Dynamic approach of spatial segregation: a framework with mobile phone data

Lino Galiana (INSEE)

With Benjamin Sakarovitch (INSEE), François Sémécurbe (INSEE) and Zbigniew Smoreda (Orange Labs)

UEA Amsterdam 2019

May 31th, 2019

Introduction

Residential segregation drivers

Residential segregation drivers: housing

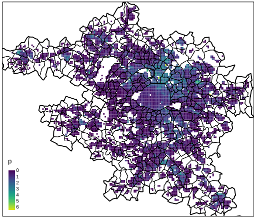
- Income gradient from housing prices (Alonso, 1964)
  - High opportunity cost of transportation: wealthiest live in city center, poorest in suburbs
  - High valuation of housing space: wealthiest live in suburbs, poorest in city center
- Social housing aims to ensure social mixing
  - Social housing clusters poor population in specific areas (Verdugo and Toma, 2018)
  - Dynamic effect: school segregation creates persistence \item People can coexist without interaction (Chamboredon and Lemaire, 1970)

Residential segregation drivers: preferences and mobility

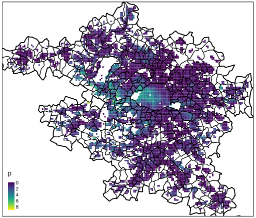
- Heterogeneity in preferences have spatial effects
  - Schelling (1969): clustering based on preference for neighborhood
  - Tiebout (1956): spatial sorting based on public goods preferences
- Mobility plays a key role to understand segregation
  - Long run: high quality public good bring people in neighborhood, affecting housing price (Black, 1999; Fack and Grenet, 2010)
  - Within-week mobility brings together people from different neighborhood
- Infraday dynamic can be strong:
  - Davis et al. (2017): outside segregation (restaurants) 50% lower than residential segregatio
  - Athey et al. (2019): similar scale for public space as parks

Residential segregation: limitations of tax data

- Good picture of residential segregation with tax & census data
- But fixed picture
  - People spend time out of their living neighborhood:
  - Experienced segregation vs residential segregation



(a) Low-income population (first decile)



(b) High-income population (last decile)

Residential segregation: limitations of tax data

- Dissimilarity index (Duncan & Duncan, 1955)

$$ID = \frac{1}{2}\sum_{j=1}^{J}\left|\frac{w_j}{W_T} - \frac{n_j - w_j}{N_T - W_T}\right|$$

- Administrative data $\Rightarrow$ residential segregation:
  - Static vision of segregation
  - Separation of income groups within residential space
  - No information on visited places
- Mobility continuously reshapes income spatial distribution
  - Need high-frequency geolocated data...
  - ... combined with traditional data to characterize individuals

Research question

Research question

- Main questions:
  - How do mobility affect urban segregation ?
  - Do high-frequency data help us in identifying patterns in segregation that cannot be understood with administrative data?
- Contribution:
  - Combining phone and traditional data
  - Proposition of a methodology to ensure combination robustness
  - Fine spatial and temporal granularity to understand segregation
  - Next step is to interpret patterns with respect to city characteristics

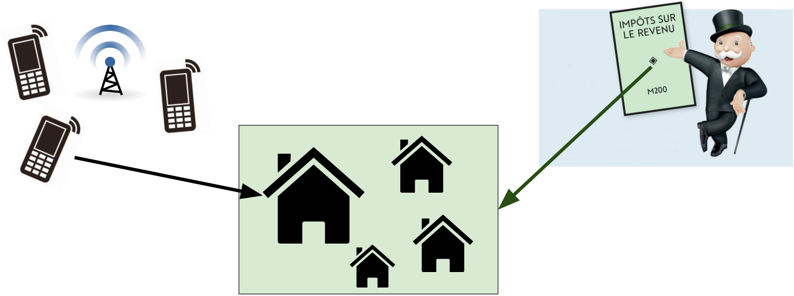
Methodology adopted

- We analyze infraday dynamic:
  - 48 points: 24 for weekdays, 24 for weekend
- Requires time depending segregation indexes
  - Dissimilarity index series for each city
- Paris, Lyon and Marseilles
  - Agglomeration level: city centers and suburbs
  - More than 13 millions people in tax data
  - More cities soon

Data

Principle

- Caracterize phone users from living environment
- Probability of belonging to first/last decile from observed income distribution in tax data



Phone data

Phone data

- Orange data September 2007
  - 18.5 millions SIM cards ($\approx 1/3$ French population)
  - Text messages and call: 3 billions events
  - Geocoding at antenna level (exact (x, y) unknown)
- Transformation into 500x500 meters cell level presence

  Methodology here
- We do not use interaction dimension
  - Plan for future research on social segregation
- Big data volume is a challenge

Phone data

- 2007 is old:
  - People were not using their phone as much as now
  - Temporal sparsity at individual level (in average 4 points a day by user)

	mean	s.d.	min	P10	P25	median	P75	P90	max
Average number of daily events per user	4.3	3.6	1	1.4	2	3.1	5.4	8.7	123
Number of distincts days users appear	20	9.2	1	5	13	23	28	30	30
Average number of events between 7PM and 9AM per user	2.4	1.7	0	1	1.3	1.9	2.9	4.4	87
Number of distincts days users appear between 7PM and 9AM	15.2	9.4	0	2	7	15	24	28	30
Number of observations:				3,024,884,663					
Number of unique phone users:				18,541,440					

Table 1: Orange 2007 CDR : summary statistics of September data
[replace and update the one in the paper]

Tax data

Tax data

- 2011 geocoded tax data at (x, y) level
  - Income by consumption unit
- Income based segregation
  - Distribution of income extremes (first and last deciles)
  - Relative definition of income: is individual wealthier/poorer than a city reference level ?
- Bimodal approach
  - First decile vs others
  - Last decile vs others

Tax data

- Sub-population (first/last decile) frequency in cell
- Spatial aggregation at cell level i

$$p_i^{D1} = \mathbb{P}(y_x < \mu^{D1}) = \mathbb{E}(\mathbf{1}_{\{y_x < \mu^{D1}\}}) = \frac{1}{n_i} \sum_{x=1}^{n_i} \mathbf{1}_{\{y_x < \mu^{D1}\}}$$

$$p_i^{D9} = \mathbb{P}(y_x > \mu^{D9}) = \mathbb{E}(\mathbf{1}_{\{y_x > \mu^{D9}\}}) = \frac{1}{n_i} \sum_{x=1}^{n_i} \mathbf{1}_{\{y_x > \mu^{D9}\}}$$

- If $p_i > 0.1$, over-representation of subpopulation in cell
- That frequency is used to simulate phone user status given their simulated residence

Tax data

- Intuitions regarding city segregation from tax data
  - e.g. Paris: more segregation at the top

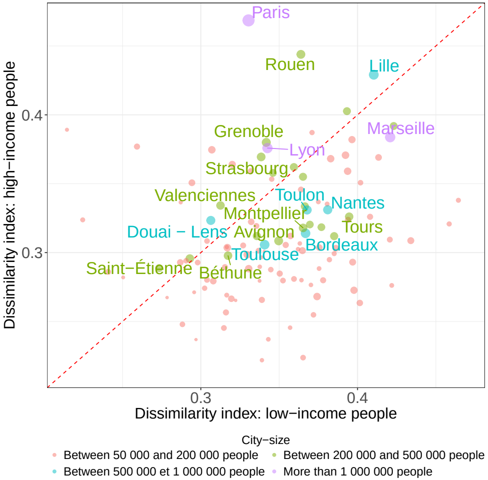


Figure 2: Dissimilarity index for main French cities

Methodology to build segregation index

Workflow

- Phone user status is simulated from his/her phone track (only personal information) and neighborhood level tax aggregates
- 3 steps to estimate segregation dynamics:
  1. Home estimation
     - Estimate probabilities that individual lives in some neighborhood given nighttime (19 pm - 9 am) phone track
  2. Home cell and income simulations
     - Home simulation knowing cell level probability sequences
     - Income simulation given first/last decile frequence appearance in tax data (p_i)
     - Test other designs to check robustness of income simulation
  3. Compute segregation indexes
     - They depend on observation time t (dynamic approach)

Details for step 1 and 2 here

Segregation index

- Two typical days: weekdays, weekend
- Individual probabilities at cell level on a given time window: $\mathbb{P}_x(c_{it})$ Details
- Probabilize dissimilarity index (Duncan & Duncan, 1955):

$$ID_t^g = \frac{1}{2}\sum_{c\in\mathcal{C}}\left|\frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\in g}}{\underbrace{\sum_{x\in\mathcal{X}}\mathbf{1}_{x\in g}}_{\substack{\text{Number people of income group } g\\ \text{that are observed at time } t}}} - \frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\notin g}}{\underbrace{\sum_{x\in\mathcal{X}}\mathbf{1}_{x\notin g}}_{\substack{\text{Number people not in income group } g\\ \text{that are observed at time } t}}}\right|$$

- Remainder, standard index:

$$ID = \frac{1}{2}\sum_{c\in\mathcal{C}}\left|\frac{w_c}{W_T} - \frac{n_c - w_c}{N_T - W_T}\right|$$

Results

Segregation dynamics

Segregation dynamics: low-income

- City-level segregation evolution across time
  - People not observed at a given hour of the night (19-9) are assumed to be at home
  - This removes downward bias in index with respect to tax data
  - Dynamic robust to other income simulation methods

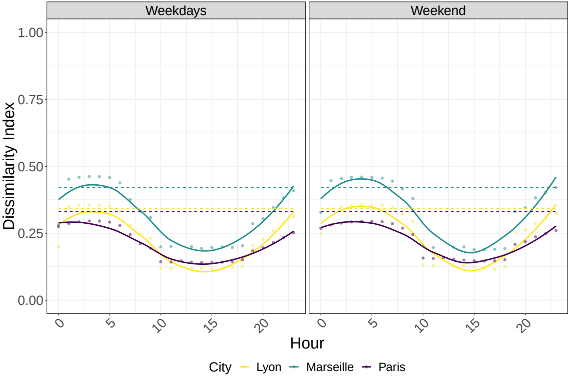


Figure 3: Low-income segregation dynamics

Segregation dynamics: high-income

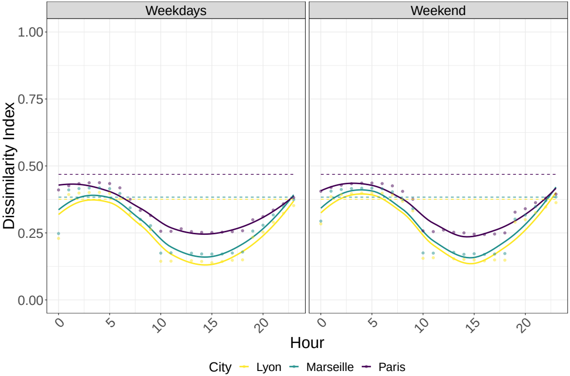


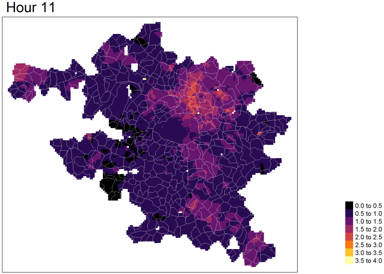
Figure 4: High-income segregation dynamics

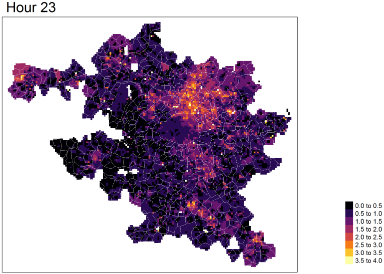
Segregation dynamics: comparing cities and income groups

- Significant difference between nighttime and daytime segregation levels
  - Segregation starts to decrease around 6-7am and goes up after 4-5pm
  - No significant difference between weekend and weekdays $\Rightarrow$ separate saturday and sunday ?
- Differences in level observed in tax data also present in phone data
  - e.g. Paris: segregation higher at the top
- Mobile phone inform us on dynamics:
  - Decrease stronger in Marseilles and Lyon than in Paris
- Further research: can we identify some inclusive/exclusive cities ?

Evolution of city structure across time

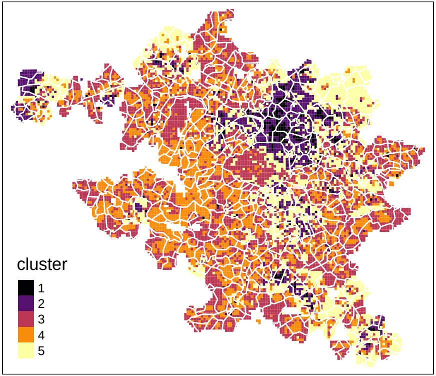
e.g. Low-income concentration at two different hours (Full sequence here)

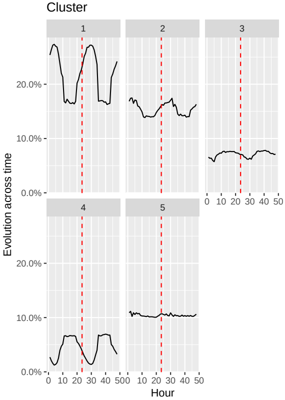




Spatial clustering [really preliminary]

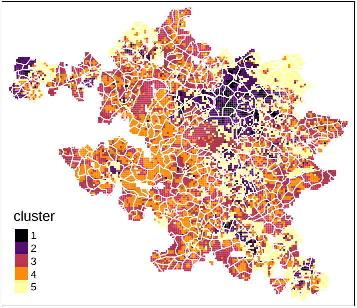
- Clustering to identify spaces that share common population composition characteristics
  - Will be related to places characteristics (infrastructures...)
- e.g.: share of population belonging to low-income group

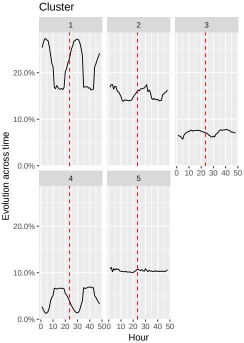




Spatial clustering [really preliminary]

Cluster	Night	Day
1	Large over-representation	Decrease
2	Large over-representation	More stable
3	Under-representation	Small increase
4	Large under-representation	Increase
5	Stable at 10%	Stable at 10%





Conclusion

Conclusion

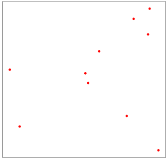
- Bringing together phone and tax data requires methodological foundations
- Segregation at its acme during nighttime/hometime
- Need interpretation of segregation spatio-temporal dynamics with respect to city amenities
- Results consistent with Davis et al (2017) and Athey et al (2019)

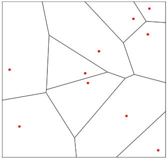
Appendix

Probabilization

Phone users' presence probabilization

Back to slide

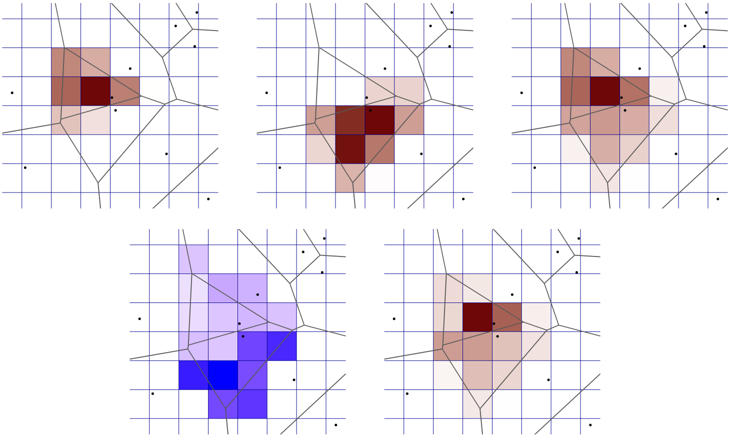




- Mobile phone litterature does not dissociate:
  - Coverage area: observations at antenna level into presence area
  - Statistical unit: economic information level
- Coverage area: Voronoi tesselation
  - Each point in space is associated with closest antenna
- However, must not be analysis statistical unit
  - Partition depends too much on antennas local density

Phone users' presence probabilization

- Cell level probabilization to abstract from voronoi
  - Knowing call has been observed from antenna v_j, probability it happened into cell c_i? (Bayes rule)
- 500x500m cell level
  - Phone data: probabilize both presence and home
  - Tax data: local aggregates at cell level



Methodology: more details

1. Home estimation

- Nighttime phone track (19h-9h) used to estimate individual residence probability for all cells
- Bayesian approach to account for the fact that all metropolitan space is not residential
  - In a coverage area, prior in most densily populated cells
  - Prior from population density computed from tax data
- Prior distribution is a reweighting for cell level home

$$\mathbb{P}_x(c_i^{\text{home}}|v_j) \propto \underbrace{\mathbb{P}(c_i^{\text{home}})}_{\substack{\text{prior from}\\ \text{population density}}} \quad \underbrace{\mathbb{P}_x(v_j|c_i)}_{\substack{\text{areas ratio:}\\ \frac{s(v\cap c)}{s(c)}}}$$

- Sequence from home probabilities: $\nu_x^{\text{home}}(c_i)$
  - Used to simulate x income

Back to presentation

2. Home and income simulations

4 methods of home simulation to check robustness of segregation indexes

Methodology	Choice of x's home
Main method	Draw home from all residence probabilities $\nu_x^{\texttt{home}}$
`One stage simulation`	Cell where probability is maximum: $c_i = \arg\max_{c_i} \nu_x^{\text{home}}(c_i)$
`cell_max_proba`	x assigned where probability of being member of group g is maximized
`cell_min_proba`	x assigned where probability of being member of group g is minimized

Last two methods: evaluate effect on segregation indexes to over- or under-estimate the share of sub-group g on population

Back to presentation

3. Segregation indexes: cell level presence

- Probability that an event measured in antenna v_j at time t occurred in cell c_i is

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

- We denote c_{it} the probability of being present at time t in cell c_i. This is a recollection of conditional probabilities

$$\forall c_{it} \in \mathcal{C}, \quad \mathbb{P}_x(c_{it}) = \sum_{v_{jt} \in \mathcal{V}} \mathbb{P}(c_{it}|v_{jt})\mathbb{P}_x(v_{jt}) \tag{1}$$

with $\mathcal{V}$ voronoi/antennas and $\mathcal{C}$ 500m cells.

Back to presentation