Residential segregation, daytime segregation and spatial frictions: an analysis from mobile phone data

Lino Galiana (INSEE)

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

JMA 2021 Virtual Conference

June 4th, 2021

Outline

Introduction

A few links

- Latest working paper version here
- A shorter version (policy brief) here
- Everything can be found in my personal website: https://linogaliana.netlify.app

Why social mixing matters ?

- Polarization in residential market
  - gentrification
  - concentration of low-income people in social housing
- Segregation has long-run effects
  - School
  - Access to public infrastructures
- Hot policy questions
  - *Loi séparatisme* in France
- Not so much quantitative knowledge when not looking from residential perspective

Why do we need mobile phone data to measure segregation ?

Residential segregation drivers: housing

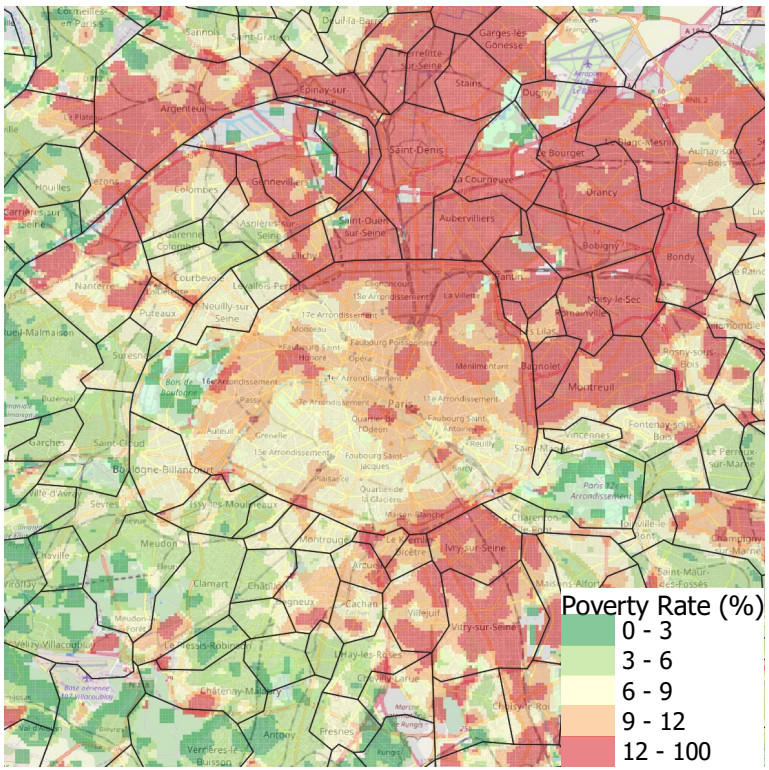
- Income gradient from housing prices (Alonso, 1964)
  - High opportunity cost of transportation: wealthiest live in city center, poorest in suburbs
  - High valuation of housing space: wealthiest live in suburbs, poorest in city center
- Social housing aims to ensure social mixing
  - Social housing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

Residential segregation drivers: preferences and mobility

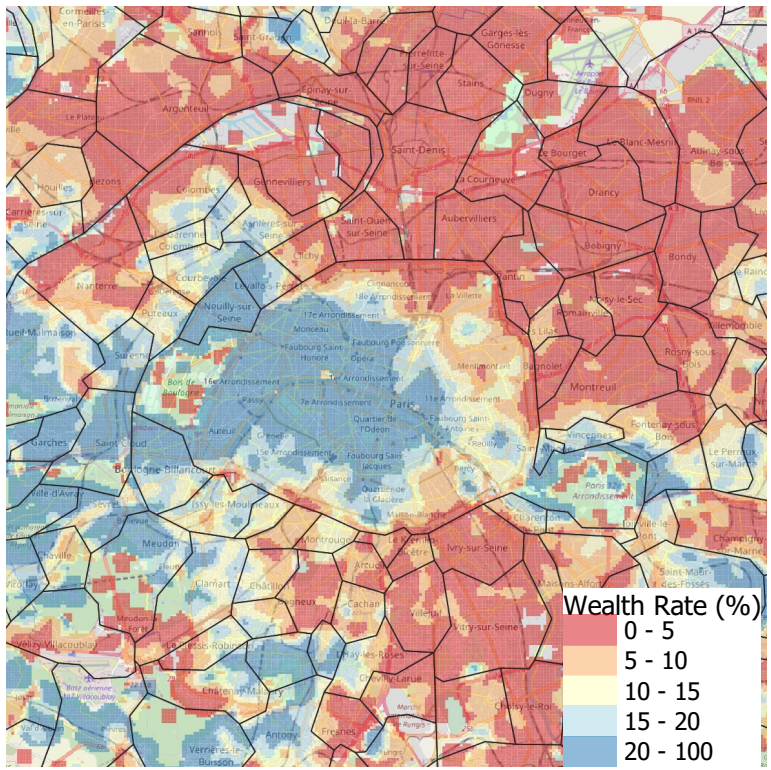
- Heterogeneity in preferences have spatial effects
  - Schelling (1969): clustering based on preference for neighborhood
  - Tiebout (1956): spatial sorting based on public goods preferences
- Mobility plays a key role to understand segregation
  - Long run: high quality public good bring people in 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. (2019): outside segregation (restaurants) 50% lower than residential segregatio
  - Athey et al. (2019): similar scale for public space as parks

Goal of the paper

From a picture



(a) Low-income population (first decile)



(b) High-income population (last decile)

to a more complete sequence

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
- Numeric traces useful to know where people go
  - Davis et al. (2019): Yelp data
  - Athey et al. (2020): GPS data

Residential segregation: limitations of tax data

- Theil index (Theil, 1984)

$$H = \frac{n_c}{N^{\text{city}}} \sum_{c=1}^{C} \frac{E(p^{\text{city}}) - E(p_c)}{E(p^{\text{city}})}$$

  - Entropy measures diversity
    $(E(p) = -p\log(p) - (1-p)\log(1-p))$
  - Compares entropy at city and cell level
- 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

Outline

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 understand with administrative data?
  - Can we measure heterogeneity in spatial frictions within a city using high resolution mobility flows ?
- Contribution:
  - Combining phone and traditional data
  - Proposition of a methodology to ensure combination robustness
  - Fine spatial and temporal granularity to understand segregation
  - Gravity approach with large scale data to measure cost of mobility

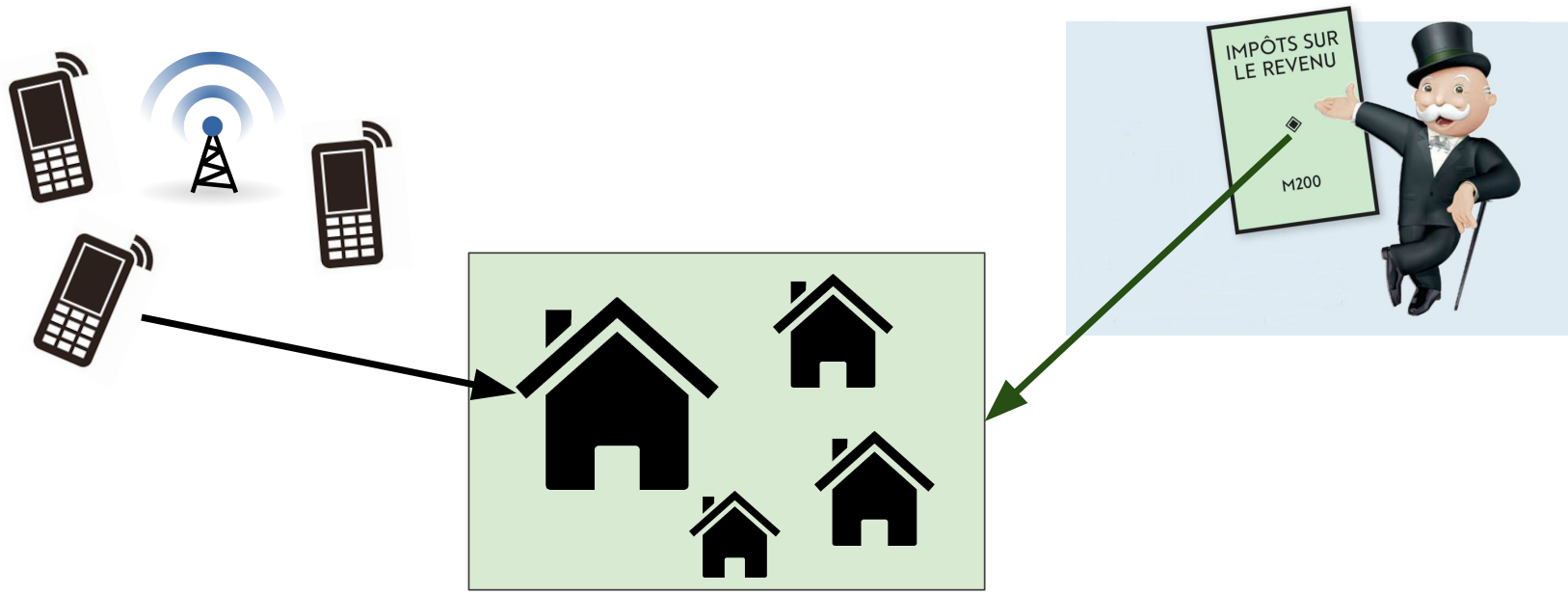
Methodology adopted

- We analyze infraday dynamic:
  - 48 points: 24 for weekdays, 24 for weekend
- Requires time depending segregation indexes
  - Theil index series for each city
- Gravity model to measure spatial frictions
  - Takes into account the zero-flows problem
- Paris, Lyon and Marseille
  - Agglomeration level: city centers and suburbs
  - More than 13 millions people in tax data

Outline

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

Outline

Tax data

- 2014 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

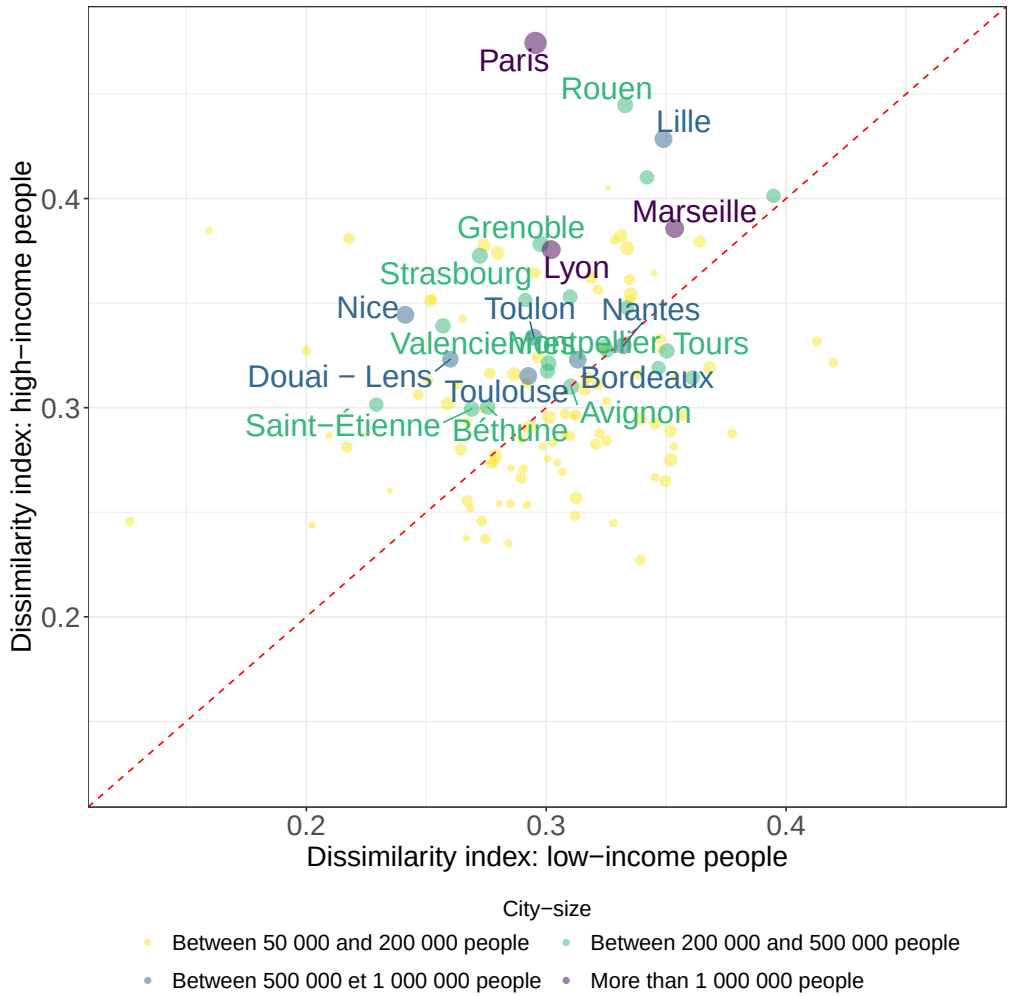


Figure 2: Dissimilarity index for main French cities

Outline

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 Theil index (as well as other indices):

$$H_t = \sum_{c=1}^{C} \frac{\overbrace{\sum_{x \in \mathcal{X}} \mathbb{P}_{x,t}(c)}^{\substack{\text{Number people of income group } g \\ \text{that are observed at time } t \text{in cell } c}}}{\underbrace{\text{card}(\mathcal{X})}_{\substack{\text{Number people of income group } g \\ \text{that are observed at time } t}}} \frac{E(p_{c,t}) - E(p_t^{\text{city}})}{E(p_t^{\text{city}})}$$

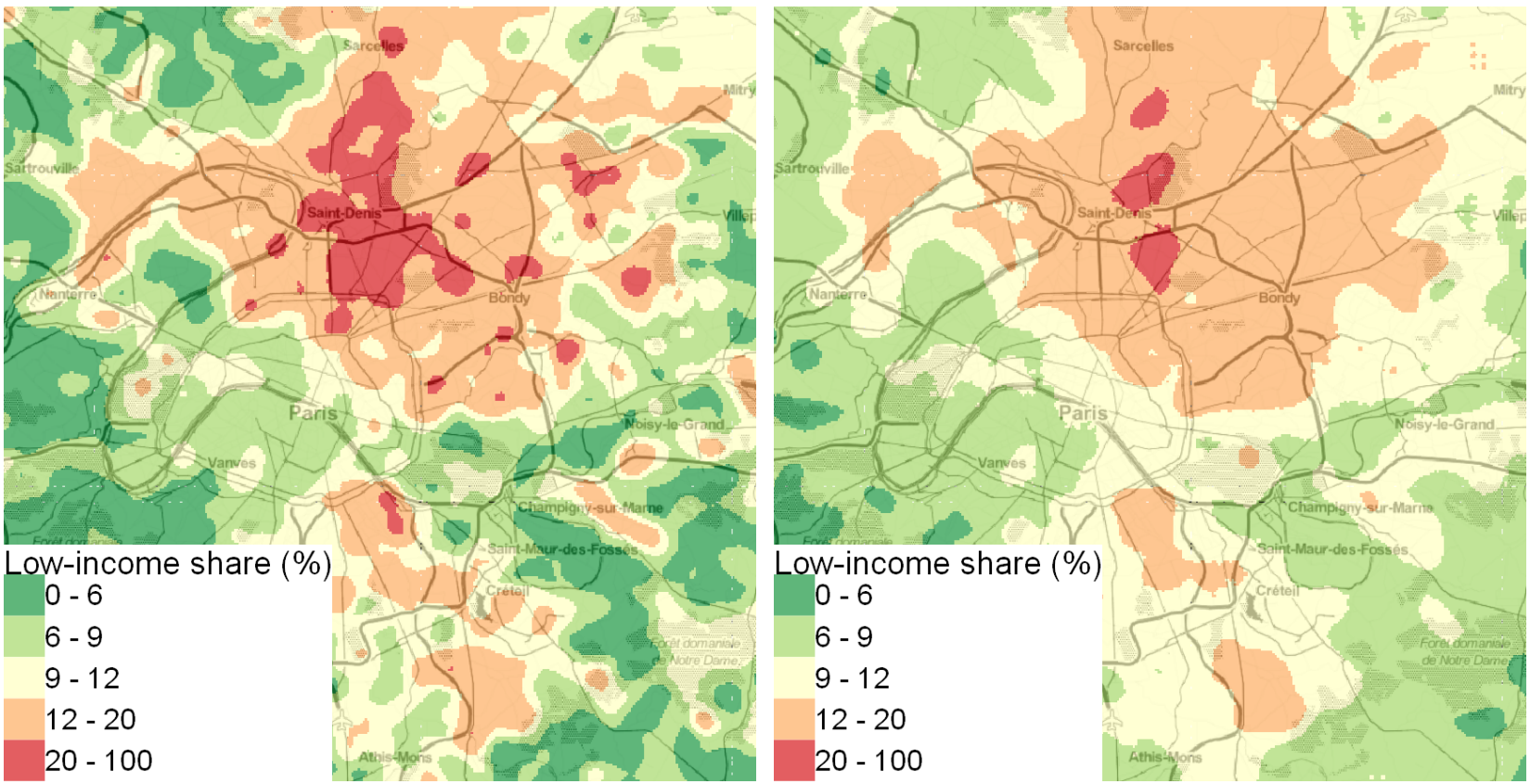
- Remainder, standard index:

$$H = \frac{n_c}{N^{\text{city}}} \sum_{c=1}^{C} \frac{E(p^{\text{city}}) - E(p_c)}{E(p^{\text{city}})}$$

Outline

Evolution of city structure during daytime

e.g. Low-income concentration at two different times (6am and 4pm). Full sequence here



Segregation dynamics

- City-level segregation evolution along 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
- Strong decrease in segregation indices
- Relative difference btw daytime and nighttime stronger for low-income

	Paris		Lyon		Marseille	
	Low-income	High-income	Low-income	High-income	Low-income	High-income
			Weekdays			
Max amplitude	0.05	0.12	0.06	0.1	0.08	0.11
Relative amplitude (%)	76.72	68.77	88.03	82.03	77.82	77.8
Within night (19h-9h) relative amplitude (%)	61.67	55.5	71.46	65.56	64.46	61.48

Max amplitude $H^{max} - H^{min}$
Relative amplitude (%): $1 - H^{min}/H^{max}$

Segregation dynamics: low-income

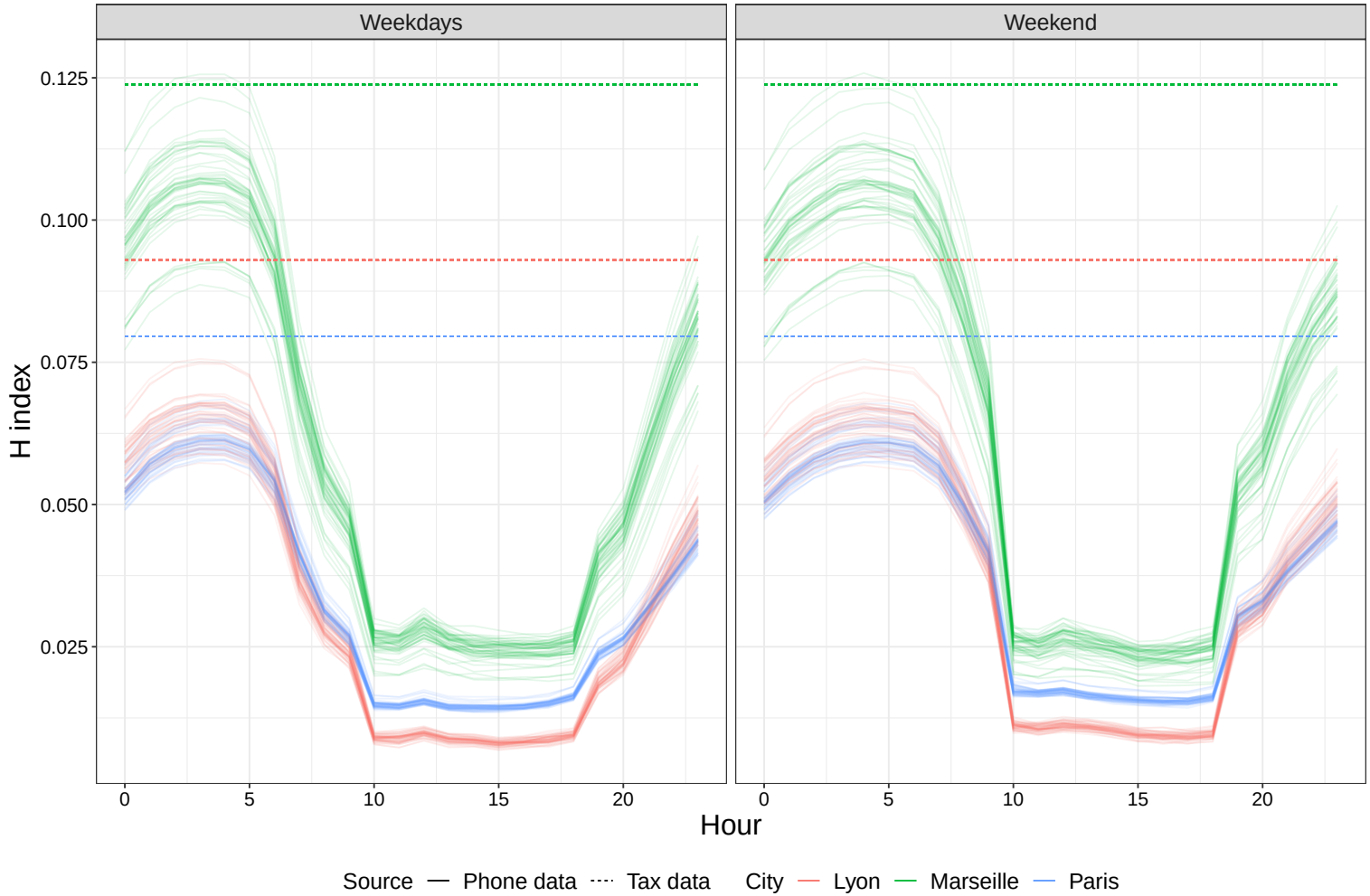


Figure 4: Low-income segregation dynamics (50 replications bootstrap)

Segregation dynamics: high-income

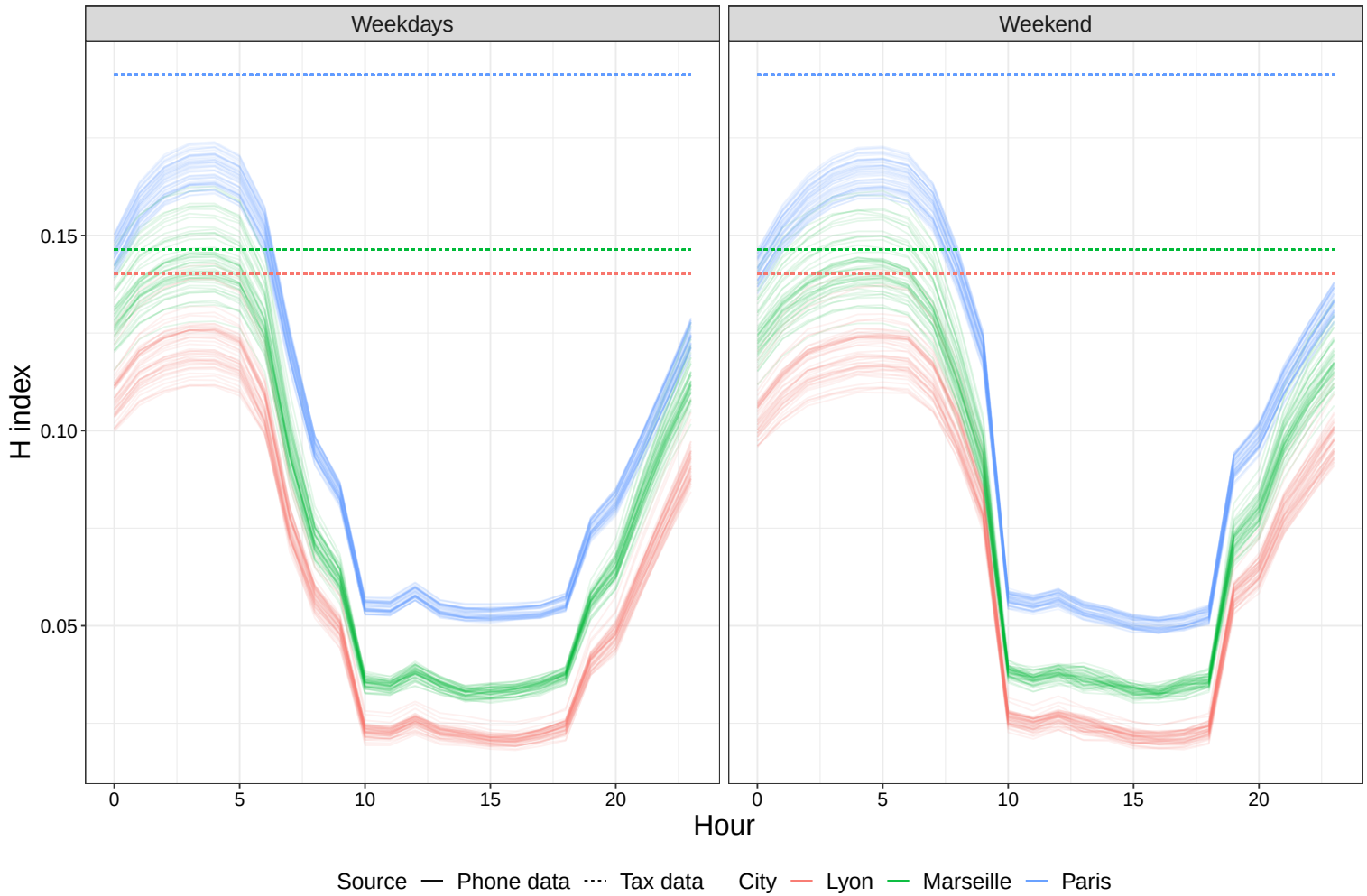


Figure 5: High-income segregation dynamics (50 replications bootstrap)

Segregation dynamics: comparing cities and income groups

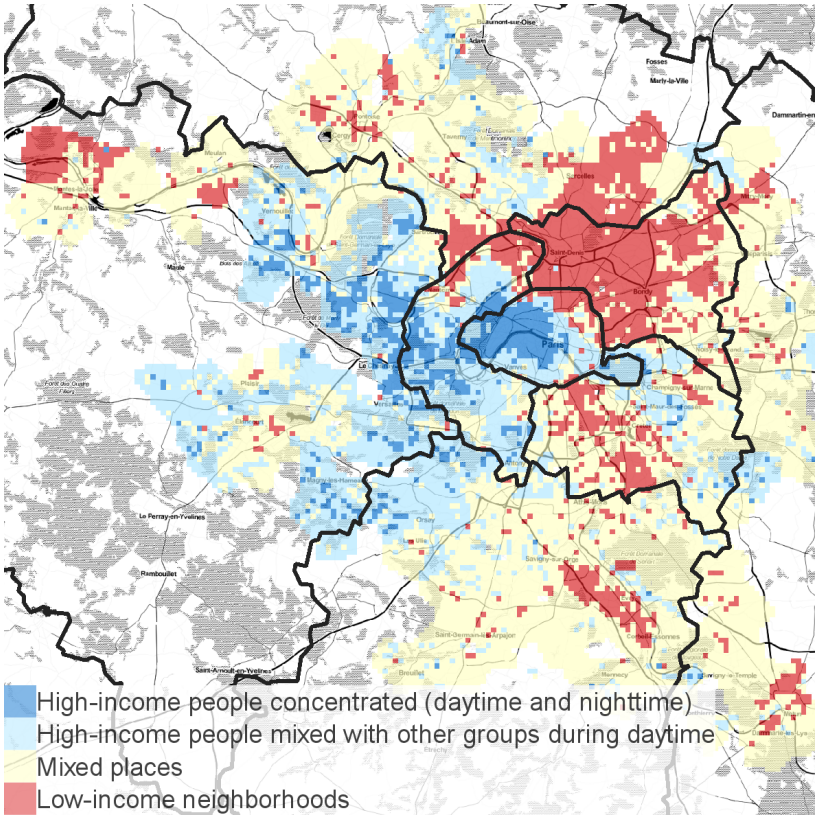
- Significant difference between nighttime and daytime segregation levels
  - Segregation starts to decrease around 6-7am and goes up after 4-5pm
  - No significant difference between weekend and weekdays $\Rightarrow$ separate saturday and sunday ?
- Differences in level observed in tax data also present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

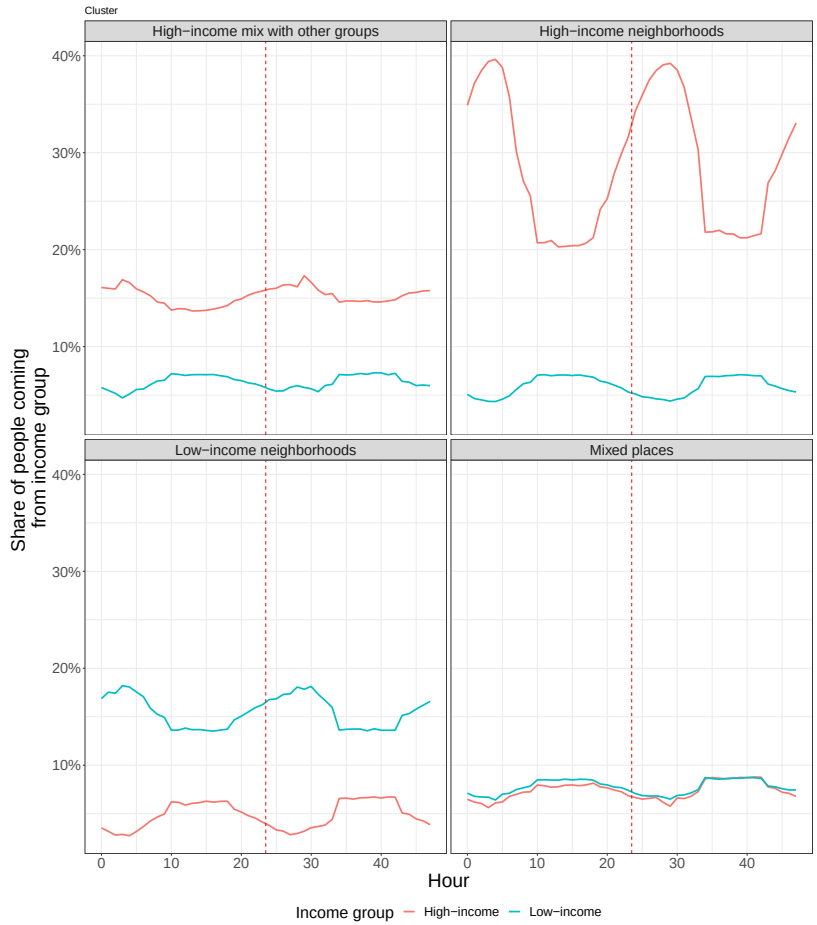
Outline

Principle

- Clustering to identify places that share common population composition characteristics
  - Will be related to places characteristics (infrastructures...)
- Sequence of low-income and high-income concentrations in our cells
- K-means classification algorithm
- 4 clusters is motivated by the trade-off between variety and parsimony.

Results


High-income people concentrated (daytime and nighttime)
High-income people mixed with other groups during daytime
Mixed places
Low-income neighborhoods


Cluster
High−income mix with other groups
High−income neighborhoods
Low−income neighborhoods
Mixed places
Share of people coming from income group
40%
30%
20%
10%
0
10
20
30
40
Hour
Income group
High−income
Low−income

Outline

Gravity model with origin-destination flows

$$p_{i \to j}^{g} = a \frac{M_i^{\beta_1} M_j^{\beta_2}}{D_{ij}^{\beta_3}} \tag{1}$$

- Mobile phone literature refer to gravity equation (e.g. Krings *et al*, 2009)
  - Does not estimate distance-decay with robust methodology
  - Some common caveats of gravity equation (e.g. zero-flows problem) need to be accounted
- We observe only strictly positive flows (censoring problem)
  - Loglinearized OLS equations are biased
- Silva & Tenreyro (2006) and Silva & Tenreyro (2011):
  - Augment observed sample with every potential flows
  - Count data models (Pseudo Poisson ML) more suited than a log-linearized OLS equation
- When large share of zeros (our case): zero-inflated count model

Gravity model with origin-destination flows

- We propose to use estimation strategies derived from international trade theory...
- ... with urban flows measured using mobile phone data
  - Likelihood of being in cell c_i knowing people live in cell c_j
  - Origin-destination flows at 500 meters level
- Estimate heterogeneity in distance costs:
  - Spatial dimension: suburbs vs center
  - Social dimension: low-income vs high-income
- Estimation on a 5% sample to speed up computations (robust to full sample)

$$\text{(selection)} \qquad \mathbb{P}(p_{i \to j} > 0) = 1 - \pi_{ij} = \frac{\exp(Z_{ij}\gamma)}{1 + \exp(Z_{ij}\gamma)}$$

$$\text{(outcome)} \qquad \lambda_i(X_{ij}) = \mathbb{E}_{f,\theta}(p_{i \to j} | X_{ij}) = \exp(X_{ij}\beta)$$

Outline

Reader's digest

- Urban structure plays a key role to understand flows intensity
  - Paris and Lyon: more costly to move from suburban areas to another suburban area
  - Marseille: travel cost higher in city center
- Related to public transportation
- Low-income people tend to live in places where spatial frictions are higher
  - Controlling for origin, income plays a more marginal role
  - Low-income people less likely to move (selection model)
  - But the gap between coefficients is more limited than differences between suburbs and center

Results (Marseille)

	Dependent variable:			
	Low-income		High-income	
	Selection	Outcome	Selection	Outcome
p_j^{D1} in destination cell (tax data)	1.505^{***}	0.428^{***}	1.861^{***}	0.322^{**}
	(0.269)	(0.148)	(0.265)	(0.145)
p_j^{D9} in destination cell (tax data)	-1.064^{***}	0.204^{**}	-0.811^{***}	-0.029
	(0.138)	(0.094)	(0.131)	(0.090)
Distance (suburbs $\rightarrow$ suburbs)	-2.439^{***}	-1.070^{***}	-2.328^{***}	-1.023^{***}
	(0.028)	(0.017)	(0.026)	(0.018)
Distance (center $\rightarrow$ suburbs)	-2.594^{***}	-1.121^{***}	-2.438^{***}	-1.069^{***}
	(0.026)	(0.014)	(0.023)	(0.013)
Distance (suburbs $\rightarrow$ center)	-2.106^{***}	-1.105^{***}	-2.041^{***}	-1.033^{***}
	(0.026)	(0.016)	(0.024)	(0.016)
Distance (center $\rightarrow$ center)	-2.424^{***}	-1.614^{***}	-2.195^{***}	-1.611^{***}
	(0.035)	(0.016)	(0.032)	(0.015)
α (dispersion)	1.6		1.5	
Observations	1,368,224		1,333,298	
Log likelihood (by obs.)	-0.1		-0.1	

Note: *p<0.1; **p<0.05; ***p<0.01

Dependent variable $p^{g}_{i\rightarrow j}$: low (resp. high) income density in cell c_j that live in cell c_i

Other controls: population in home cell, population in destination cell, employment in home cell, employment in destination cell

Results (Paris)

	Dependent variable:			
	Low-income		High-income	
	Selection	Outcome	Selection	Outcome
p_j^{D1} in destination cell (tax data)	0.302^{***}	-1.038^{***}	0.419^{***}	-1.101^{***}
	(0.093)	(0.049)	(0.092)	(0.047)
p_j^{D9} in destination cell (tax data)	1.518^{***}	0.773^{***}	1.488^{***}	0.695^{***}
	(0.054)	(0.031)	(0.053)	(0.030)
Destination cell belongs to cluster 1 (where high-income people are over-represented)	0.597^{***}	0.066^{***}	0.656^{***}	0.054^{***}
	(0.019)	(0.009)	(0.019)	(0.009)
Destination cell belongs to cluster 2 (where low-income people are over-represented)	-0.102^{***}	-0.224^{***}	-0.023^{*}	-0.270^{***}
	(0.014)	(0.008)	(0.014)	(0.008)
Destination cell belongs to cluster 3 (where high-income mix with middle class)	-0.154^{***}	0.111^{***}	-0.222^{***}	0.114^{***}
	(0.014)	(0.010)	(0.014)	(0.010)
Distance (suburbs $\rightarrow$ suburbs)	-2.623^{***}	-1.649^{***}	-2.567^{***}	-1.665^{***}
	(0.009)	(0.004)	(0.008)	(0.004)
Distance (center $\rightarrow$ suburbs)	-2.398^{***}	-1.377^{***}	-2.299^{***}	-1.371^{***}
	(0.009)	(0.004)	(0.009)	(0.004)
Distance (suburbs $\rightarrow$ center)	-1.682^{***}	-1.430^{***}	-1.629^{***}	-1.438^{***}
	(0.009)	(0.004)	(0.009)	(0.004)
Distance (center $\rightarrow$ center)	-1.995^{***}	-1.115^{***}	-1.959^{***}	-1.195^{***}
	(0.014)	(0.007)	(0.015)	(0.007)
α (dispersion)	1.6		1.5	
Observations	8,430,820		8,426,330	
Log likelihood (by obs.)	-0.1		-0.1	

Note: *p<0.1; **p<0.05; ***p<0.01

Dependent variable $p^{g}_{i\rightarrow j}$: low (resp. high) income density in cell c_j that live in cell c_i

Other controls: population in home cell, population in destination cell, employment in home cell, employment in destination cell

Results (Lyon)

	Dependent variable:			
	Low-income		High-income	
	Selection	Outcome	Selection	Outcome
p_j^{D1} in destination cell (tax data)	2.187^{***}	-0.479^{***}	1.009^{***}	-0.392^{***}
	(0.330)	(0.151)	(0.321)	(0.146)
p_j^{D9} in destination cell (tax data)	-0.573^{***}	0.538^{***}	-0.973^{***}	0.790^{***}
	(0.134)	(0.087)	(0.144)	(0.088)
Distance (suburbs $\rightarrow$ suburbs)	-3.023^{***}	-1.860^{***}	-3.367^{***}	-1.858^{***}
	(0.035)	(0.020)	(0.041)	(0.020)
Distance (center $\rightarrow$ suburbs)	-2.665^{***}	-1.525^{***}	-3.038^{***}	-1.530^{***}
	(0.030)	(0.014)	(0.037)	(0.015)
Distance (suburbs $\rightarrow$ center)	-1.713^{***}	-1.553^{***}	-1.628^{***}	-1.583^{***}
	(0.031)	(0.016)	(0.041)	(0.018)
Distance (center $\rightarrow$ center)	-1.939^{***}	-1.251^{***}	-2.345^{***}	-1.242^{***}
	(0.034)	(0.017)	(0.039)	(0.017)
α (dispersion)	1.4		1.5	
Observations	860,362		849,734	
Log likelihood (by obs.)	-0.1		-0.1	

Note: *p<0.1; **p<0.05; ***p<0.01

Dependent variable $p^{g}_{i \rightarrow j}$: low (resp. high) income density in cell c_j that live in cell c_i

Other controls: population in home cell, population in destination cell, employment in home cell, employment in destination cell

Outline

Conclusion

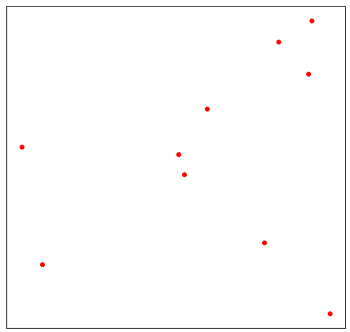
- Bringing together phone and tax data requires methodological foundations
- Segregation:
  - Acme during nighttime/hometime
  - Goes down by $\approx 70\%$ by daytime
  - Results consistent with Davis et al (2019) and Athey et al (2019)
- Mobility cost:
  - Depends on urban structure: Marseille vs Paris/Lyon
  - Some heterogeneity given neighborhood income level: e.g. low-income neighborhood in Marseille
  - Low-income people more concentrated where spatial frictions are higher

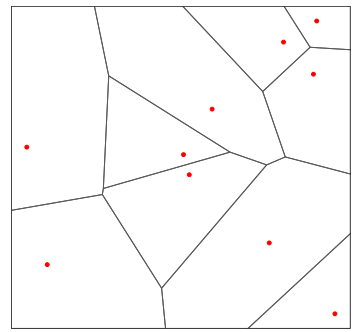
Outline

Appendix

Phone users' presence probabilization

Back to slide



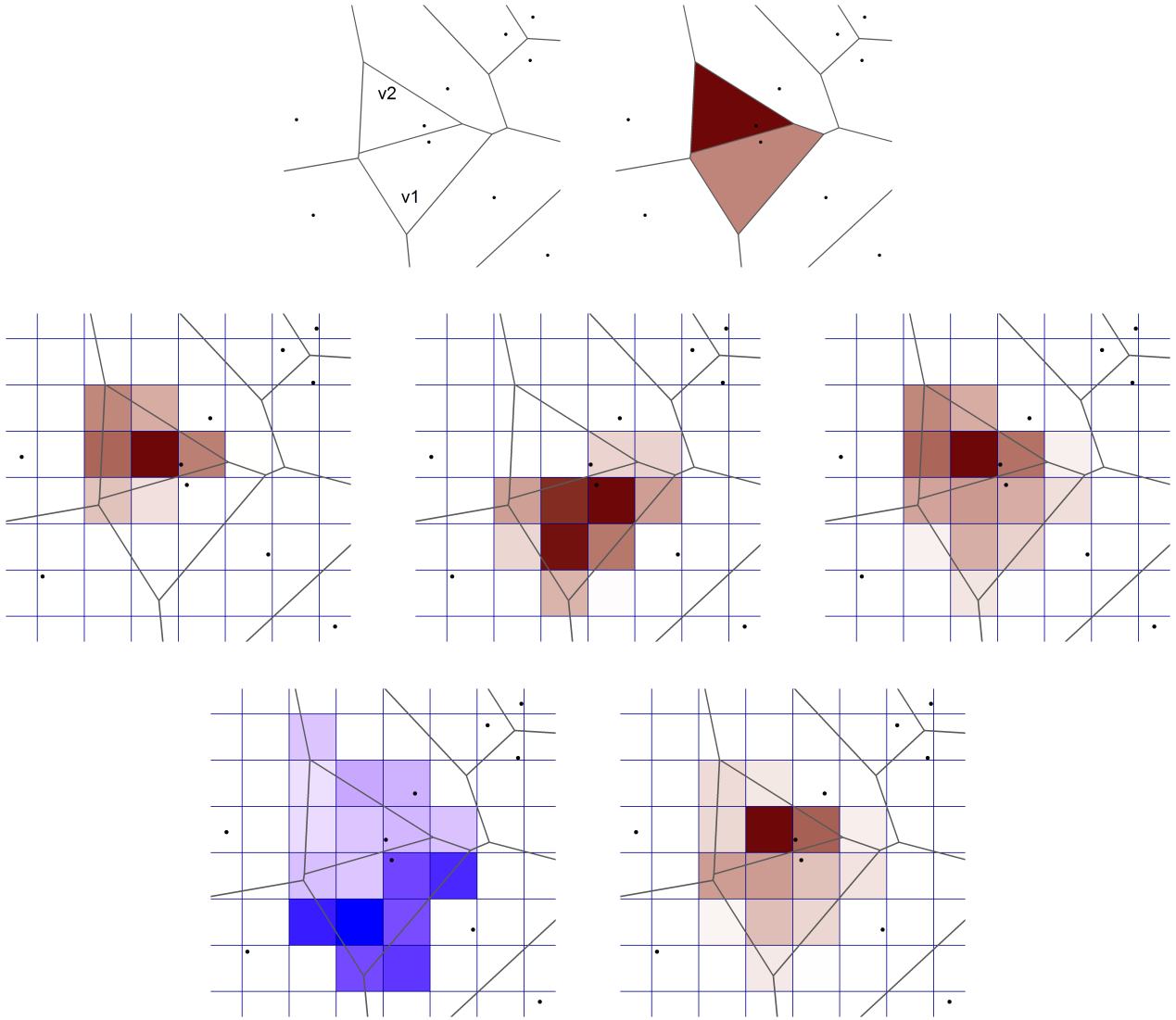


- Mobile phone litterature does not dissociate:
  - Coverage area: observations at antenna level into presence area
  - Statistical unit: economic information level
- Coverage area: Voronoi tesselation
  - Each point in space is associated with closest antenna
- However, must not be analysis statistical 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
- Illustration in next slide for home detection:
  - 2/3 events located in v_2 ; 1/3 located in v_1
  - Grid probabilities $(\mathbb{P}(c_i|v_j))_{i,j}$ via Bayes' rule (see (c) and (d))
  - With uninformative prior, home detection given by (e)
  - If population denser in tiles that intersect v_1 (f), home detection is modified (g)

Phone users' presence probabilization



Outline

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

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

- 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: $c_i = \arg\max_{c_i} \nu_x^{\texttt{home}}(c_i)$
`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

$$\forall c_{it} \in \mathcal{C}, \quad \mathbb{P}_x(c_{it}) = \sum_{v_{jt} \in \mathcal{V}} \mathbb{P}(c_{it}|v_{jt}) \mathbb{P}_x(v_{jt}) \tag{2}$$

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

Back to presentation

Outline

Additional elements: spatial clustering

Back to slide

- 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