Insee

Complementing Official Statistics with Mobile Phone Data

BigSurv 2018

Lino Galiana

Friday 26th October

Introduction

Motivation

- **Individual matching** between NSI's microdata and private actor's data **not possible**:
  - Individual privacy
  - Lack of common identifiers between databases

- Solution: **Combination from spatial aggregates**
  - Need common benchmark: *residence*

- This work has several objectives:
  - Show that INS data can be combined with mobile phone data to produce new information
  - Give a methodological framework to combine sources

Data

- Phone data: **2007 Orange's CDR**
  - Around 18 millions SIM cards, $\approx$ 4 billions observations a month
  - $\approx$ 18 000 antennas associated to voronoi cells
  - Events measured at antenna level
- Tax data: **2011 Filosofi database**
  - Geocoded tax data at (x, y) residence level
  - French households that declared income or received social benefits (almost exhaustive)
  - Enables to study population and income distributions at fine granularity levels
- Forbidden to communicate data on aggregate based on less than 11 households:
  - Spatial aggregation

Plan

I - Methodological points regarding spatial granularity

II - Combining income information and phone users data

III - Example of official statistics problematic with mobile phone: mobility and income inequality

I - Home detection and spatial granularity

Structure of phone data and home detection

- **2 dimensions** must be dissociated when combining sources:
  - *Spatial unit adopted*: voronoi or regular grid ?
  - *Antenna coverage model*: measurements are conditional on observation at antenna level

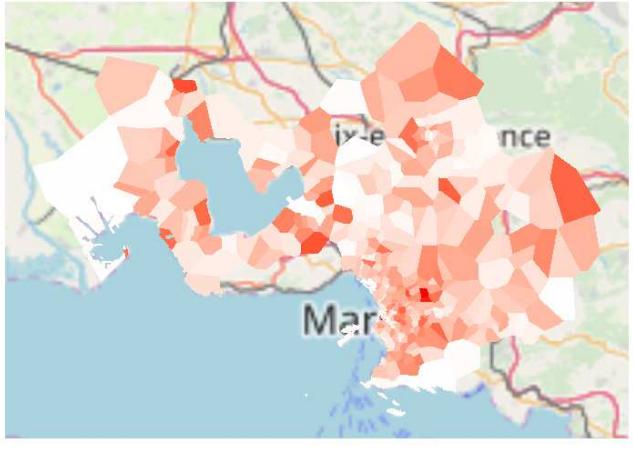
- Our work:
  - No hypothesis change regarding antenna coverage model
  - But we want to determine the sensitivity to spatial unit adopted

- Spatial unit definition must be the **first step**, not a reprojection after measuring events at voronoi level

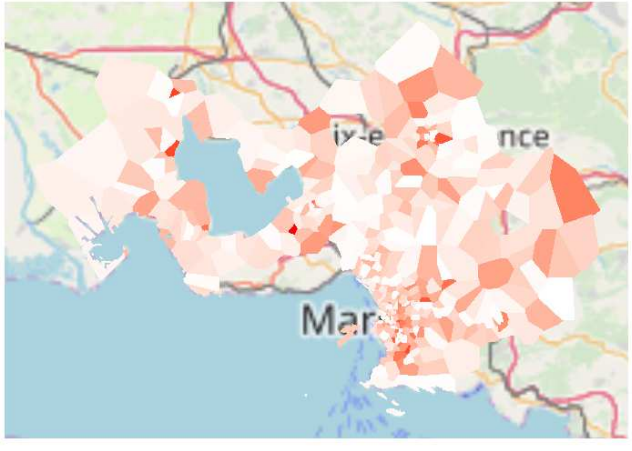
- **Residence**:
  - Benchmark to combine individual phone data and spatial aggregates
  - Home detection: first step for any further use of mobile phone

Comparing residential maps between tax and phone data

- Too uniform population distribution in phone data



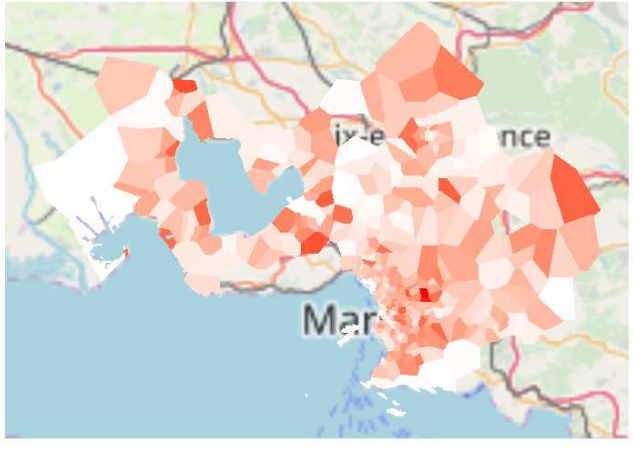
Population in tax data



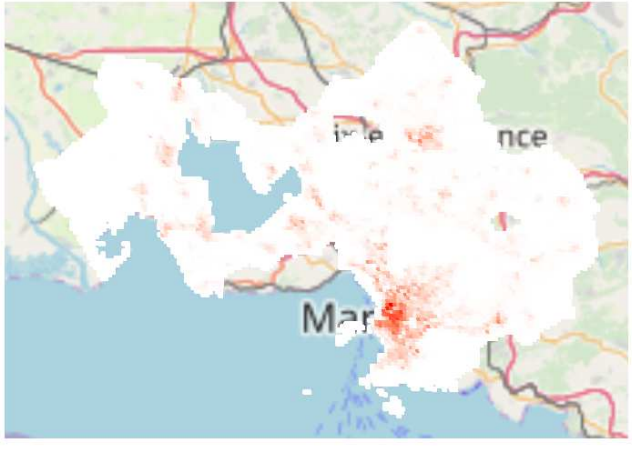
Population in phone data

Comparing residential maps between granularities

- Population in tax data not concentrated enough



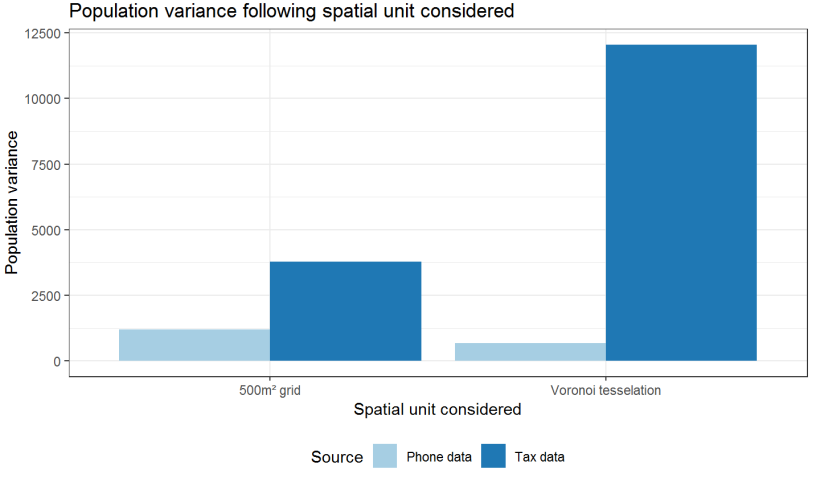
Voronoi tesselation



500m² grid

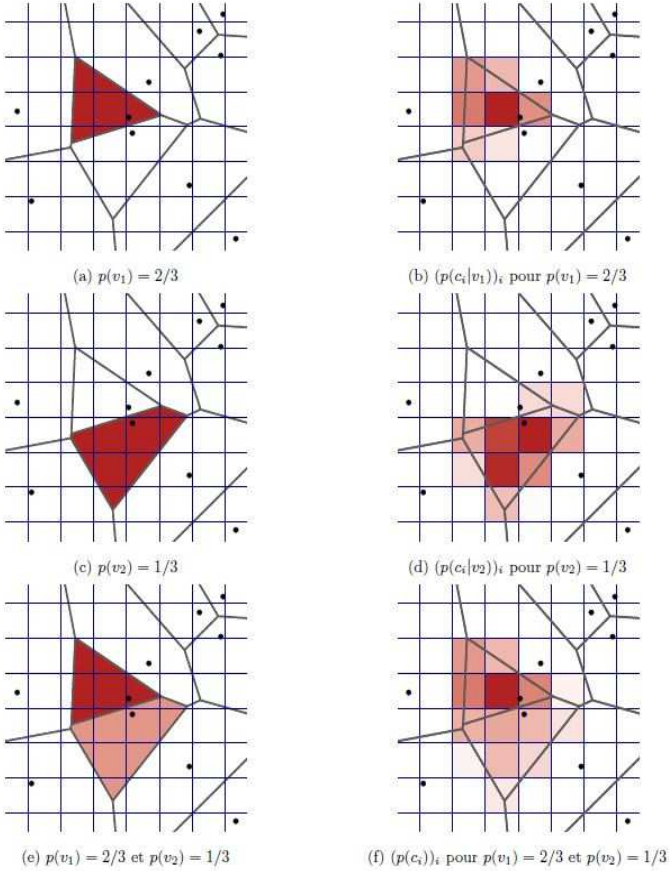
Limitations of Voronoi

- Voronoi have several limitations:
  - Large population variance in tax data…
  - Too uniform population distribution in phone data
  - Too many huge spatial units (Appendix on spatial units)
- Want a **more regular spatial units** than voronoi
  - Cell sizes to be less heterogeneous
  - Limit heterogeneity in population size when going to administrative data
- Want to probabilize phone users presence



Voronoi tesselation is not desirable when matching phone and income or population data

Probabilization of user's presence



(a) $p(v_1) = 2/3$ (b) $(p(c_i|v_1))_i$ pour $p(v_1) = 2/3$

(c) $p(v_2) = 1/3$ (d) $(p(c_i|v_2))_i$ pour $p(v_2) = 1/3$

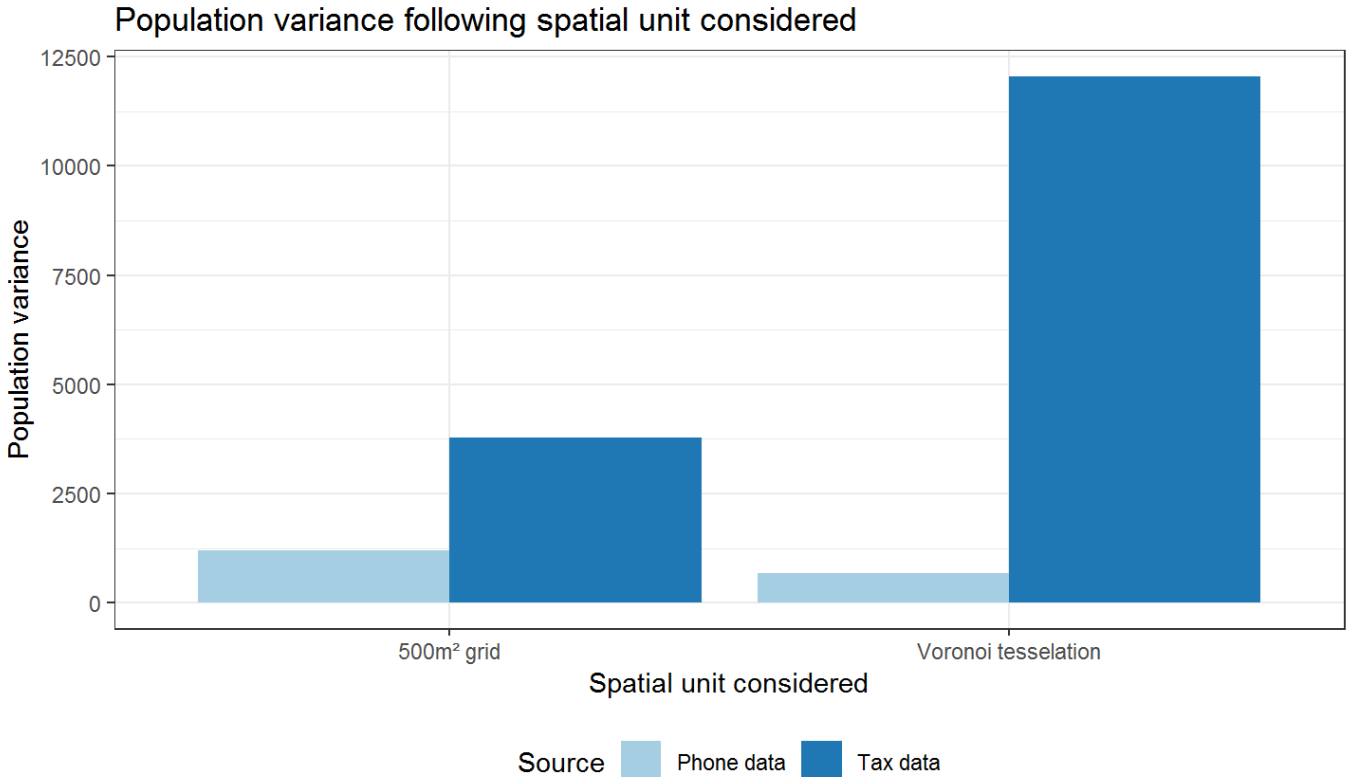
(e) $p(v_1) = 2/3$ et $p(v_2) = 1/3$ (f) $(p(c_i))_i$ pour $p(v_1) = 2/3$ et $p(v_2) = 1/3$

Figure 2: Voronoi tesselation and grid: example

Lecture : events are allocated to the cells as follow: a phone is located in voronoi 1 with probability 2/3 (a) and in voronoi 2 with probability 1/3 (c) (i.e 2/3 of its phone interactions are transmitted by the antenna in v1 and 1/3 by the antenna in v2). We split these probabilities according to the cell surfaces (see (b) and (d)), and then recover the global repartition for voronois (e) and cells (f).

Voronoi and grid: comparison

- More stable population in fiscal data
- Less uniformly distributed in space



II - Combining phone and income tax data

How to assign income to phone users?

- **Residence** can be used to map phone data with INSEE's data:
  - Requires to match from spatial aggregates
  - No information revealed for cells with less than 11 households

- Simulation based on income quantiles within cell
  - Other methods have been implemented, see Appendix on income distribution

- Adapt income assignment to the underlying population in tax data
  - Finer vision of income distribution when population is large
  - Fix number of quantiles as a function of population:

$$m_j = \left\lfloor \frac{N_j}{11} \right\rfloor$$

Income simulation method

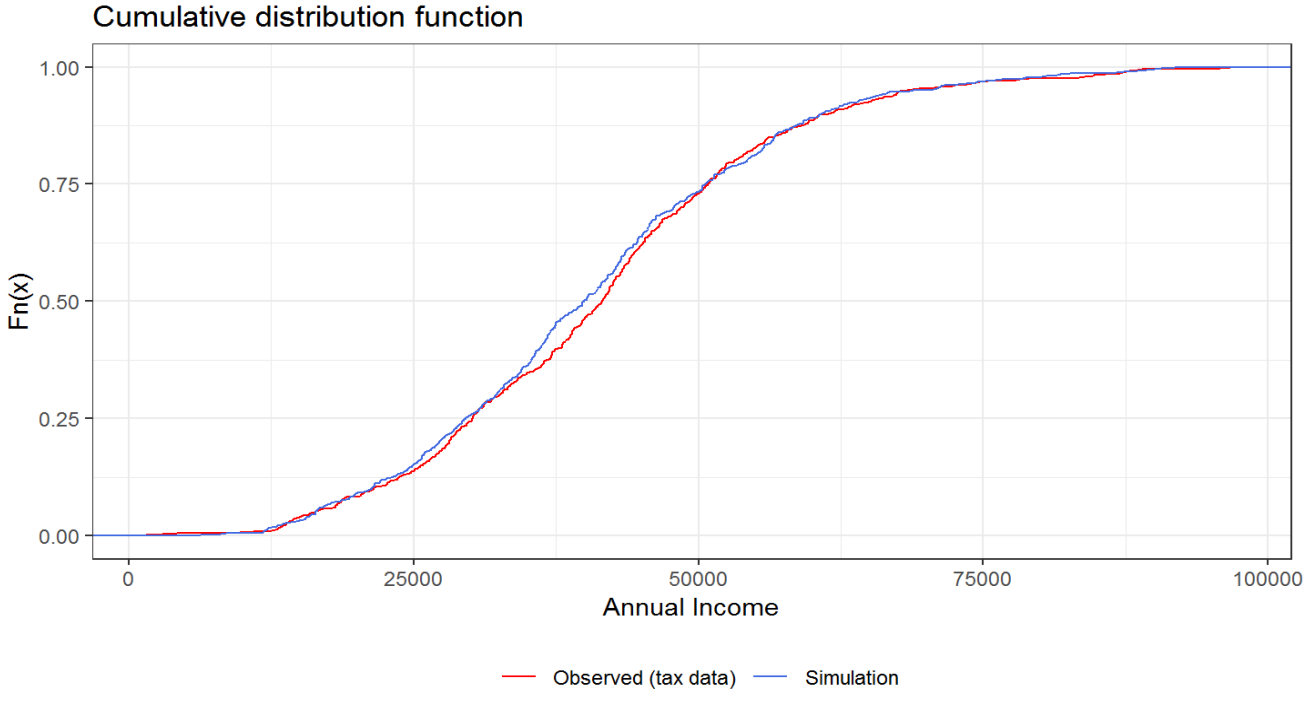
- Quantile interpolation: linear interpolation between the quantiles above and below uniform draw u

$$y = q_{\lfloor u \rfloor} + (q_{\lfloor u \rfloor + 1} - q_{\lfloor u \rfloor})(u - \lfloor u \rfloor)$$

- Test 1 (between variability): **reproduce median income distribution**:
  - Re-aggregate income by cell (median) in simulated and observed data
  - Compare cells median income distributions $(Y_j^{\text{tax}})_j$ and $(Y_j^{\text{simu}})_j$
- Test 2 (within cell): **cell by cell adequation test**
  - Test adequation for each cell with Kolmogorov-Smirnov test
  - Cell by cell KS adequation test
  - Test statistics: $D = \sup \left| F_n^{\text{tax}} \left(\widetilde{Y} \right) - F_n^{\text{simul}} \left(\widetilde{Y} \right) \right|$

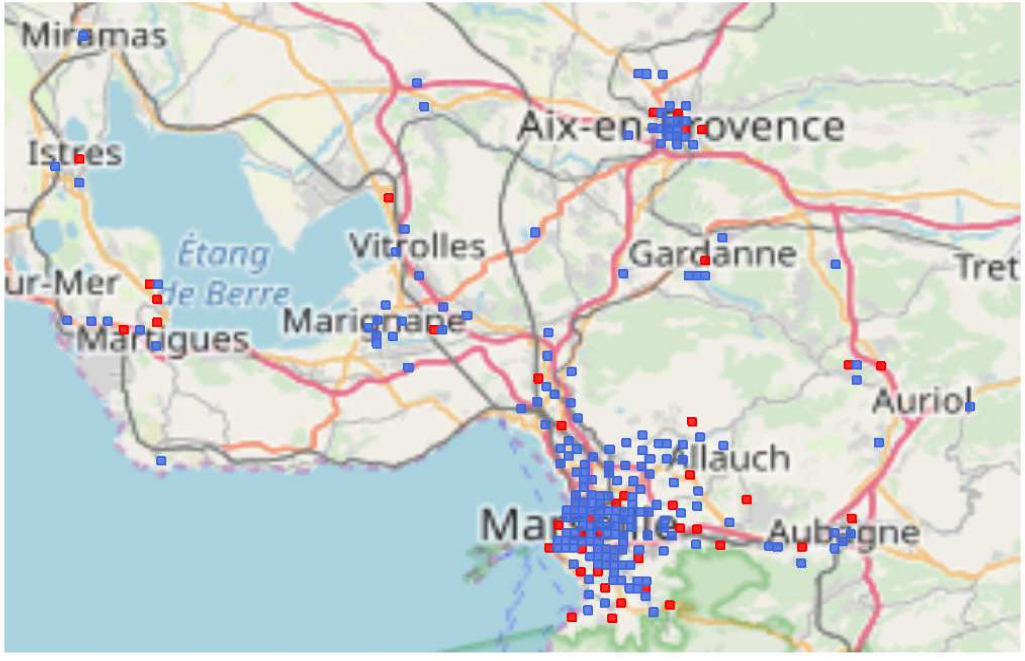
Test 1 (between variability): reproduce median income distribution

- Re-aggregate income by cell (median) in simulated and observed data
  - Compare cell's median income distributions $(Y_j^{\text{tax}})_j$ and $(Y_j^{\text{simu}})_j$



Test 2: generalized adequation test

- Reject adequation hypothesis (null hypothesis) in 18% of cells
- **Simulation less robust in suburbs** (where population is less dense)



III - Application

Measuring mobility with phone data (Paris)

- Mobility is traditionally studied with surveys:
  - Small samples
  - Omission biases
  - Phone data open new perspectives in that field

- Measure mobility by radius of gyration

$$rg_x = \sqrt{\frac{\sum_{i=1}^{n_x}(l_i - \bar{l}_x)^2}{n_x}}$$

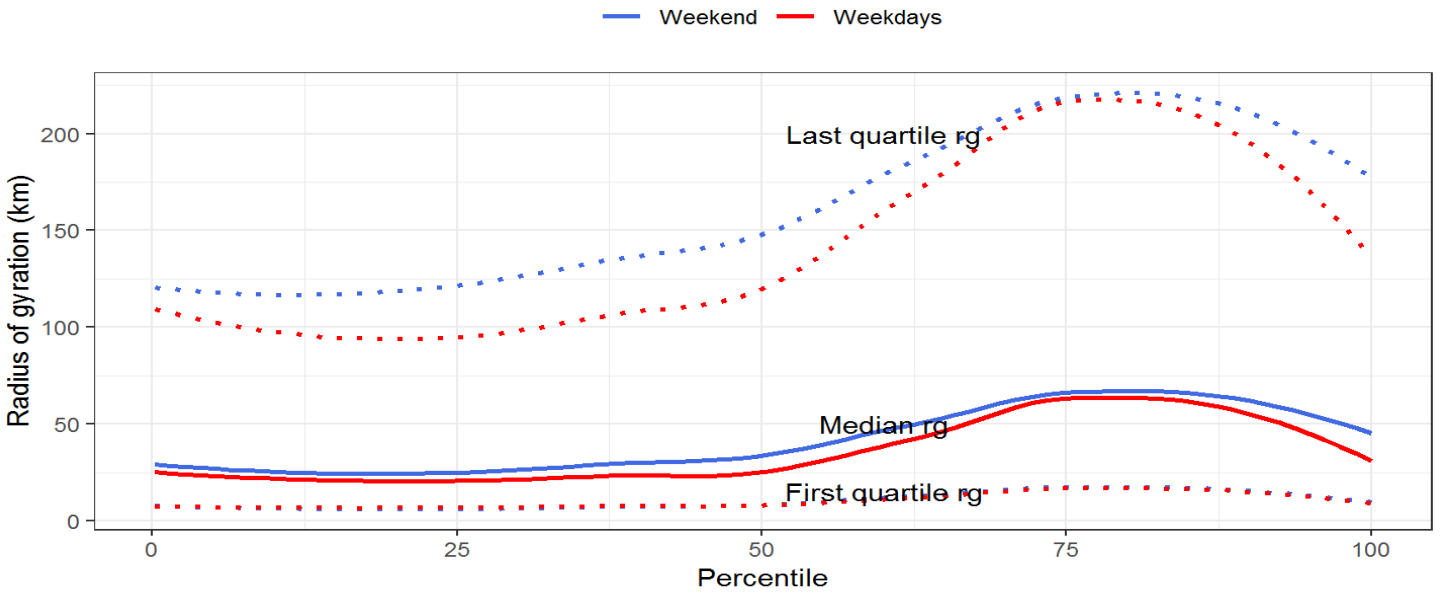
where l_i are two dimensionals locations coordinates and $\bar{l}_x$ a reference point

- Rather than taking $\bar{l}_x$ as barycenter of individual locations, we define $\bar{l}_x$ to be the centroid of the cell where x September home is located.

- Income definition: **home cell median income**

Mobility and income

- Mobility is different between income groups
- People are more mobile during weekends (with difference following income groups)



500 income groups (percentiles) are defined.
Median and quartiles computations are led by income group
Number of phone users by group around 6665
Total number of phone users: 3 405 190

Mobility pattern changes with holidays

- People tend to move more during summer (home is defined from September locations in both cases)
- People have similar mobility pattern between weekend and weekdays in summer while they change behavior in September
- Wealthier people tend to move more in summer Appendix

Conclusion

Conclusion

- Possibility for INS to **bring value to phone data**
  - Produce studies on INS scope
  - Example with insights regarding mobility in Paris

- Requires **carefulness in methodology**
  - Distinguish between spatial units and coverage model
  - Phone users' income assignment

Appendix on spatial units

Probabilization of user's presence

- Change spatial unit: **grid with 500m² cells**
  - $\approx$ 2 millions cells for France

- Probability an event occurred in cell i is

$$\mathbb{P}(c_i|v_j) = \frac{\mathcal{S}(c_i \cap v_j)}{\mathcal{S}(v_j)}$$

  where $\mathcal{S}(v_j)$ is the voronoi surface and $\mathcal{S}(c_i \cap v_j)$ the intersection area between c_i and v_j
  - Introduce uncertainty in measurement
  - More complex bayesian approaches for future research

- Probabilities are always **conditional on observation at antenna level**
  - antenna real coverage is unknown,
  - assumption that points are covered by closest antenna
  - abstraction from voronoi tesselation cannot be total

Probabilization of user's presence

- Bayes rule to get cell level probabilities

$$\forall c_j \in \mathcal{C}, \quad \mathbb{P}_x(c_j) = \sum_{v_j \in \mathcal{V}} \mathbb{P}\left(c_j \middle| v_j\right) \mathbb{P}_x(v_j)$$

(probabilities are indexed by x to show that this computation is made for every phone user x)

- Probabilization is made **before allocating home**
- Affect home detection: no longer an antenna of residence but a cell
  - For max action heuristic: choice based on $\mathbb{P}(c_i)$
  - For distinct day heuristic: $\mathbb{P}(c_i)$ values do not matter while they are strictly positive

Decomposition of population by spatial granularity

Table 1: Summary statistics of population in tax data by spatial raster

Cell size	Marseille		Lyon		Paris	
	Grid	Voronoï	Grid	Voronoï	Grid	Voronoï
Empty cell	0 (44.41)	0 (0.45)	0 (24.06)	0 (0)	0 (29.46)	0 (5.2)
Less than 11 household	1.04 (21.15)	0.01 (1.36)	0.9 (26.87)	0.01 (2.1)	0.14 (11.98)	0.01 (2.33)
Between 11 household and 50 individuals	0.95 (5.57)	0.01 (0.45)	0.79 (6.82)	0 (0.47)	0.11 (2.4)	0.01 (1.2)
Between 50 and 200 individuals	5.66 (11.88)	0.13 (4.77)	5.51 (16.39)	0.08 (3.04)	1.17 (9.04)	0.19 (6.07)
Between 200 and 1000 individuals	25.81 (11.64)	1.88 (13.18)	26.35 (18.27)	2.27 (15.19)	13.21 (20.89)	2.93 (19.41)
More than 1000 individuals	66.55 (5.35)	97.97 (79.77)	66.46 (7.59)	97.64 (79.21)	85.37 (26.23)	96.85 (65.78)
Total	1 758 986 (8091)	1 758 986 (440)	1 740 388 (5412)	1 740 388 (428)	10 990 852 (12 617)	10 990 852 (2998)

For each category, the first row corresponds to the share of population represented while the second row (between parenthesis) corresponds to the share of spatial units considered. In the Total category is reported total city population and number of spatial units of the tesselations in columns.

Lecture: In Paris, when adopting voronoi tesselation, 65.78% of the polygons present more than 1000 people. 96.85% of Paris people live inside those polygones. When adopting grid, only 26.23% have more than 1000 people. This still represents of 85.37% Paris population.

Appendix on income distribution

Income simulation methods

- Assignation can be parametric or not:
  - Non-parametric : use quantiles to approximate income distribution
  - Parametric: simulate from a known distribution given estimated parameters

- **Non-parametric methods**:
  - Quantile interpolation: given an ordered set of quantiles $(q_1, \dots q_{m_j})$ and a uniform draw $u \sim \mathcal{U}[0, m_j - 1]$ and apply linear interpolation between the quantiles above and below u. In other words,

$$y = q_{\lfloor u \rfloor} + (q_{\lfloor u \rfloor + 1} - q_{\lfloor u \rfloor})(u - \lfloor u \rfloor)$$

  - Jitterized quantiles: We draw from a discrete uniform distribution. Given an ordered set of quantiles $(q_1, \dots q_m)$, we take the u^{th} positioned quantile, denoted q_u.After drawing q_u, we jitterize it to get income y as $y = q_u + \epsilon$ where ϵ is a jitter

- **Parametric methods**:
  - Normal simulation
  - Log-normal simulation (not yet implemented)

How to evaluate simulations ?

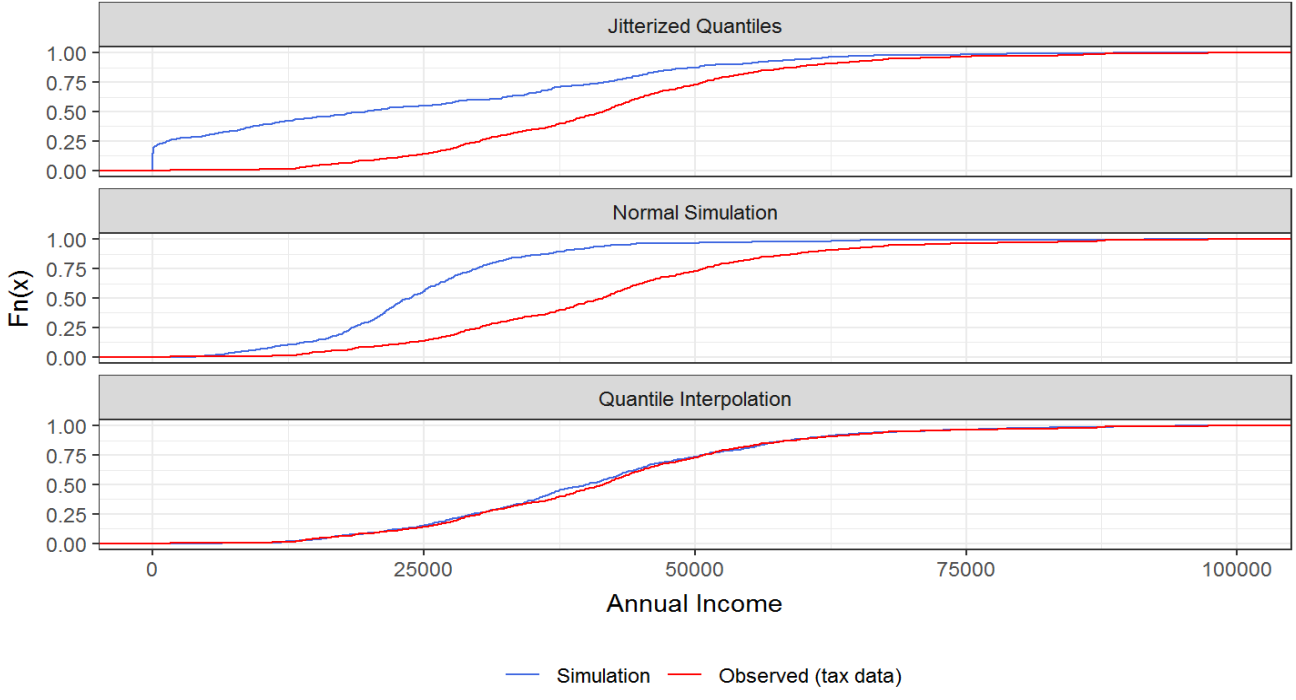
- We want to reproduce income distribution in several dimensions:
  1. Between cells variability
  2. Within cells variability
  3. Spatial autocorrelation

- We propose tests for 1. and 2.
  - Between variability: cell by cell, do we reproduce median income ?
  - Within variability: cell by cell, do simulated distribution look as observed one ?

- Examples for Marseille

Test 1 (between variability): reproduce median income distribution

- Re-aggregate income by cell (median) in simulated and observed data
  - Compare cell's median income distributions $(Y_j^{\text{tax}})_j$ and $(Y_j^{\text{simu}})_j$

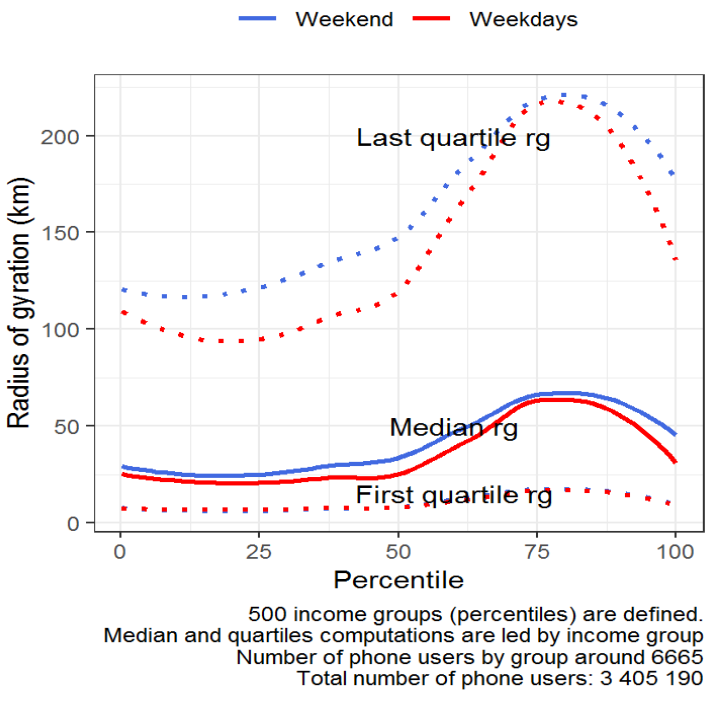


Test 2 (within cell): cell by cell adequation test

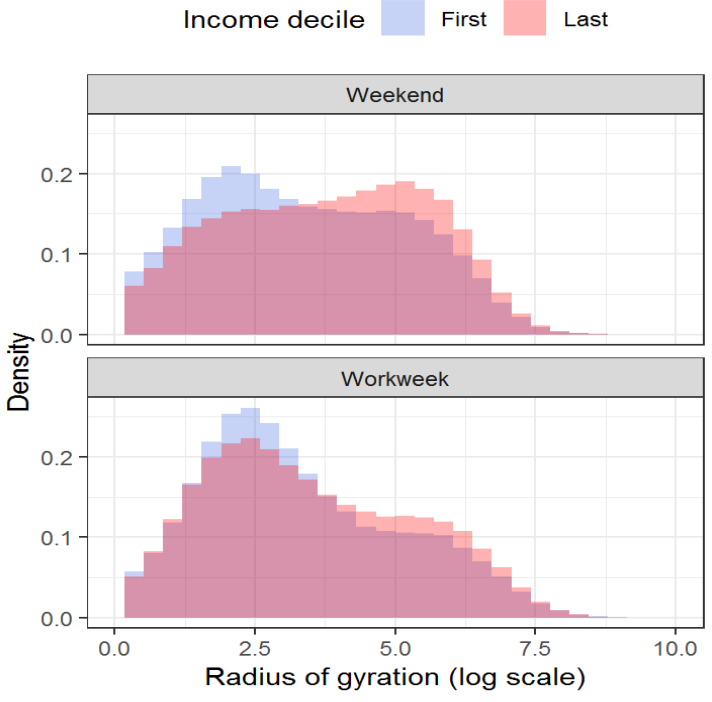
- Quantile interpolation is the best, by far

Appendix on mobility

Mobility and income



500 income groups (percentiles) are defined.
Median and quartiles computations are led by income group
Number of phone users by group around 6665
Total number of phone users: 3 405 190



Wealthier people tend to move more in summer