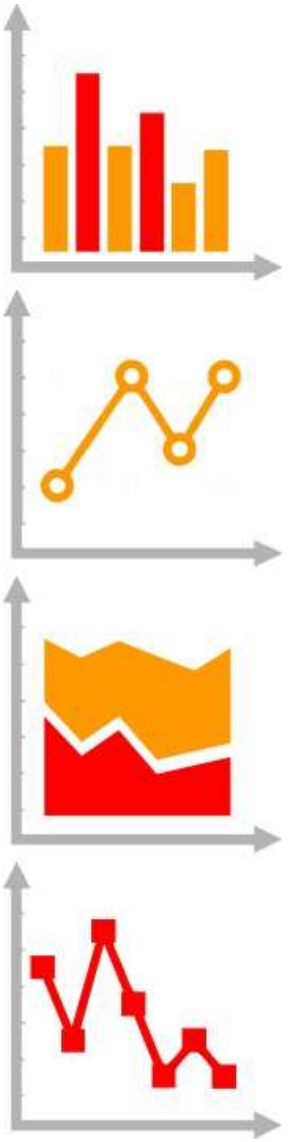
Sampling methodology for French social surveys



EU Twinning Project IL/12 CRIS 2015/370-467

Activity D7 : Integration of sampling, workload allocation, management and monitoring in a multi-field survey approach

Pierre-Arnaud Pendoli (INSEE)

Insee

Mesurer pour comprendre

Jerusalem
6 Feb 2018

Summary

Summary

Introduction

- National Institute of Statistics and Economic Studies (Insee) manages several social studies :
  - Quarterly :
    - Labour Force Survey ==> unemployment rate
    - Rent Survey ==> rent index
  - Annually :
    - Income and living conditions survey
    - Victimization survey
  - Not regular frequency :
    - Housing Survey
    - Adult Education Survey
    - Household Expenditure Survey
    - Working Conditions Survey …

Introduction

- In France, a large part of social surveys samples are drawn in a Master Sample (except for Labour Force Survey)

- French Master Sample is a set of areas sampled
  - to represent the best as possible the French population
  - to limit the costs for the collect of households surveys

- This sampling methodology is called “two stages sampling”
  - 1st stage : drawing the areas (primary sample unit = PSU)
  - 2nd stage : drawing the dwellings or the individuals in the PSU

Introduction

- Four principal issues :
  - What sampling frame do we use at Insee ?
  - How do we build our primary units ?
  - What sampling design do we choose for the 1st stage ?
  - What sampling design do we choose for the 2nd stage ?

Summary

Quality requirements for a sampling frame

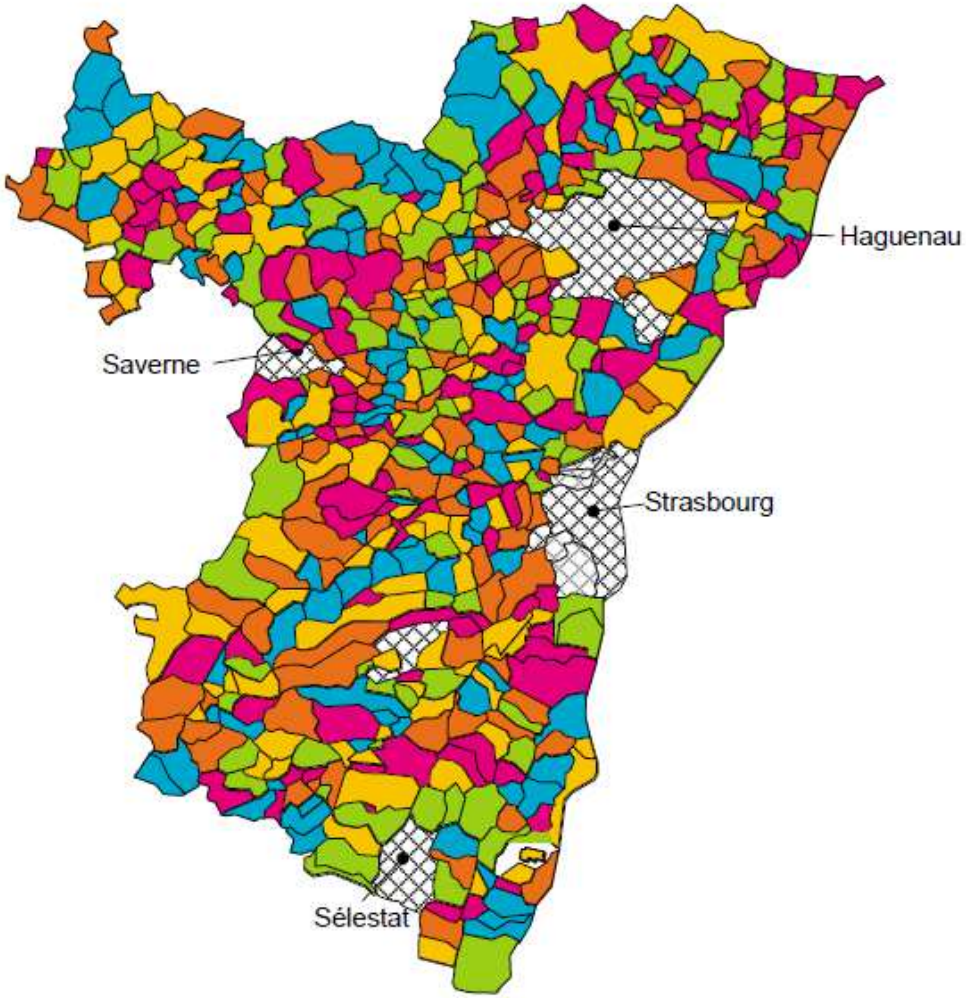
- To build a good sampling frame for a face-to-face survey we need
  - Concerning the units (dwellings and/or inhabitants) :
    - an exhaustive basis updated regularly
    - an identifier for each unit of the basis
  - Concerning their characteristics (variables) :
    - Geographical variables to locate precisely each unit
    - Social and demographic variables related to the theme of the survey. These variables can help us
      - to define the sampling design (by creating strata e.g.)
      - to improve the quality of the samples (by using calibrations methods e.g.)

An usual sampling frame : the census

- Annual census surveys :
  - Since 2004
  - Every year a dwelling sample is collected
  - Small localities (< 10,000 inhabitants) :
    - Random distribution between five rotative groups
    - Every year we collect all the localities of one of the groups
  - Big localities (> 10,000 inhabitants) :
    - In each city we get a dwellings register
    - In each city : random distribution of the dwellings between five rotative groups
    - Every year we draw 40% of the dwellings of one of the groups and we collect all the housings and the inhabitants of the sample

An usual sampling frame : the census

- Small localities : random distribution between 5 rotative groups



Exemple du Bas-Rhin

"communes 2004" (2009, 2014, ...)
(moins de 10 000 habitants)

"communes 2005" (2010, 2015, ...)
(moins de 10 000 habitants)

"communes 2006" (2011, 2016, ...)
(moins de 10 000 habitants)

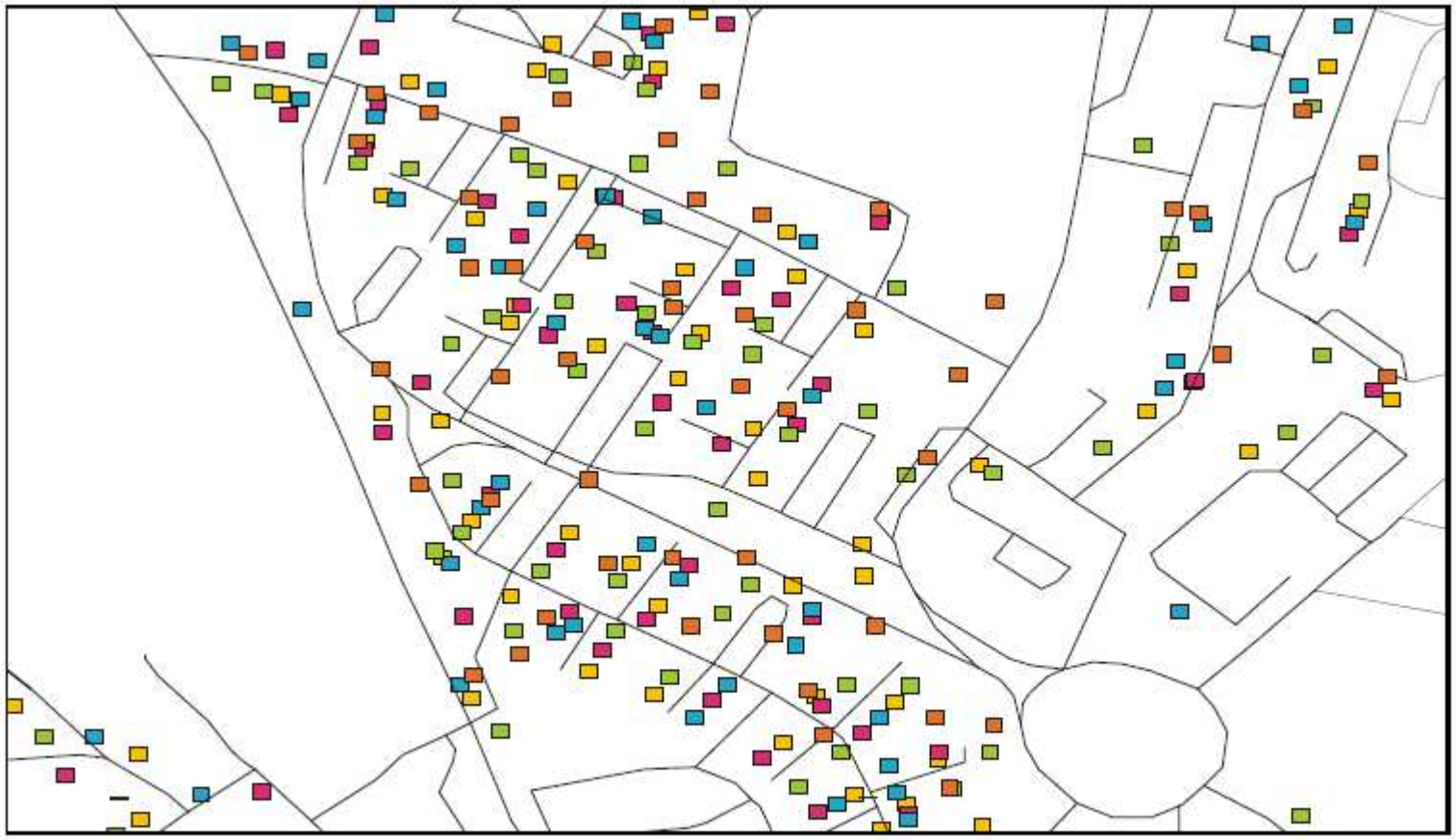
"communes 2007" (2012, 2017, ...)
(moins de 10 000 habitants)

"communes 2008" (2013, 2018, ...)
(moins de 10 000 habitants)

Communes concernées chaque année
(10 000 habitants ou plus)

An usual sampling frame : the census

- Big localities :
  - (1) Random distribution of the dwellings between 5 rotative groups



An usual sampling frame : the census

- Big localities :
  - (2) Random drawing of the dwellings in the annual rotative group

An usual sampling frame : the census

- Annual census surveys have been used since 2009 to draw the samples for the main social surveys

- But this sampling frame gets two disadvantages
  - It's not an exhaustive basis. At the end of a five year cycle :
    - all housings and inhabitants are collected in small cities ...
    - ... but only 40% in big cities due to annual dwellings sampling
  - The basis is only updated in fifths year by year, thanks to each new annual census survey

- Between 2013 and 2017, we collected about 47 millions inhabitants, that is about 72% of the French population

A new sampling frame : FIDELI

- Fideli : French acronym meaning “demographic file of housings and individuals”
  - Since 2012
  - It’s a register (contrary to the annual census which is an annual survey partly coming from a sampling)
  - This register is an aggregation of different tax files (housing taxes e.g.)
  - It contains two files :
    - A register of housings
    - A register of individuals
  - With social, demographic et geographic variables
  - These files are updated yearly thanks to annual tax returns

A new sampling frame : FIDELI

- Fideli checks the main quality requirements to be a good sampling frame
- Insee is beginning to develop a computing application to manage social surveys samples using Fideli as sampling frame : NAUTILE
- Nautile : French acronym meaning “New application used for housings and individuals surveys sampling”
- Nautile will replace the current application (Octopusse) that uses the annual census surveys as sampling frames
- From 2020 Fideli will be used as the main sampling frame for all the social surveys at Insee

Summary

Finding the appropriate geographical level to create the PU

- When we create primary units (PU), we have to remember the following points :

  - Two stages sampling design is used for social survey to limit the interviewers' travel time

    ==> the PU have to be defined to a geographical point of view

    E.g. one or an aggregating of
    - cities
    - sub-municipal areas

Finding the appropriate geographical level to create the PU

- When we create primary units (PU), we have to remember the following points :
  - For the first stage sampling, we have to calculate some variables for each PU in order to respect the PU sampling design

    E.g. to stratify the PU we need to calculate the stratification variables for each PU

  - For the second stage sampling, we have to match PU and secondary units (housings/individuals)

    ==> We need a common identifier between the PU and the secondary units

- At Insee we build our PU from cities because the city identifier is known in all our files (e.g. Fideli)

Minimising PUs' extent under constraints

- Our aim : minimising PUs' extent to limit the interviewers' travel times
- PU have a minimal size, 2,500 main residences, in order to not re-interview a dwelling after it has already been drawn in a sample
- We want ideally build (intra) heterogeneous PU in order to limit the cluster effect among the sampled PU

Minimising PUs' extent under constraints

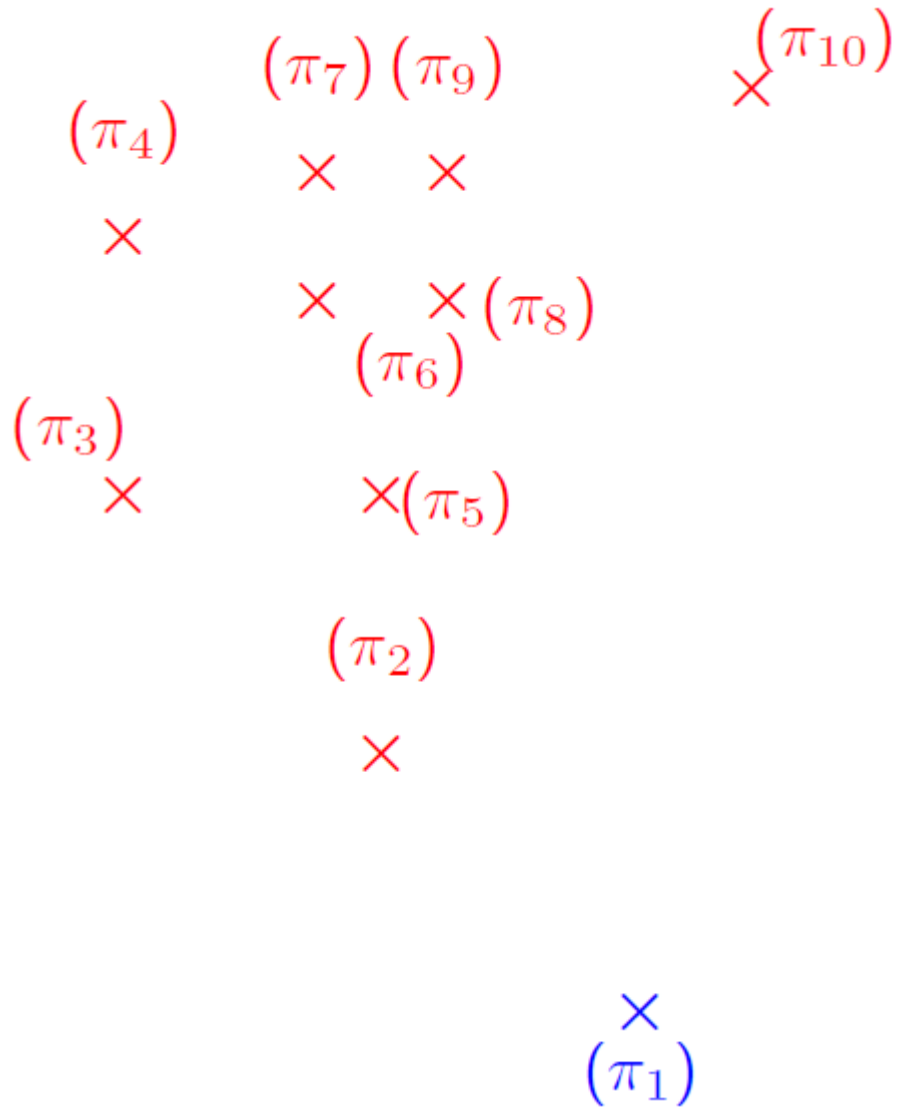
- To minimise the extent of the PU we use the Travelling salesman problem
  - The idea : this algorithm approximates the shortest path from a starting point to visit all the cities of a region (e.g. a district for Israel)
  - We create the set of PU in the following way :
    - We need the distance matrix by road between all the cities of the region
    - For each starting city, we get a path that is the solution to the algorithm
    - From a certain starting point we travel the cities located on the path – which is the solution – until respecting the minimal size of a PU (2,500 main residences) ==> we build the first PU
    - We keep going the travel to build the following PU

Minimising PUs' extent under constraints

- Graphic representation of the Travelling salesman problem

Minimising PUs' extent under constraints

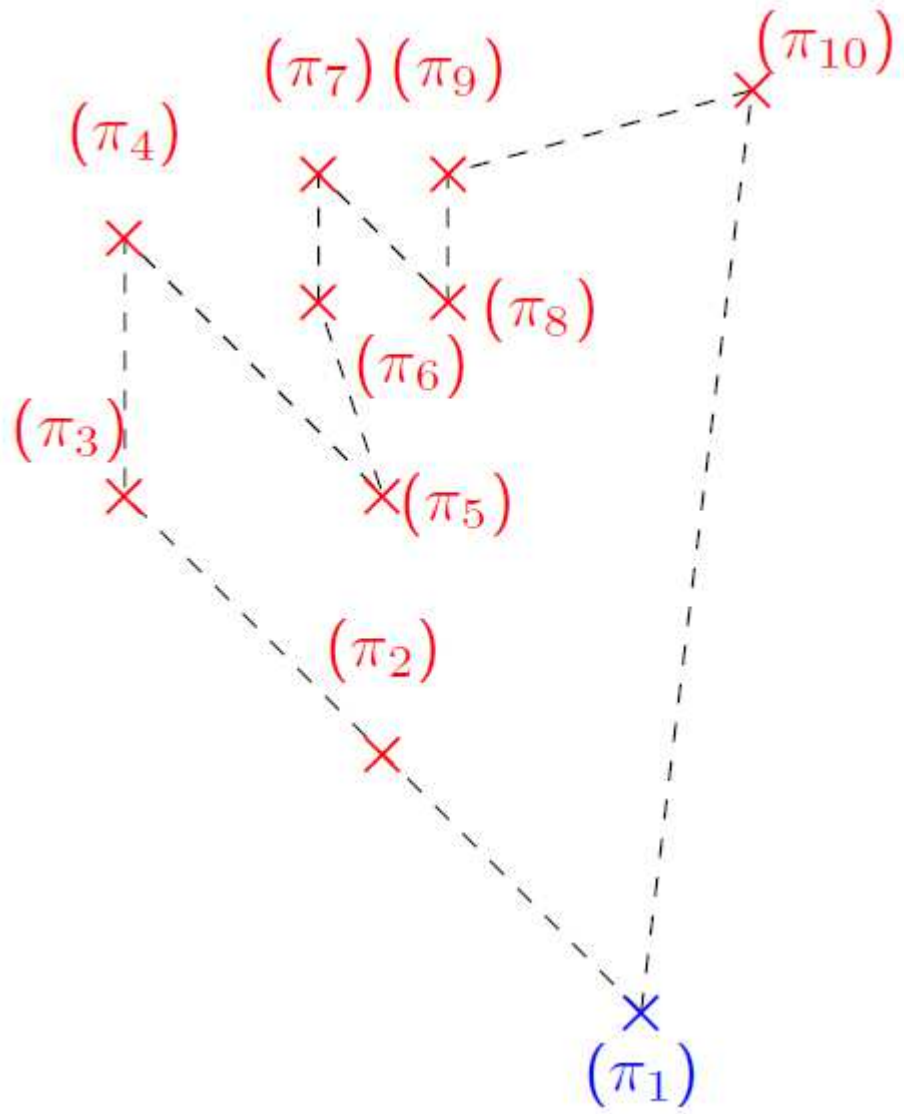
- Graphic representation of the Travelling salesman problem



Insee

Minimising PUs' extent under constraints

- Graphic representation of the Travelling salesman problem



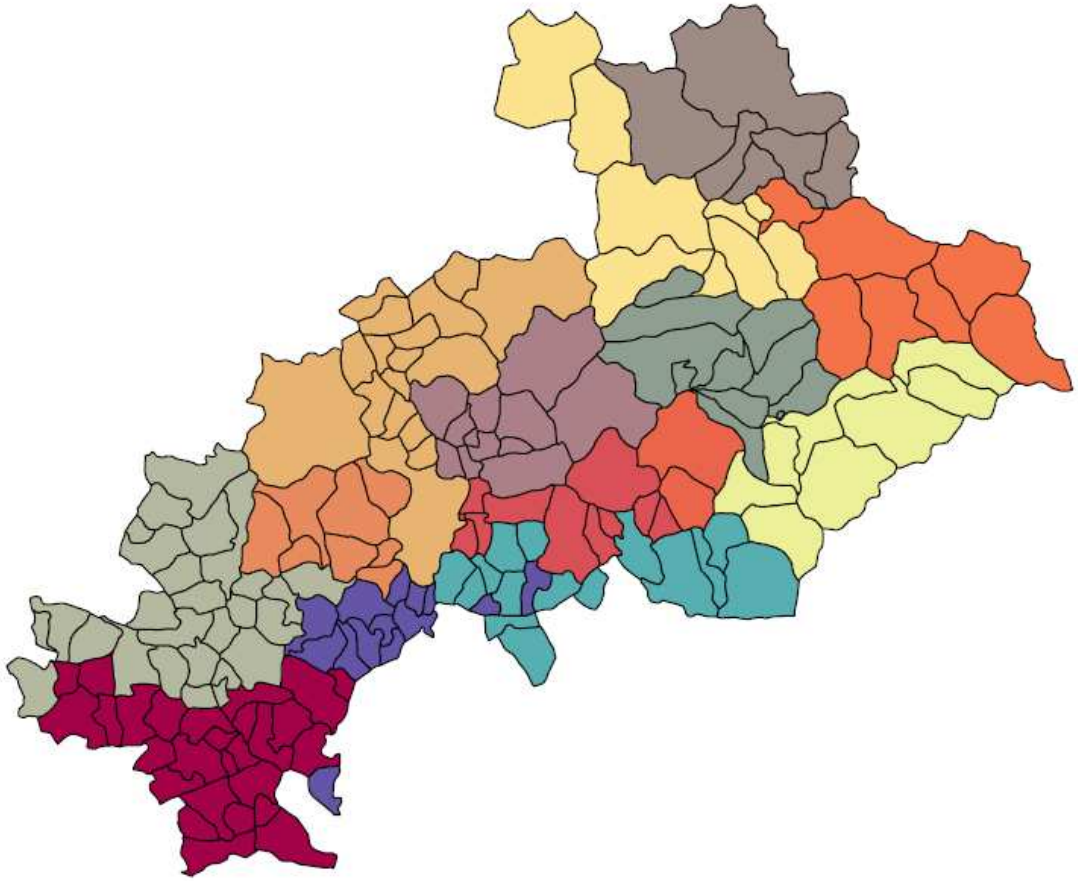
Insee

Minimising PUs' extent under constraints

- Using the Travelling salesman problem for the new French Master Sample :
  - Method implemented by Cyril Favre-Martinoz & Thomas Merly-Alpa (Insee)
  - Package TSP in R
  - In order to limit the dependence on the starting point and the resulting solution we make 1,000 realizations of the algorithm
    - For each realization, the starting point is got by a random drawing from the cities of the region
  - We get 1,000 sets of PU
  - We keep the set that minimises the average extent of the PU

Minimising PUs' extent under constraints

- Example for the French Master Sample : a French department is an administrative subdivision of a region
- This map represents the PU created for the French department Hautes-Alpes



Minimising PUs' extent under constraints

- Result : we built 5,128 PU which partition the French territory

- Remarks on the use of the Travelling salesman problem to build PU :
  - The regional level used to consider the cities for each realization of the algorithm needs to be small enough to enable a fast computation
    - The size of the distance matrix between the cities increases with the number of the cities of the region
  - In France, the road network is more optimized at the departmental level than at the regional level
    - The departmental borders are linked to natural borders

Summary

The PU sample size

- The first question we have to answer to draw the PU is the following : how much to draw of PU ?

- 1st principle : PU sample is perpetuated to draw the main social surveys samples for the decade 2020-2030

- 2nd principle : at Insee, Interviewer Action Areas (IAA) and PU partly overlap : except to big PU, Insee's policy is to consider each sampled PU as an IAA

- 3rd principle : PU are stratified by regions
  - Insee is structurally organized by regions
  - Regional directions manage the collect of Insee's social survey and the interviewers' workload

The PU sample size

- Firstly, we compute the number of interviewers for each region :
  - Let N_{reg} the number of main residences of a region
  - Let R be the average sampling rate for a survey sample (proportion of main residences interviewed). R = 1:2,000
  - Let I be the average number of interviews realized by an interviewer for a survey. I = 20

$$n_{\text{reg}}^{\text{interviewers}} = \frac{N_{reg}.\text{R}}{I}$$ is the number of interviewers to mobilize for each survey in the region

$$n_{\text{reg}}^{\text{interviewers}} = \frac{N_{reg}}{40{,}000}$$

The PU sample size

- Secondly, we compute the number of interviewers for each PU :
  - Considering that ***R*** = 1:2,000 is the average sampling rate for a survey sample and that ***I*** = 20 is the average number of interviews realized by an interviewer for a survey, we get that an area of 40,000 main residences is the coverage capability of an interviewer
  - Sampled PU for which $N_{PU} < 40{,}000$: one interviewer is assigned
  - Sampled PU for which $N_{PU} > 40{,}000$: we assign a number of interviewers equal to $\frac{N_{PU}}{40{,}000}$ rounded to the closest integer

The PU inclusion probabilities

- The second question we have to answer to draw the PU is the following : how to set the inclusion probabilities ?
- One of the aims of the surveys sampling is to minimize the sampling weights dispersion
- For that, we can use the **self-weighting two-stages sampling methodology**
- For the first stage sampling, it involves that we set the PU inclusion probabilities proportional to their sizes (number of dwellings, number of main residences …) since it is worth less than 1

The PU inclusion probabilities

- **Inclusion probabilities computation**

Let be a PU belonging to a French region ***reg***

1. Initialization : $n_{\text{reg}}^{\text{interviewers}} = \frac{N_{reg}}{40{,}000}$ and $rest_{\text{reg}} = N_{reg}$
2. PU are sorted by descending number of main residences
3. For each PU, do :
   - If $n_{\text{reg}}^{\text{interviewers}} \times \frac{N_{UP}}{rest_{reg}} \leqslant 1$ then do :

$$\left\{ \begin{array}{l} \pi_{UP} = n_{\text{reg}}^{\text{interviewers}} \times \frac{N_{UP}}{rest_{reg}} \\ n_{\text{reg}}^{\text{interviewers}} \leftarrow n_{\text{reg}}^{\text{interviewers}} - 1 \\ rest_{\text{reg}} = N_{reg} - N_{UP} \end{array} \right\}$$

   - Else do :

$$\left\{ \begin{array}{l} \pi_{UP} = 1 \\ n_{\text{reg}}^{\text{interviewers}} \leftarrow n_{\text{reg}}^{\text{interviewers}} - round\left(\frac{N_{PU}}{40{,}000}\right) \\ rest_{\text{reg}} = N_{reg} - N_{UP} \end{array} \right\}$$

The PU inclusion probabilities

- Results for the actual French Master Sample
  - We get 80 PU with $\pi_{UP}=1$
    - We can call them “certain PU”
    - About 130 interviewers are sharing the certain PU
  - We get 487 PU with $\pi_{UP}<1$
    - We can call them “probable PU”
    - 487 interviewers are sharing the probable PU
  - Current PU sample has been drawn in 2008
    - Social surveys samples (except for the Labour Force Survey) have been drawn in this Master Sample for ten years (2009-2019)
    - Interviewers have been hired in these areas for this period

The PU sampling design

- The third question we have to answer to draw the PU is the following : which sampling design to choose ?

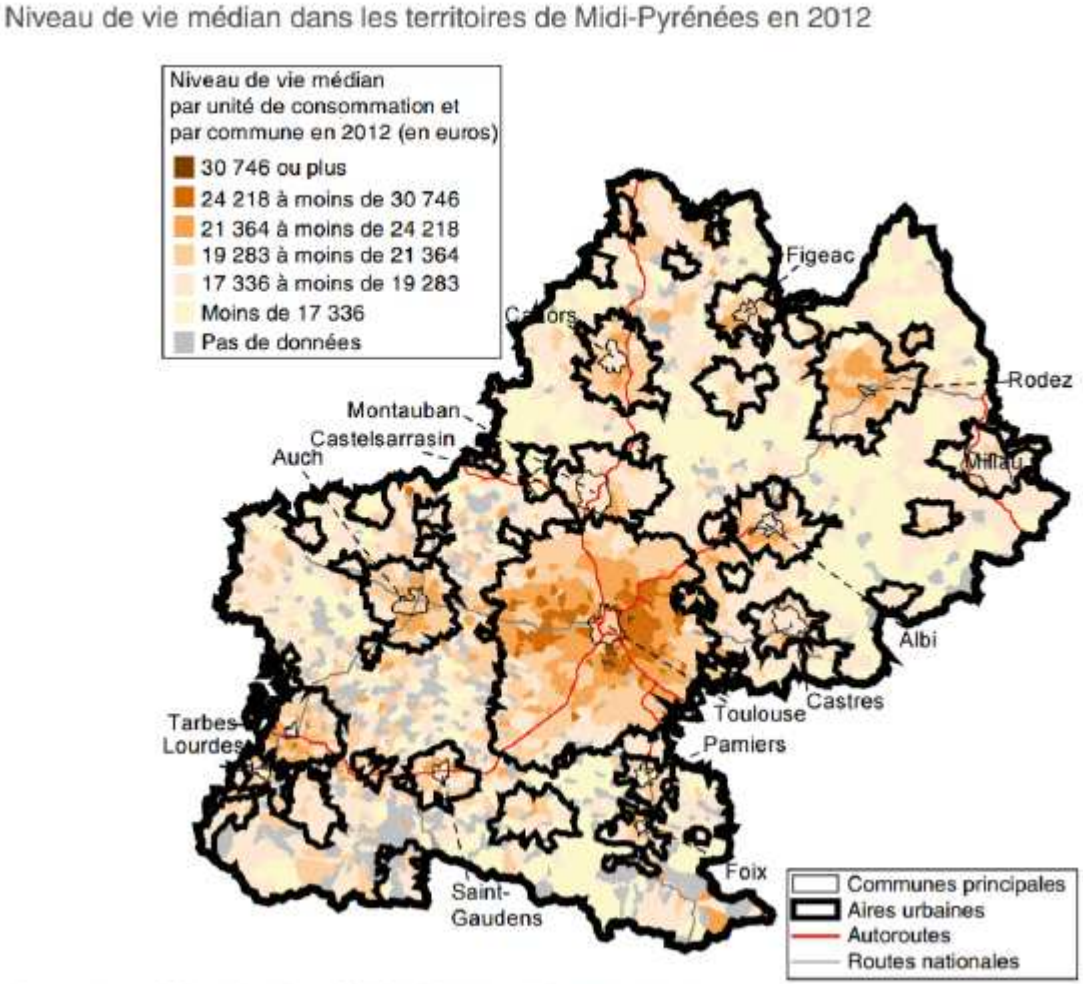
  - The sampling design has to respect the PU inclusion probabilities

  - We would like the PU sample to be as “representative” as possible of all the primary units

  - We would like the PU sample to be as geographically dispersed as possible to cover the entire French territory

The PU sampling design

- Example of the French region Midi-Pyrénées (illustrated by Cyril Favre-Martinoz - 2017)



The PU sampling design

- Example of the French region Midi-Pyrénées

  - In this example, we observe a strong spatial autocorrelation for the median income by unit of consumption and by city

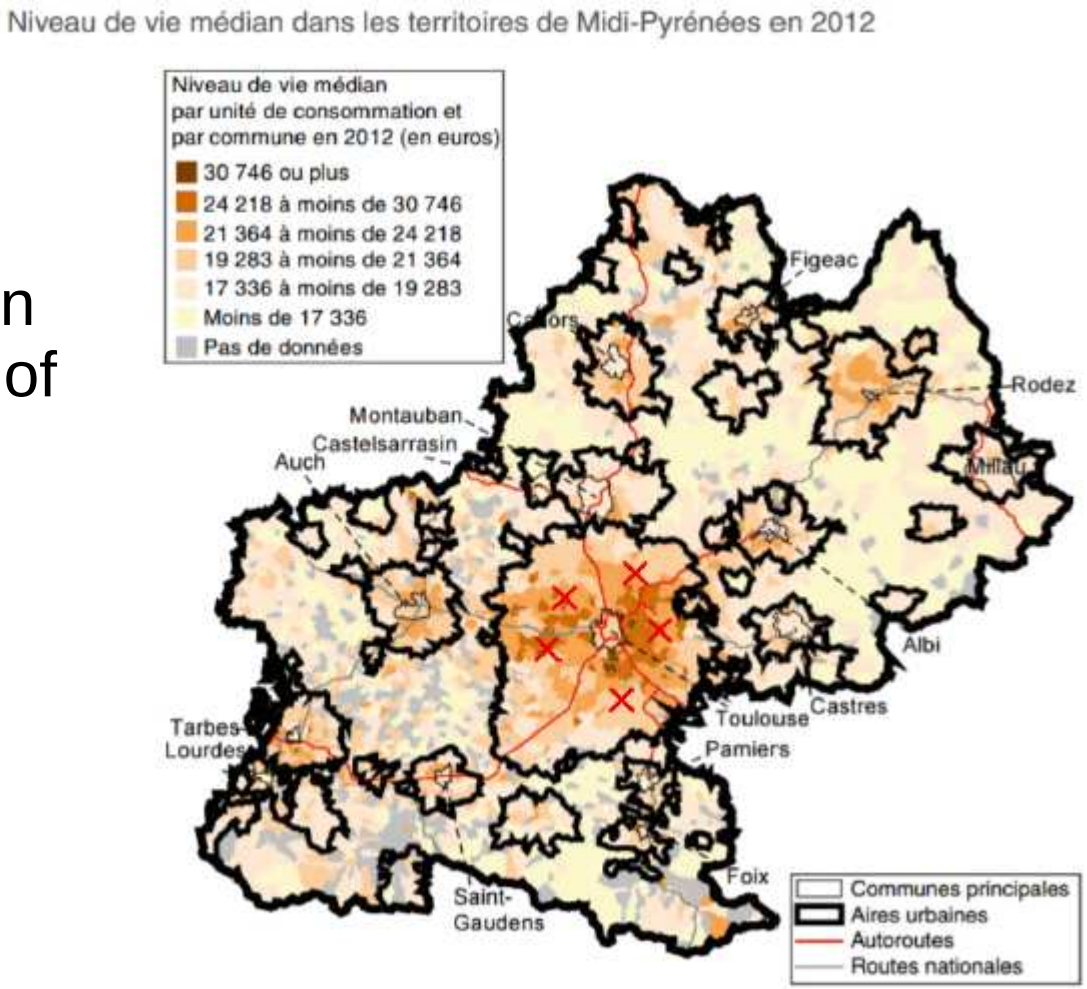
  - If we had to select some cities in order to estimate the average expenditure of a household, we should select them geographically dispersed

  - The underlying idea is that it is preferable to not sample a neighbouring unit of a selected unit due to spatial autocorrelation

The PU sampling design

- Example of the French region Midi-Pyrénées

PU sample concentrated in high standard of living areas

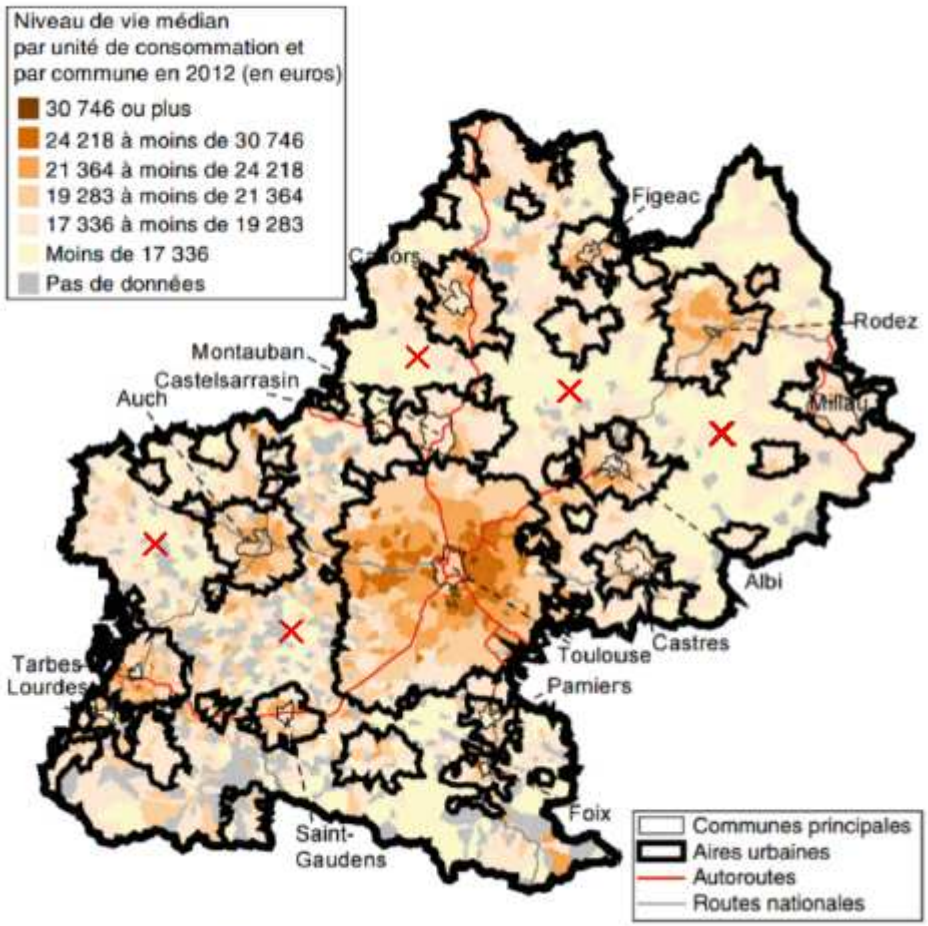


The PU sampling design

- Example of the French region Midi-Pyrénées

PU sample concentrated in low standard of living areas

Niveau de vie médian dans les territoires de Midi-Pyrénées en 2012



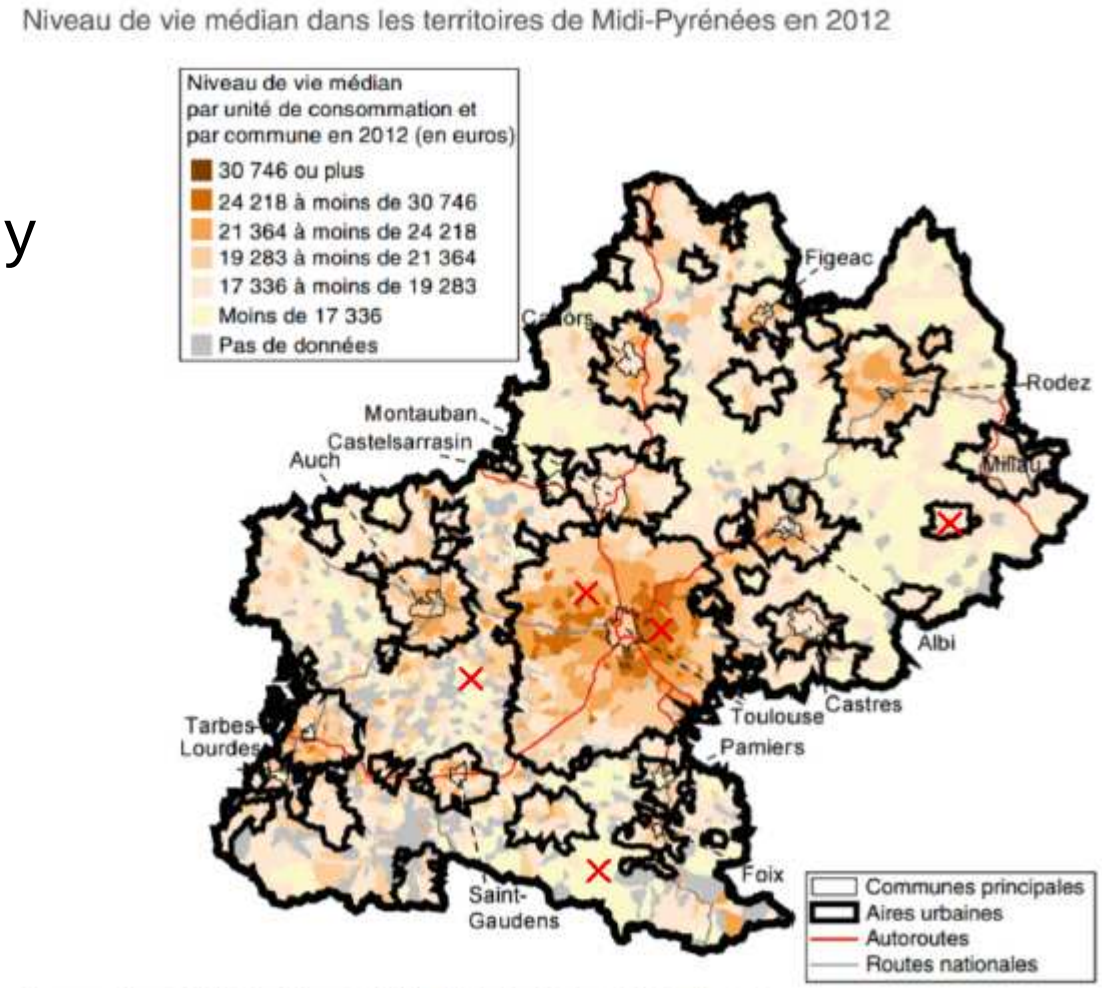
Sources : Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012

Insee

The PU sampling design

- Example of the French region Midi-Pyrénées

Geographically dispersed PU sample



The PU sampling design

- First stage sampling design we chose to use for the new French Master Sample (2020-2030) :

  Spatially balanced sampling method

- This sampling design is a process whose aim is to supply a sample that respects the following constraints :
  - PU inclusion probabilities are respected
  - The PU sample is approximately balanced on ***p*** auxiliary variables
  - The PU sample is geographically dispersed in order to cover the entire French territory

The PU sampling design

- **Spatially balanced sampling method**
  - It has been proposed by Grafström & Tillé (2013)
  - It has been developed by Grafström & Lisic (2016) with the package “BalancedSampling” in R
  - Several methods are currently tested at Insee by the Sampling Division to draw the PU for the new French Master Sample
  - Deadline to draw the new PU sample : March 2018

The PU sampling design

- **Measuring the gains brought by this sampling design**
  - 10,000 drawings of PU samples are realized
    - For a simple balanced sampling design
    - For a spatially balanced sampling design
  - Stratification by the region
  - Balancing variables :
    - Number of housings by type of urban zoning
    - Tax revenue
    - 3 age groups : <20, 20-60, >60 years old
    - Number of HLM housings
    - Type of households : single-parent families, large families
    - **+** geographical coordinates (X,Y) of the biggest city of the UP (only for the spatially balanced sampling design)

The PU sampling design

- **Results in terms of accuracy : Mean Squared Error**
  - For each balancing variable and other variables of interest
  - We estimate MSE by the mean thanks to the 10,000 simulations of simple balanced or spatially balanced PU samples drawings (Monte-Carlo method)
  - For balancing variables : MSE are very slightly higher with the spatially balanced method because of the adding of one balancing constraint for the drawing
  - For other variables of interest : MSE are lower with the spatially balanced method
    - The higher spatial autocorrelation of a variable is, the lower MSE is with the spatially balanced method than with the simple balanced method

The PU sampling design

- **Results in terms of geographical dispersion**
  - The Voronoi polygon of a sampled PU gathers all the geometrical coordinates that are closer to that PU than to an other sampled PU
  - Let δ_i be the sum of the inclusion probabilities of the PU located in the polygon ***i***
  - Let $\Delta = \frac{1}{n}\sum_{i \in S} (\delta_i - 1)^2$ be the Voronoi indicator
  - The more PU sample is geographically dispersed, the smaller Δ is
  - The value of Δ is estimated by the mean on the 10,000 simulations of simple balanced or spatially balanced PU samples drawings (Monte-Carlo method)
  - As expected, we note that the spatially balanced sampling is more geographically dispersed than the simple balance sampling

Summary

The annual sampling frame

- For each social survey, the sampling frame will include all the housings/individuals (according to the scope of the survey) that ...
  - belong to the cities situated in the sampled PU

    and
  - belong to the the last known Fideli files
    - Fideli files are updated every year thanks to the annual tax returns
    - Data on housings and individuals will be updated every year in Fideli files and therefore also in new French Master Sample computing application (Nautile)

Allocation by PU

- According to the **self-weighting two-stages sampling methodology** :
  - For the first stage sampling, we set the PU inclusion probabilities proportional to their sizes since it is worth less than 1
  - For the second stage sampling, under the assumption that PU inclusion probabilities are less than 1, we draw the secondary units by using a simple random sampling without replacement (SRSWOR) with a constant sampling size in each sampled PU
  - It ensures equality of the sampled secondary units' sampling weights

Allocation by PU

- For each social survey sample, the second stage allocations are computed in the following way :

  1. We allocate the sample between the set of the certain PU and the set of the probable PU proportionally to the size of each set (in terms of number of housings, of individuals)

  2. We allocate the sample between the sampled PU

     - For the set of the certain PU, we allocate the sample proportionally to the PU sizes

     - For the set of the probable PU, each PU (one interviewer) has the same allocation

This method enables to create approximatively the same workloads for all the interviewers

Housings/individuals sampling design

- Several sampling designs are possible to draw the secondary units according to the second stage allocations :
  - (stratified) SRSWOR
  - A (stratified) systematic sampling in the sampling frame previously sorted by selected variables
    - This sampling design is the current method to draw the housings for the surveys in the annual Census sampling frame
  - A balanced sampling
    - By being careful with the number of balancing variables with regard to the size of the sample
    - Besides the variance of the estimators is not easily computable for a balanced sampling at both stages

Housings/individuals sampling design

- Auxiliary variables available in Fideli files for stratification, sorting or balanced sampling
  - Individual level :
    - Age
    - Gender
    - Incomes/salaries
    - Marital status (married, single, ...)
    - ...
  - Housing level :
    - Type of building (house, flat)
    - Type of residence (main, secondary)
    - Occupancy status (landlord, renter)
    - Number of rooms
    - ...

Summary

Opportunities

- The previously presented methodology gets two disadvantages :
  - At the second stage of the sampling : by drawing independently several samples in the same sampling frame, a unit has a probability to be drawn in several samples
    - We wish to use a method to spread out the survey burden on the households of the Master Sample
  - At the first stage of the sampling : the use of a Master Sample (a single PU sample for a decade of surveys) concentrates almost all the surveys in the same PU for ten years
    - Is the Master Sample able to support all the samples, especially if we wish to draw samples that overlap as little as possible ?

Possibility of using negative coordination for the second stage sampling

- At the second stage sampling
  - We wish to use a method to spread out the survey burden on the households of the Master Sample
  - Insee currently uses **negative sample coordination methods** for Business surveys
  - The purpose of negative coordination is to take into account the previous surveys when drawing a new sample
    - in order to minimise the overlap between business samples
    - by fostering the selection of businesses that have not already been selected in recent surveys
    - while preserving the unbiasedness of the samples

Possibility of using negative coordination for the second stage sampling

- The process of coordinated sampling for Business surveys is the following
  - Each unit gets a permanent random number (PRN) : for a unit, as long as it exists, its number will never change
  - PRN are transformed by a coordination function wisely chosen to take into account the cumulative response burden linked to previous surveys
  - The selection of sample depends on the values of the transformed PRN
  - The inclusion probability of a unit is a decreasing function of the transformed PRN

Specific PSU for the Labour Force Survey

- At the first stage sampling
  - Drawing different PU samples according to the survey would enable to keep from running out the sampling frame, especially if we wish to draw samples that overlap as little as possible
  - However the choice of a single Master Sample (i.e. a single PU sample) is linked to the will to limit the costs of the surveys
    - Insee's interviewers' network is more or less permanent for a decade
    - We can't draw PU samples very far from each other without increasing the interviewers travel times

Specific PSU for the Labour Force Survey

- To solve this problem the Sampling Division is currently testing **positive sample coordination methods** to draw a specific PU sample for the Labour Force Survey :
  - The New French Master Sample will be used for all the other social surveys
  - A specific sample of areas will be drawn for the Labour Force Survey
    - This survey is quarterly
      - Quarterly sample contains about 108,000 respondent individuals
      - Each quarter, one sixth of the these individuals is renewed
    - Drawing the Labour Force survey sample in the PU of the New French Master Sample is not an option

Specific PSU for the Labour Force Survey

- To draw the Labour Force Survey sample :

1. We build for each PSU of the Master Sample a Coordination Unit (CU) that includes several PU
2. Each CU has to include a minimal number of housings
3. We draw the Labour Force Areas (primary units of this survey) in the CU

PU sampling

Selection of the linked CU

Drawing of the Labour Force Areas in the CU

Specific PSU for the Labour Force Survey

- Positive coordination between Master Sample PSU and Labour Force Survey Areas
  - enables to allocate the survey workload on a bigger sampling frame
  - while limiting the travel time of the interviewers
  - Indeed, to build the coordination units we use the path constituted by the Travelling salesman algorithm through the primary units
- Deadline to validate the methodology and to draw the new Labour Force Survey Areas : spring 2018

Summary

Conclusion

- The sampling methodology used at Insee for social surveys depends on several constraints of organization :
  - Primary units are constituted then drawn once for all during a decade (2020-2030 for the new French Master Sample)
  - Except for big cities (> 40,000 main residences) each sampled PU is assigned to a single interviewer
  - For each survey we have to provide approximately equal workloads among the interviewers
  - We shouldn’t interview the same household for more than one survey over a period of time (5 years with the actual system)

Conclusion

- Methodological choices which allow to respect these constraints
  - The algorithm of the Travelling salesman problem enables to minimize PUs' extent under constraints and thus to minimize the travelling time of the interviewers
  - The self-weighting two-stages sampling methodology enables
    - To get a second stage sample with equal sampling weights
    - To provide approximately equal workloads among the interviewers (equal allocations among the PU at the second stage sampling)

Conclusion

- Methodological choices which allow to respect these constraints
  - The negative sample coordination, currently used for Business surveys at Insee, would also allow to spread out the survey burden on the households

- At the first stage sampling : the spatially balanced sampling design allows
  - To respect the PU inclusion probabilities
  - To improve the estimators of parameters in terms of accuracy
  - To get a geographically dispersed PU sample

Literature

- Dickson, M. M., & Tillé, Y. (2016). Ordered spatial sampling by means of the travelling salesman problem. *Computational Statistics*
- Favre-Martinoz, C. (2017). An overview of spatially balanced sampling methods with an application to the French Master Sample
- Favre-Martinoz, C., & Merly-Alpa, T. (2017). Constitution et Tirage d'Unités Primaires pour des sondages en mobilisant de l'information spatiale
- Grafström, A. and Lisic, J. (2016). BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.5.2. http://CRAN.R-project.org/package=BalancedSampling
- Grafström, A., & Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*
- Gros, E., & Le Gleut, R. (2017). Sample Coordination and Response Burden for Business Surveys: Methodology and Practice of the Procedure Implemented at Insee. European Establishment Statistics Workshop 2017 – Southampton
- Guggemos, F., & Sautory, O. (2012). Sampling Coordination of Business Surveys Conducted by Insee. Fourth International Conference on Establishment Surveys – Survey Methods for Businesses, Farms and Institutions - Montréal

Sampling methodology for French social surveys

Thank you for your attention

Do you have some questions ?

Insee
Mesurer pour comprendre

Pierre-Arnaud PENDOLI
Expert in survey and census methodology
Census Methods and Processing Division
(+33)01.41.17.58.52.
pierre-arnaud.pendoli@insee.fr

French National Institute of Statistics and Economic studies

www.insee.fr

@InseeFr