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)

Workshop on Social Mobility, CEPR-AMSE-Banque de France

November 15th, 2019

Introduction

Introduction

Residential segregation drivers: housing

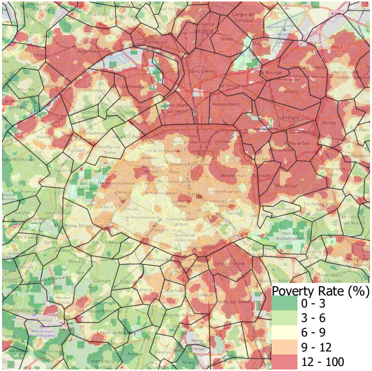
- Income gradient from housing prices (Alonso, 1964)
  - High opportunity cost of transportation: wealthiest live in city center, poorest in suburbs
  - High valuation of housing space: wealthiest live in suburbs, poorest in city center
- Social housing aims to ensure social mixing
  - Social housing 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

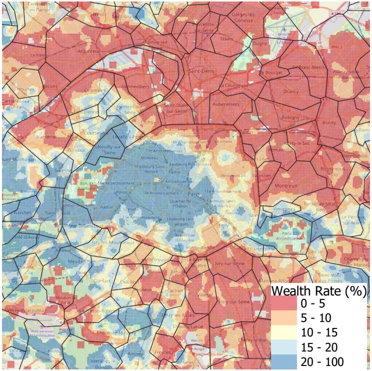
- Heterogeneity in preferences have spatial effects
  - Schelling (1969): clustering based on preference for neighborhood
  - Tiebout (1956): spatial sorting based on public goods preferences
- Mobility plays a key role to understand segregation
  - Long run: high quality public good bring people in 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

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

- 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?
  - 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

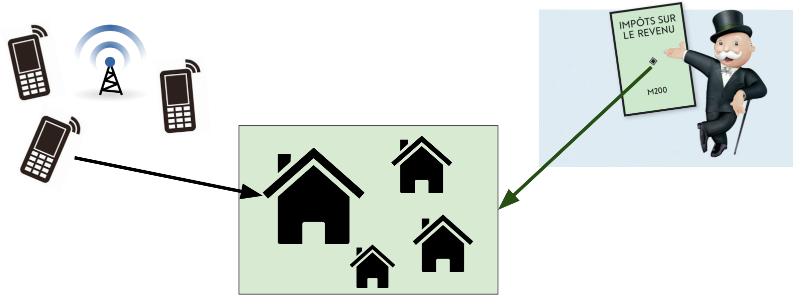
Methodology adopted

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

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

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

Tax data

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

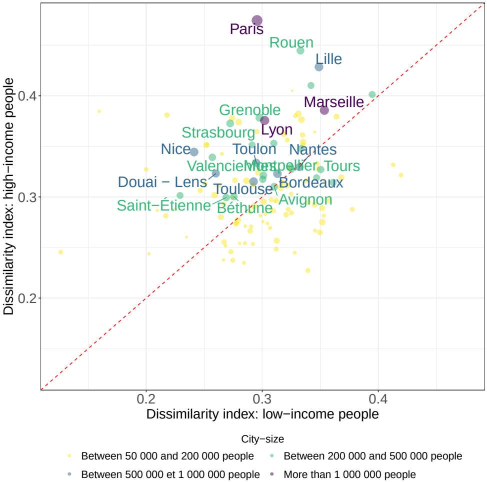


Figure 2: Dissimilarity index for main French cities

Dynamic segregation

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

- 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
  - Alternative simulation: nighttime level affected but dynamics keep the same pattern

	Paris		Lyon		Marseille	
	Low-income	High-income	Low-income	High-income	Low-income	High-income
			Weekdays			
Max amplitude	0.13	0.18	0.15	0.2	0.16	0.19
Relative amplitude (%)	51.9	41.95	61.58	55.24	47.77	45.95
Within night (19h-9h) relative amplitude (%)	37.18	31.53	43.65	39.58	36.8	32.69
			Weekend			
Max amplitude	0.13	0.19	0.15	0.19	0.17	0.19
Relative amplitude (%)	49.59	43.73	59.09	53.22	49.18	45.64
Within night (19h-9h) relative ampli-	28.29	24.99	30.6	28.23	28.3	25.22

Segregation dynamics: low-income

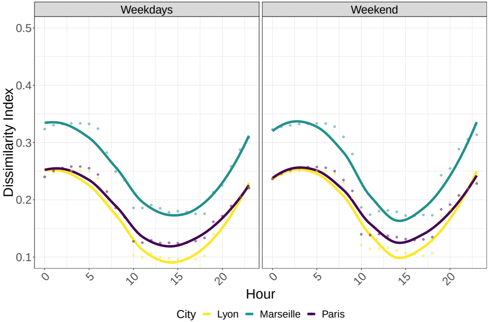


Figure 3: Low-income segregation dynamics

Segregation dynamics: high-income

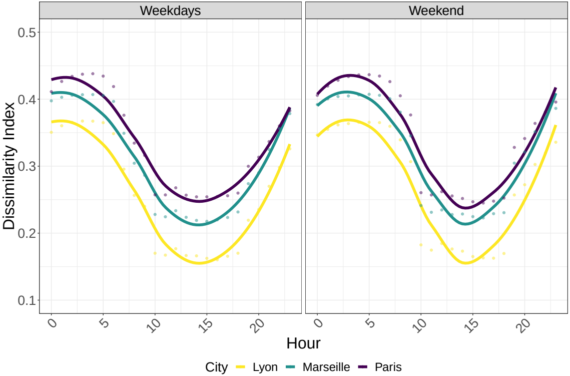


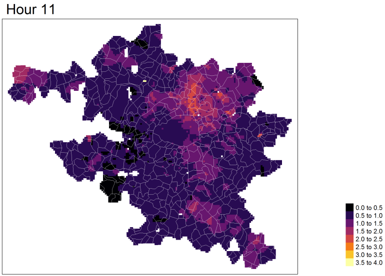
Figure 4: High-income segregation dynamics

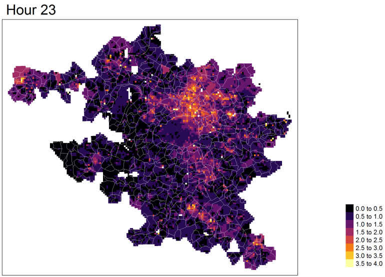
Segregation dynamics: comparing cities and income groups

- 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 ⇒ 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 Marseille and Lyon than in Paris
  - Track neighborhood composition Results here
  - 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)





Gravity model from urban flows

Specification

Gravity model with origin-destination flows

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

- 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
  - ML Count data models 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

$$
\begin{aligned}
& & p^{g}_{i\to j} &\sim \pi\delta_0 + (1-\pi)\mathcal{NB}(\lambda) \\
&\text{(Selection)} & \text{probit}(\pi) &= \textstyle\sum_i \alpha_i z_i + u_i \\
&\text{(Outcome)} & \lambda(X_{ij}) &= \mathbb{E}_{f,\theta}(p^{g}_{i\to j}|X_{ij}) = \exp(\beta^\intercal X_{ij})
\end{aligned}
\tag{2}
$$

- BIC: Choice of negative binomial (outcome) with probit link (selection)

Results

Results (Marseille)

	Dependent variable:			
	Low-income		High-income	
	Selection	Outcome	Selection	Outcome
	(1)	(2)	(3)	(4)
p_j^{D1} in destination cell (tax data)	-0.610^{***}	-0.364^{***}	-0.622^{***}	-0.347^{***}
	(0.036)	(0.063)	(0.050)	(0.046)
p_j^{D9} in destination cell (tax data)	0.104^{***}	0.560^{***}	0.080^{***}	0.575^{***}
	(0.020)	(0.044)	(0.024)	(0.029)
Distance (suburbs $\rightarrow$ suburbs)	0.999^{***}	-1.545^{***}	1.133^{***}	-1.833^{***}
	(0.004)	(0.003)	(0.005)	(0.008)
Distance (center $\rightarrow$ suburbs)	1.142^{***}	-1.440^{***}	1.194^{***}	-1.895^{***}
	(0.003)	(0.004)	(0.004)	(0.008)
Distance (suburbs $\rightarrow$ center)	0.872^{***}	-1.389^{***}	1.031^{***}	-1.701^{***}
	(0.004)	(0.004)	(0.004)	(0.007)
Distance (center $\rightarrow$ center)	1.248^{***}	-2.258^{***}	1.223^{***}	-1.899^{***}
	(0.005)	(NaN)	(0.006)	(0.007)
Observations	11,503,616		11,503,504	
Bayesian information criterion	1,832,368		1,814,426	
Log Likelihood	-915,997.0		-907,026.2	

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
	(1)	(2)	(3)	(4)
p_j^{D1} in destination cell (tax data)	-0.351^{***}	-0.804^{***}	-0.356^{***}	-0.817^{***}
	(0.028)	(0.019)	(0.027)	(0.019)
p_j^{D9} in destination cell (tax data)	-0.940^{***}	0.923^{***}	-0.935^{***}	0.909^{***}
	(0.016)	(0.012)	(0.015)	(0.011)
Distance (suburbs $\rightarrow$ suburbs)	1.480^{***}	-2.181^{***}	1.455^{***}	-2.277^{***}
	(0.002)	(0.002)	(0.002)	(0.002)
Distance (center $\rightarrow$ suburbs)	1.406^{***}	-1.815^{***}	1.416^{***}	-1.833^{***}
	(0.002)	(0.002)	(0.002)	(0.002)
Distance (suburbs $\rightarrow$ center)	1.020^{***}	-1.701^{***}	1.083^{***}	-1.700^{***}
	(0.002)	(0.002)	(0.002)	(0.002)
Distance (center $\rightarrow$ center)	1.120^{***}	-1.648^{***}	1.143^{***}	-1.637^{***}
	(0.004)	(0.003)	(0.004)	(0.003)
Observations	114,769,412		114,533,911	
Bayesian information criterion	20,149,001		20,247,335	
Log Likelihood	-10,074,287.2		-10,123,454.3	

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
	(1)	(2)	(3)	(4)
p_j^{D1} in destination cell (tax data)	-0.411^{***}	-0.231^{***}	-0.432^{***}	-0.295^{***}
	(0.050)	(0.049)	(0.066)	(0.046)
p_j^{D9} in destination cell (tax data)	0.062^{***}	0.680^{***}	0.041^{***}	0.660^{***}
	(0.024)	(0.029)	(0.030)	(0.027)
Distance (suburbs $\rightarrow$ suburbs)	1.620^{***}	-2.018^{***}	1.640^{***}	-2.039^{***}
	(0.005)	(0.007)	(0.007)	(0.007)
Distance (center $\rightarrow$ suburbs)	1.536^{***}	-1.818^{***}	1.539^{***}	-1.815^{***}
	(0.005)	(0.005)	(0.006)	(0.005)
Distance (suburbs $\rightarrow$ center)	0.946^{***}	-1.626^{***}	0.926^{***}	-1.618^{***}
	(0.005)	(0.005)	(0.007)	(0.005)
Distance (center $\rightarrow$ center)	1.069^{***}	-1.476^{***}	1.081^{***}	-1.484^{***}
	(0.006)	(0.006)	(0.007)	(0.005)
Observations	10,795,189		10,691,215	
Bayesian information criterion	2,234,960		2,223,427	
Log Likelihood	-1,117,293.9		-1,111,527.6	

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Dependent variable $p_{i \rightarrow j}^{g}$: 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

Conclusion

Conclusion

- Bringing together phone and tax data requires methodological foundations
- Segregation:
  - Acme during nighttime/hometime
  - Goes down by $\approx 50\%$ 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

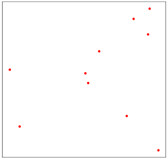
Appendix

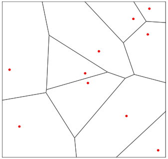
Appendix

Probabilization

Phone users' presence probabilization

Back to slide



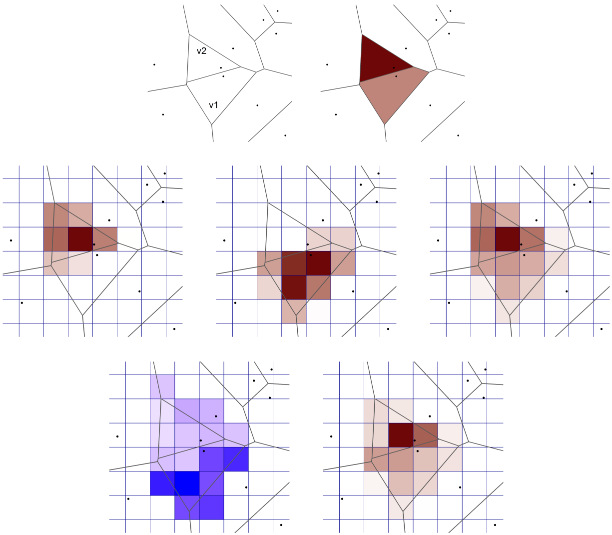


- Mobile phone litterature does not dissociate:
  - Coverage area: observations at antenna level into presence area
  - Statistical unit: economic information level
- Coverage area: Voronoi tesselation
  - Each point in space is associated with closest antenna
- However, must not be analysis statistical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



Methodology: more details

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 ν_x^{home}
One stage simulation	Cell where probability is maximum: $c_i = \arg\max_{c_i} \nu_x^{\mathrm{home}}(c_i)$
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

$$\forall c_{it} \in \mathcal{C}, \quad \mathbb{P}_x(c_{it}) = \sum_{v_{jt} \in \mathcal{V}} \mathbb{P}(c_{it} | v_{jt}) \mathbb{P}_x(v_{jt}) \tag{3}$$

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

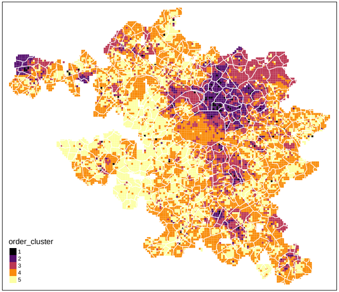
Back to presentation

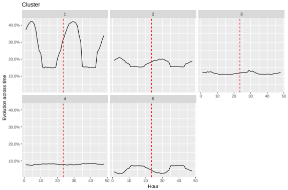
Additional elements: spatial clustering

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





Additional elements: spatial clustering

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%