Why geocoding a sampling frame might prove useful ?

Household Surveys in a Changing Data Landscape Challenges, Opportunities and an Agenda for the Future

28 February 2020, New York

Vincent Loonis

February 28, 2020

Insee
Mesurer pour comprendre

Background

- Over the last few years a lot of initiatives have been undertaken to better integrate geospatial and statistical information, including within the scope of sampling frames.
- Insee, the French National Institute of Statistics and Economic Studies, has benefited from these initiatives to try and build a fully geocoded statistical information system.
- Insee has also wanted to contribute more actively to the global or regional reflexions, while taking advantage of one of its specific features: Economic studies.
- This is the starting point of a Handbook of Spatial Analysis.

Insee
Mesurer pour comprendre

The handbook of spatial analysis

- The handbook results from a collaboration between Insee, Eurostat and the EFGS (European Forum for Geography and Statistics). It is available for free (in French or English) on their websites.
- The handbook raises (and tries to answer) questions such as:
  - What use should be done of geocoded data sources ?
  - When should spatial location be taken into account ?
  - How spatial statistical methods should be applied ?
- The handbook consists of 14 chapters written by 24 authors.
- The bulk of the chapters presents already well-documented methods (spatial econometrics, smoothing) while drawing on their implementation in R.
- Some of the chapters are specifically devoted to NSIs: spatial sampling, small areas estimation and spatial correlation, confidentiality of spatial data.
- One chapter raises the question of spatial econometrics on survey data.

Insee
Mesurer pour comprendre

Spatial sampling

1. The chapter on *Spatial sampling* provides guidelines to:
   - Build Primary Statistical Units (PSU) that are relevant for the field work while preserving the accuracy of the estimations.
   - Spread a sample over a territory in order to improve the accuracy of the estimations.
2. The chapters on *small area estimation and spatial correlation*, or that on spatial econometrics on survey data show that, once the sample is drawn, it might be difficult to implement the various methods of spatial statistics.

Insee
Mesurer pour comprendre

Building Primary Statistical Units

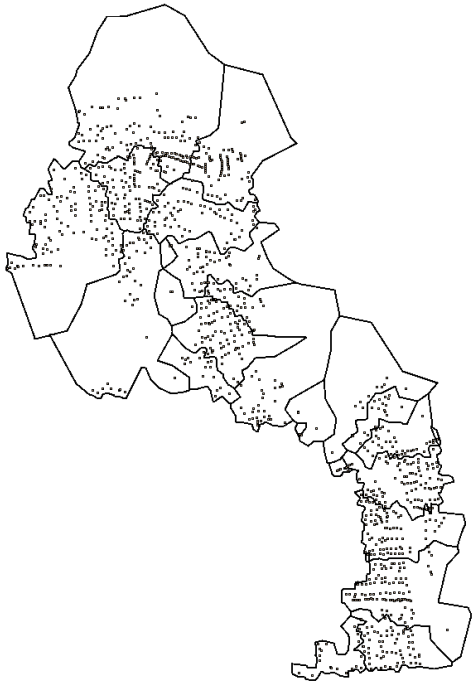
- For face-to-face surveys, the sampling design is often a 2-stage sampling design.
- The first step consists in selecting a sample of PSUs.
- Building and selecting the PSUs is often a trade-off between field work and accuracy requirements.
  - To help organise the field work,
    - A PSU is often assigned to one interviewer.
    - The geographical extent of the PSUs has to be as small as possible in order to reduce the travel costs (time and money).
    - The number of statistical units within a PSU has to be large enough to provide a sufficient work load.
  - To preserve the accuracy of the estimations,
    - The number of PSUs has to be as large as possible, and their size as small as possible to avoid a too strong *design effect*.
    - The size of the PSUs should remain constant to benefit from the so-called MINIMAX property (to avoid a major desaster during the estimation process).

Insee
Mesurer pour comprendre

Building PSUs

Building PSUs is equivalent to spatial clustering with size constraints.

(a) Creating a path passing through all the statistical units



(b) Following the path to build PSUs of 120 dwelling

Figure: Building PSUs for the French Labour Force Survey (LFS)

Spreading the sample

The First Law of Geography, according to Waldo Tobler, is everything is related to everything else, but near things are more related than distant things

+

The unwritten law of sampling: You shall avoid selecting statistical units that have the same features

=

The unwritten law of spatial sampling: You shall spread your sample as much as possible.

Insee
Mesurer pour comprendre

Spreading the sample

Spatial sampling has attracted a lot of attention over the last years. A lot of methods are available, that lead to the same conclusion: the gain in terms of variance is all the greater as the things are spatially related (autocorrelated).

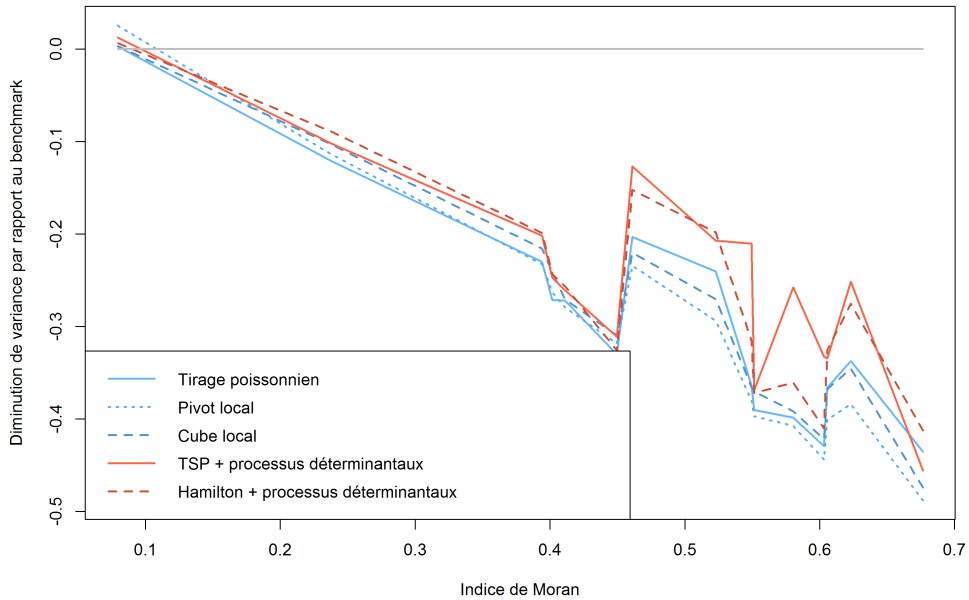


Figure: Gain in variance of estimation according to spatial autocorrelation (Moran's Index) and to the spatial sampling method. The reference line is the benchmark (SRS).

Thank you for your attention !!!

Figure: credit : Cyril Ruoso

Insee
Mesurer pour comprendre