling on a regular grid



- A regular grid is randomly placed on the area under study.
- A sample of points is selected on the grid via systematic sampling.
- The sample is spread over space, but we may face some unexpected periodicity.

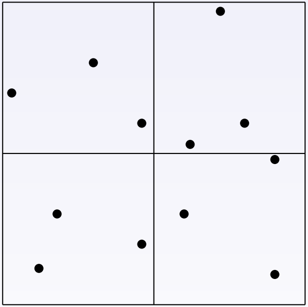
Generalized Random Tesselation Sampling (GRTS)

Stevens and Olsen (2004)



Generalized Random Tesselation Sampling (GRTS)

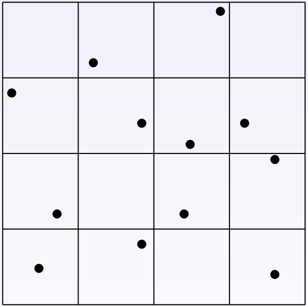
Stevens and Olsen (2004)



- Tesselation of the area on a regular grid, with "addresses".

Generalized Random Tesselation Sampling (GRTS)

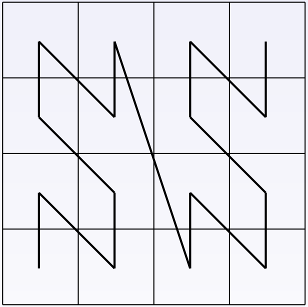
Stevens and Olsen (2004)



- Tesselation of the area on a regular grid, with "addresses".

Generalized Random Tesselation Sampling (GRTS)

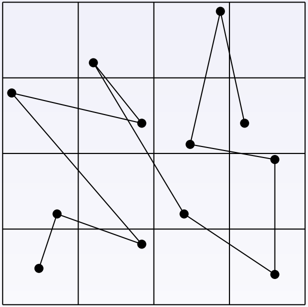
Stevens and Olsen (2004)



- Tesselation of the area on a regular grid, with "addresses".
- The addresses are ranked on a line.

Generalized Random Tesselation Sampling (GRTS)

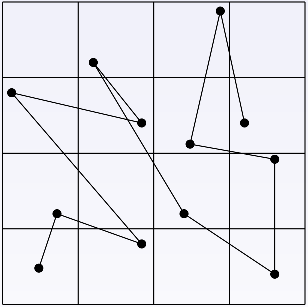
Stevens and Olsen (2004)



- Tesselation of the area on a regular grid, with "addresses".
- The addresses are ranked on a line.

Generalized Random Tesselation Sampling (GRTS)

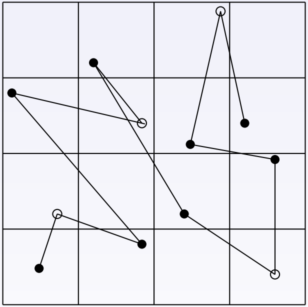
Stevens and Olsen (2004)



- Tesselation of the area on a regular grid, with "addresses".
- The addresses are ranked on a line.
- Sample selection on the line via systematic sampling after (partial) randomization.

Generalized Random Tesselation Sampling (GRTS)

Stevens and Olsen (2004)



- Tesselation of the area on a regular grid, with "addresses".
- The addresses are ranked on a line.
- Sample selection on the line via systematic sampling after (partial) randomization.

Pivotal Tesselation Method

The GRTS method gives samples well spread over space (Stevens and Olsen, 2004), but with systematic sampling the study of the statistical properties of the HT-estimator is made difficult (and not sure to hold), even with a partial randomization.

We propose to use the tesselation method, but by replacing systematic sampling by pivotal sampling. This leads to the Pivotal Tesselation Method (PTM).

The sample is still well spread over space + HT-estimator consistent and asymptotically normal.

Alternatively, pivotal sampling can be used with any spatial sampling design with some form of ranking on units (e.g., Dickson and Tillé, 2016).

A small simulation study

Example 5 of Grafström et al. (2012). Divide the unit square according to a 20×20 grid $\Rightarrow$ population of $N = 400$ units.
Variable $y_k \equiv$ area within the cell under $f(x1, x2) = 3(x1 + x2) + \sin\{6(x1 + x2)\}$.

Samples of size $n = 16, 32$ or 48 with equal probabilities. Spatial sampling designs :

- pivotal tesselation method (PTM),
- generalized random tesselation sampling (GRTS),
- local pivotal methods (LPM1 and LPM2 ; Grafström et al., 2012).
- pivotal method through Traveling Salesman Problem order (TSP, Dickson and Tillé, 2016).
- simple random sampling (SRS).

Computation of an indicator of spatial balance (Voronoi polygons) + variance associated to each sampling strategy.

Results

Table – Monte Carlo Mean of the spatial balance and Monte Carlo Variance of the Horvitz-Thompson estimator for Population 1

	PTM	GRTS	LPM1	LPM2	TSP	SRS
	$E_{MC}(\Delta)$					
$n = 16$	0.07	0.12	0.08	0.09	0.11	0.33
$n = 32$	0.08	0.11	0.07	0.07	0.10	0.30
$n = 48$	0.09	0.11	0.07	0.07	0.10	0.29
	$V_{MC}(\hat{t}_{y\pi})$ $(\times 100)$					
$n = 16$	1.53	2.49	1.94	1.96	2.65	12.48
$n = 32$	0.39	0.89	0.54	0.57	0.65	6.18
$n = 48$	0.16	0.34	0.26	0.27	0.28	3.91

Future work

Pivotal sampling is a particular case of the cube method (Deville and Tillé, 2004), which enables to select balanced samples. A sampling design is balanced on a set x_k of auxiliary variables if

$$\hat{t}_{x\pi}(s) = t_x \quad \text{for all } s \text{ such that} \quad p(s) > 0.$$

Other spatial sampling methods introduce more complex dependencies in the selection of units :

- local pivotal method (Grafström et al., 2012) : at each step of the pivotal method, the 2 nearest remaining units are treated.
- local cube method (Grafström and Tillé, 2013) : at each step of the cube method, the $p+1$ nearest remaining units are treated.

Similar statistical properties are needed, but these sampling designs are more difficult to study.