Documents de travail

N° G2020/12
Novembre 2020

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

Lino GALIANA - Benjamin SAKAROVITCH
François SÉMÉCURBE - Zbigniew SMOREDA

Insee
Mesurer pour comprendre

Institut national de la statistique et des études économiques

G2020/12

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

Lino GALIANA* Benjamin SAKAROVITCH** François SÉMÉCURBE** Zbigniew SMOREDA***

Novembre 2020

Département des Études Économiques – Timbre G201
88, avenue Verdier – CS 70 058 – 92 541 MONTROUGE CEDEX – France
Tél. : 33 (1) 87 69 59 54 – E-mail : d3e-dg@insee.fr – Site Web Insee : http://www.insee.fr

* Insee-Dese – Département des études économiques – Division « Redistribution et politiques sociales »
** Insee-Dmcsi – SSP-lab
** Insee-Dmcsi – SSP-lab
*** Orange Labs, SENSE

Remerciements :
Authors are grateful to Gregory Verdugo, Kurt Schmidheiny, Florence Goffette-Nagot, Donald Davis, Lionel Wilner, Sébastien Roux, Elise Coudin, Didier Blanchet and Milena Suarez-Castillo for valuable comments. Authors would like to thank participants from UEA Amsterdam 2019, CEPR 2019 Workshop on Social Mobility, Insee's D2E Seminar, BigSurv 2018 conference, Netmob 2019.

Evolution de la ségrégation pendant la journée et frictions spatiales : une analyse à partir de données de téléphonie mobile

Ce travail exploite des données de téléphonie mobile afin de dépasser une vision statique (ou résidentielle) de la ségrégation et d'en proposer une vision plus dynamique en étudiant les mobilités infra-journalières des individus. Comme la diffusion des individus dans l'espace réduit la ségrégation, ce travail s'intéresse également aux frictions spatiales déterminant ces flux de population.

Par rapport à la situation nocturne, les bas revenus côtoient plus de personnes n'appartenant pas à leur groupe pendant la journée que ne le font les hauts revenus. La morphologie des villes, notamment la manière dont sont organisées les infrastructures de transport entre le centre et la périphérie affecte les distances parcourues. Les bas revenus sont principalement concentrés dans les espaces où les freins à la mobilité sont les plus forts (centre à Marseille, périphérie à Lyon et Paris).

Mots-clés : Ségrégation, *big data*, données de téléphonie mobile, modèle gravitaire, économie urbaine

---

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

We bring together mobile phone and geocoded tax data on the three biggest French cities to shed a new light on segregation that accounts for population flows. Mobility being a key factor to reduce spatial segregation, we build a gravity model on an unprecedent scale to estimate the heterogeneity in travel costs.

Residential segregation represents the acme of segregation. Low-income people spread more than high-income people during the day. Distance plays a key role to limit population flows. Low-income people live in neighbourhoods where the spatial frictions are strongest.

Keywords: Segregation, big data, phone data, gravity model, urban economics

Classification JEL : R23 ; R41 ; C55

District division is part of the mental map of every Parisian who masters its social symbolism. Living in the 19th or 16th district does not have the same meaning. In Paris, it is usual to link someone's social level with the neighborhood he or she lives in. In the city's division lines, we recognize those of society.
Pinçon and Pinçon-Charlot (2009)

1 Introduction

The growing access to high-frequency geocoded data represents an opportunity to understand phenomena that could not be studied before (Einav and Levin, 2014; Glaeser et al., 2018). In the field of urban economics, where mobility represents a dimension of individual choice, access to individual positions is fundamental for empirical research. A large literature has been devoted to the mechanisms and effects of spatial segregation since the pioneering studies of the Chicago school of sociology (Park, 1926). Segregation is known to play a key role in many social outcomes (Sampson, Morenoff, and Gannon-Rowley, 2002) and have long-term effects on the welfare of children that were raised in poor neighborhoods (Chetty and Hendren, 2018).

Until recently, because they were based on available data, most of the studies on segregation were focused on where people live. However, individuals tend to visit places outside their living environment, exposing them to people from other neighborhoods. Mixing with individuals from other neighborhoods creates opportunity for interactions that can affect an individual's welfare and thus partially offset the effects of residential segregation. To better understand *experienced segregation* (Athey et al., 2019), one needs to take into account where people go and who they mix with in those places. Since it is less costly to visit a place than to live there, experienced segregation might be less influenced by income than residential segregation. Because spatial frictions limit encounters of people from different neighbourhoods, they play a key role in understanding how experienced segregation may differ from residential segregation (Davis et al., 2019). With the surge of positional data, assessing the effect of mobility on segregation and the factors that hinder urban flows becomes an empirical question. This raises new opportunity to understand the dynamics in place within cities.

We explore the evolution of segregation in the three French biggest cities within a typical day using individual mobile phone data combined with traditional data sources. As mobile phone data are anonymous, it is necessary to detect the residence of mobile phone customers and use the surrounding area to characterise people in the income hierarchy. We propose a methodological framework to infer from residential data the characteristics of individuals present in anonymized positional data. Results stability is tested through a series of robustness checks. Our approach aims to limit the ecological fallacy, which is the incorrect imputation of income at the individual level from aggregated data. Since more and more big data sources are used in academic research (Einav and Levin, 2014; Glaeser et al., 2018), this imputation method combining aggregate and individual data is a relevant framework for social scientists and economists.

Experienced segregation is less extreme than residential segregation. Our results are consistent with the literature and results on residential segregation as well as recent estimates based on positional data (Davis et al., 2019; Athey et al., 2019). Comparing nighttime and daytime segregation provides a better understanding of how population movements and urban structure affect experienced segregation. For instance, in Paris, even during daytime, low-income people remain concentrated in the suburban areas where spatial frictions are strongest. The latter are measured using a gravity model to identify the factors determining flows between neighbourhoods. The results suggest that spatial frictions, inherited from the centre-periphery structure of cities, which in turn affects the social geography of cities, play a crucial role to determine the flows between neighbourhoods. Low-income people tend to live in places where transportation costs are higher. Income has a stronger effect on the likelihood of moving than on the intensity of flows. An econometric model without selection cannot account for this heterogeneity.

Section 2 justifies the need to measure segregation from an infra-day perspective. An important aspect of our approach is to adapt traditional measures to the structure of mobile phone data. In particular, we discuss how segregation can be measured from phone data. Section 3 presents the data used in this paper. We propose in Section 4 the framework adopted to combine phone and tax data in order to understand better how people spread-out. Results are discussed in Section 5. Section 6 introduces a gravity model to measure the effect that distance has on population flows. More details can be found in the Appendices.

2 Residential segregation: drivers and measurement

2.1 Segregation drivers

Segregation is an indirect end point in almost any model where spatial heterogeneity is introduced. Selection in the real estate market on the basis of income or wealth is a classic segregation mechanism. One of the conclusions of the Alonso's bid-rent model (Alonso, 1964; DiPasquale and Wheaton, 1996) is that housing prices drive household location choices. In a situation where opportunity costs of transportation are high, wealthy people will locate themselves close to city centers while poorer individuals will live in suburbs because of higher housing prices and rent in the center.[1] At a broader scale, models that deal with city productivity advantages (Duranton and Puga, 2003) can also generate segregation taking different forms, income based segregation as well as human capital concentration in a limited number of clusters. Housing price is a classical channel for public policies to reduce residential segregation. In France, a national law (Law on Urban Solidarity and Renewal, known as SRU law) forces municipalities with more than 50 000 people to ensure that at least 25% of the housing supply is social housing.[2] In neighborhoods where social housing is important (e.g. 20th district of Paris), gentrification leading to low-income people exclusion from city center has been more limited (Pinçon and Pinçon-Charlot, 2010). However, the effect of social housing on segregation is ambiguous. Public housing can strengthen social mixing but does not ensure that interactions between groups take place (Chamboredon and Lemaire, 1970). Social housing might even lead to the concentration of poor clusters (Verdugo and Toma, 2018). In that case, segregation would increase because of public housing.

An alternative approach consists in identifying households' preferences as driving segregation through different channels. Schelling (1969) proposed one of the most popular models in that strand of literature. If households value interacting with people with whom they share characteristics, extreme segregation between homogeneously constituted groups is an end point of residential mobility process. In that case, housing prices are not the driving force but reflect other factors, some of them being unobservable. The empirical investigation by Card, Mas, and Rothstein (2008) suggests that a white flight happens in most American cities when minority shares ranges between 5 to 20%. In a different perspective, Halbwachs (1906) considers, using early 20th century land data, that supply and demand in land market are not independent from political and sociological representations. Tiebout

[1] With lower transportation costs and higher households valuation for space, the reverse situation occurs: poor households live in city centers while wealthier people live in suburban areas. That configuration is more typical in the US than in Europe.

[2] The initial 20% social housing objective has been reevaluated to 25% in 2014. Municipalities that do not satisfy this requirement must pay a fine.

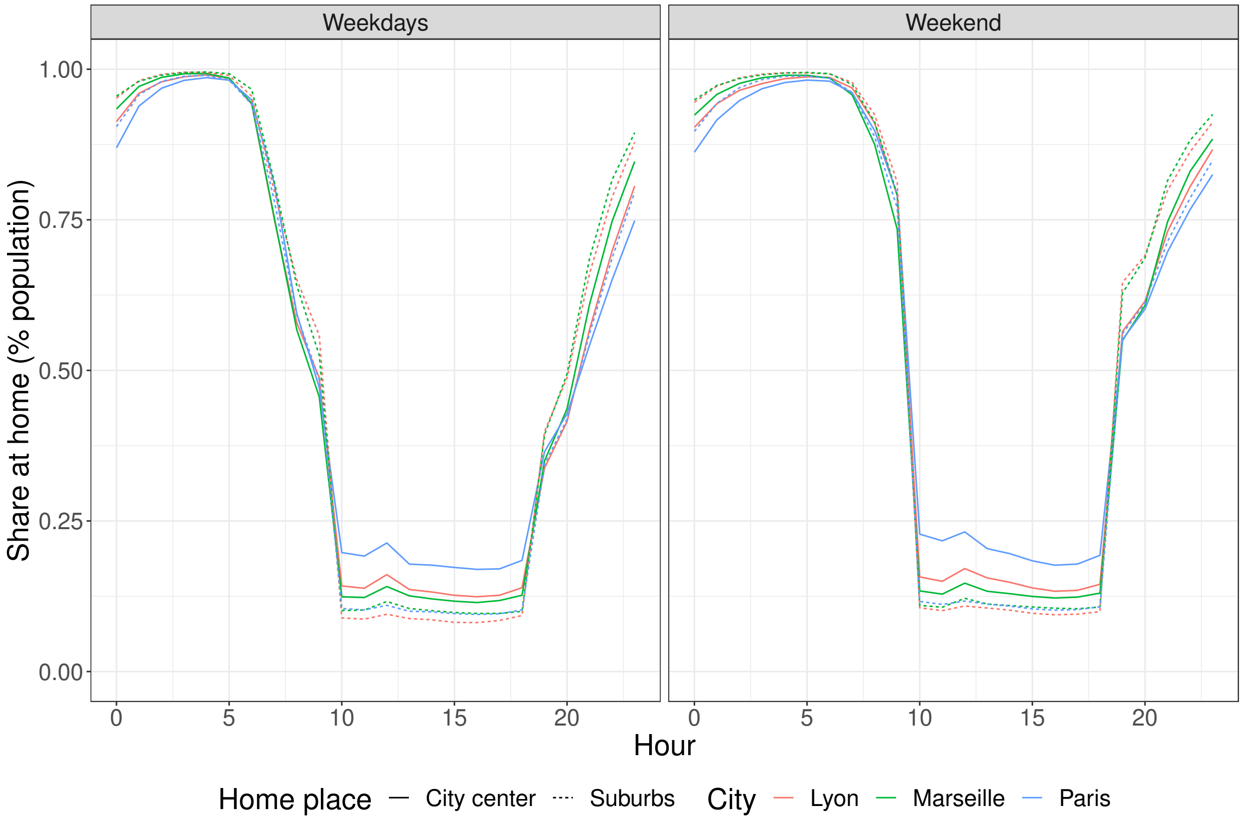


Figure 1 – Estimated share of people at home

(1956) model is also useful to understand how households' location choices creates segregation. When spaces compete over public good provision and agents' moving costs are low, homogeneous groups in terms of preferences also arise in residential spaces. For instance, using school districts borders, Black (1999) in the US and Fack and Grenet (2010) in the French case showed that housing prices are affected by school quality, leading to unevenly distributed income groups in those cities. Gentrification is an illustration of Tiebout (1956) sorting: local amenities renovations bring wealthier people into the neighborhood, leading to rising housing prices gradually excluding low-income people (Banzhaf and Walsh, 2013).

However, residential segregation is only part of the whole segregation story. Despite living on different neighborhoods, social groups can mix up at different occasions, for instance when they commute or for some activities that require moving from their living neighborhood. As shown by Figure 1, we estimate (cf. Section 4 for more details) that the share of people staying at home during daytime varies dramatically within the day. Thus, to produce a more complete view of segregation, it is important to take into account the effect of population mobility on socioeconomic groups mixing into urban space. This could be seen as a generalization of Tiebout (1956) approach where residential location as well as other spaces consumption would be suggest to choice. Recently, using Yelp data, Davis et al. (2019) have confronted consumption and residential segregation to conclude that the latter is twice higher than the former. The conclusion of the paper is quite clear on that point: *"Life in NYC is less segregated than one might infer from looking at residential segregation alone"* (Davis et al., 2019). People tend to choose restaurant on the basis of the community they belong to rather than from the neighborhood they come from. Mobility plays a key role in that conclusion because spatial frictions, which would lead people to go to restaurants closer from home, are not the main driver for choosing a restaurant in New York.

Using GPS data, Athey et al. (2019) also show that experienced segregation is significantly lower than residential segregation. Their results suggest that segregation starts to go down at 5am and increases to its acme in the evening, when most people go back to their homes. They also show that income mixing is more important in some public spaces than others. With respect to residential areas, segregation is reduced by 45% in parcs, 50% in restaurants (consistent with Davis et al., 2019). Places where diversity is maximal are movie theatres (25% of residential segregation). Employment places are also more diverse than residential areas. The results of Le Roux, Vallée, and Commenges (2017), based on survey data, also show a decrease of segregation in Paris during the day with respect to residential segregation. Since it is cheaper to travel somewhere than to live there, it is important to understand how travel changes the spatial distribution of social groups. This approach is more data-demanding. First, it requires precise data regarding people location and frequent recording of an individual's location. Second, to ensure that inference on population is robust, it is necessary to have a large sample to avoid selection biases. Census data cannot match those requirements without huge collection costs. Numeric traces generated by mobile phone activities as well as some smartphone applications data match those criteria.[3] The automatic collection of position at small marginal cost creates the possibility to record locations on a large sample.

2.2 Segregation measurement

Spatial segregation is the degree to which some groups live separately from others in a city. This general definition opens the door to a wide variety of measurement methods. The most widely used segregation indices are the dissimilarity index (Duncan and Duncan, 1955) and the information theory index also known as Theil H index (Theil, 1972). The dissimilarity index is a measure of how the distribution of a group deviates from a spatially uniform distribution. The information theory index measures the departure of neighborhood levels entropy with respect to city entropy. Both indices are evenness measures in Massey and Denton's typology (Massey and Denton, 1988) because they are based on the idea that a group is segregated if it is unevenly distributed over the space. Both indices range from 0 (all neighborhoods have the same composition) to 1 (all areas contain one group only). The dissimilarity index has the attractive feature of being interpretable as the proportion of minority members that would have to change their positions to achieve an even distribution. The Theil index value is less easy to interpret. However, the dissimilarity index is not decomposable while the Theil index is. In particular, it is possible to separate city level inequalities into within and between components and compute their contributions to overall inequality. We favor the Theil index and report dissimilarity index results, as well as other indices, into Appendix B.3.

Our Theil index is bimodal. We compute indices using the following population decomposition: low-income individuals vs others and high-income individuals vs others. Because Theil indices are invariant to overall population composition, sub-population indices can be compared together. For instance, with two mutually exclusive groups g_1 and g_2, a higher H index for g_1 means that people belonging to that group are more clustered than people from g_2.

To simplify notations, we will not use superscripts for income-group g and subscripts for time dimension t. We assume p_c is the proportion of the group of interest g into cell c at time t. We assume the city is composed of C cells. We denote p^{city} the proportion of the group of interest into the city at time t. Total population in cell c is denoted n_c. Population in the city is N^{city}. Entropy is generally defined as

$$E(p) = p \log\left(\frac{1}{p}\right) + (1-p)\log\left(\frac{1}{1-p}\right) \tag{1}$$

[3] Apps are not immune to selection bias. Because applications answer a specific need and are associated with smartphones technologies more used by youngsters, apps users datasets do not reflect population as a whole.

From that operator, it is possible to define cell-level entropy ($E(p_c)$ for $c = 1, ..., C$) and city-level entropy $E(p^{\text{city}})$. We take the convention that $0\log(\infty) = 0$. Theil index is the weighted average deviation of cell-level entropy from the city-wide entropy, expressed as a fraction of the city's total entropy[4]:

$$H = \sum_{c=1}^{C} \frac{n_c}{N^{\text{city}}} \frac{E(p^{\text{city}}) - E(p_c)}{E(p^{\text{city}})} \tag{2}$$

Since this index is based on the relative divergence from city-level composition, it is comparable between cities that do not exhibit the same income distribution or between groups whose proportions at overall level differ.

Decomposition of a city into cells usually comes from administrative or census areas. In that case, the entropy index will compare differences between census tracts and overall city composition. The problem is that administrative borders do not necessary match spatial divisions. The effect of that phenomenon on segregation indices is known as the modifiable area unit problem (MAUP) identified by Openshaw (1984).[5] In this paper, since we have access to geocoded tax data at residence level, we will not bind ourselves to adopt census tracts. This should limit the MAUP.

Decomposability of Theil index arises from the additivity of logarithm. Assume the neighborhoods are grouped together into $K < C$ clusters. Let's denote p_k^{cluster} the proportion of minority group g into cluster k. $n_{\text{k}}^{\text{cluster}}$ is the population into cluster k. We generally define the between component as

$$H_B = \sum_{k=1}^{K} \frac{n_k^{\text{cluster}}}{N^{\text{city}}} \frac{E(p^{\text{city}}) - E(p_k^{\text{cluster}})}{E(p^{\text{city}})} \tag{3}$$

It compares entropy of the clusters ($E(p_k^{\text{cluster}})$) with city level entropy ($E(p^{\text{city}})$). With the within component, it is a question of comparing the entropy of a group with that of the units that define it. We define H_k^W for each cluster k as the weighted average of entropy deviations within cluster k:

$$H_k^W = \sum_{c \in \text{cluster k}} \frac{n_c}{n_k^{\text{cluster}}} \frac{E(p_k^{\text{cluster}}) - E(p_c)}{E(p_k^{\text{cluster}})} \tag{4}$$

The first term represents the share of the population of cluster k that comes from cell c. The second term gives the difference between the group-level entropy ($E(p_k^{\text{cluster}})$) and the cell-level entropy ($E(p_c)$). Finally, one can rewrite the overall H index as the sum of a between-cluster and within-cluster components:

$$H = H_B + \sum_{k=1}^{K} \frac{n_k^{\text{cluster}}}{N^{\text{city}}} \frac{E(p_k^{\text{cluster}})}{E(p^{\text{city}})} H_k^W \tag{5}$$

More details can be found in Reardon, Yun, and Eitle (2000).

[4] By contrast, the dissimilarity index writes as

$$D = \frac{1}{2N^{\text{city}} p^{\text{city}}(1 - p^{\text{city}})} \sum_{c=1}^{C} n_c \left| p_c - p^{\text{city}} \right|$$

[5] This problem is also known as spatial aggregation problem. It is related to the fact that indices are aspatial measures of segregation because they consider that two individuals living near one another but in separate spatial units are more distant from one another than are two individuals living relatively far from one another but within the same spatial unit. One solution proposed by Reardon and O'Sullivan (2004) is to build indices from neighborhood contiguity matrices.

2.3 Residential segregation in France: what do we know from tax data ?

Before exploring the evolution of segregation within the day, it is essential to understand residential segregation. Floch (2017) proposes an interesting survey on the question. Because we use tax data, we adopt an income based segregation definition.[6] In French cities, there is an over-representation of the two extremes of income distribution with respect to the national territory. Low-income people are systematically over-represented in the biggest French cities. However, their distribution is not homogeneous as one can see in Paris (Fig. 2), Lyon (Fig. 3) and Marseille (Fig. 4). In Paris, the dichotomy between the west (wealthy places) and the east (poorer areas) is quite visible. In Marseille, low-income households are concentrated in the city center while wealthier people tend to live in the south of the city or in Aix-en-Provence, a wealthy suburban municipality. Marseille is one of the rare French cities where poor people are closer from city center than rich people. In Lyon, low-income individuals are more concentrated in the east of the agglomeration while high-income people live in the city center or at the west of the urban cluster. The difference in center-periphery structures in those three cities have effect on the distribution of high- and low-income population. In Paris, at the overall level, segregation at the top is far higher than at the bottom (Dos Santos, L'Horty, and Tovar, 2010; Oberti and Préteceille, 2016; Floch, 2017).

Time comparisons of indices are generally limited by survey releases. With the exception of Davis et al. (2019), Athey et al. (2019) and Le Roux, Vallée, and Commenges (2017), most studies regarding city diversity only understand that phenomenon through long-run evolution or city comparisons. The access to high resolution data, where individuals are observed repetitively at different places, opens new perspectives for segregation studies. In particular, this enables to determine the dynamics of segregation at fine temporal granularity and, in particular the effect of mobility on social mixing.

3 Data

3.1 Tax data

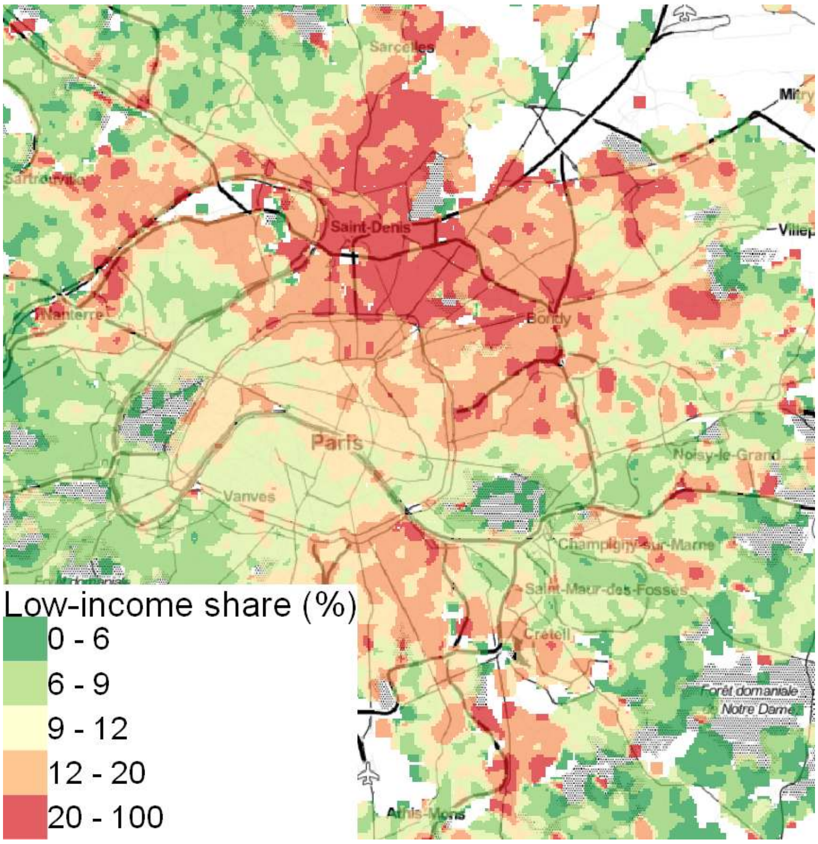
We use an exhaustive household tax dataset geocoded at (x, y) level called `Filosofi`.[7] This geocoded database gives information on the taxable income at the residence level for more than 26 million households (61.5 million individuals). We consider individuals' standard of living as defined by the ratio of disposable income over the number of consumption units.[8] By abuse of language, we will use indifferently income and standard of livings terms. Income information is used to characterize phone users' economic status following the place she is estimated to live. Geographic coordinates are used to perform aggregations at very detailed geographic level. Since household residence is geocoded into (x, y) coordinates, we define our own custom spatial granularity and do not rely on administrative borders. To ensure privacy, characterization of city neighborhoods income status is performed only if at least 11 households live inside that neighborhood.

We use the 2014 income tax year. This is the first available year where tax data were exhaus-

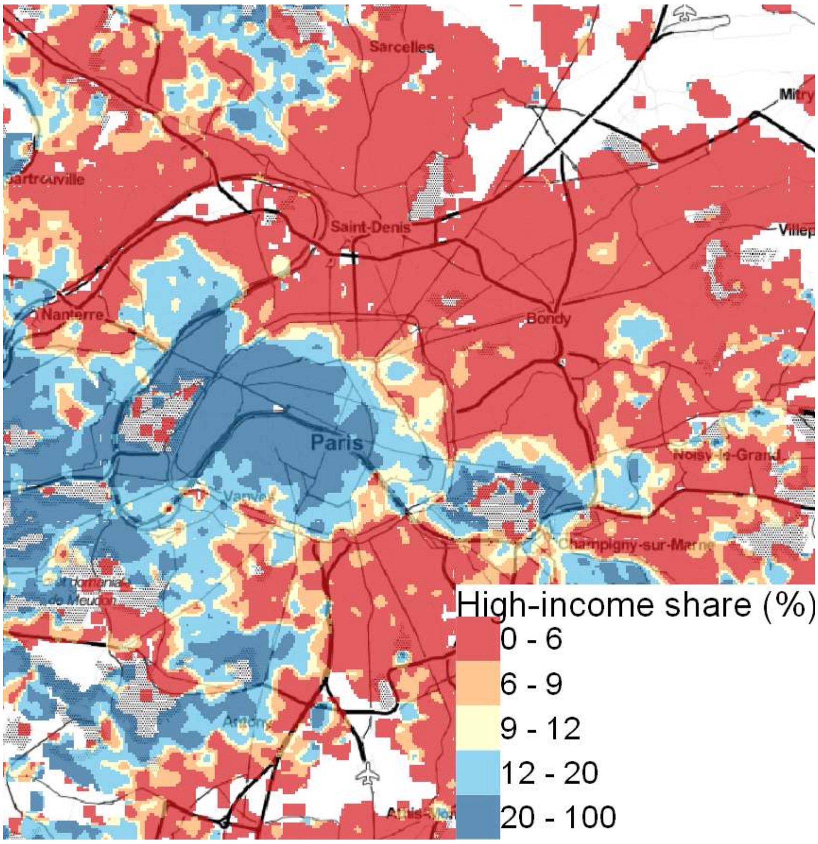
[6] Census data make possible to study other dimensions, especially the distribution of population on socioprofessional status. Several reasons led us to favor an income based approach. First, French census data are not exhaustive as the tax data we use. Second, Floch (2017) shows that indices derived from tax and census data lead to similar conclusions. Third, selection on housing market, thus on residential space, is strongly influenced by income while the link is less direct with socioprofessional status.

[7] Every French household that fills a tax form, receives social benefits or pensions belongs to that database. Thus, almost every French households is in that administrative dataset. Home is geocoded thanks to land register data.

[8] According to the modified OECD scale, first adult counts for one consumption unit. 0.5 consumption unit is added for each other person of the household aged 14 or older and 0.3 CU for the children under 14.

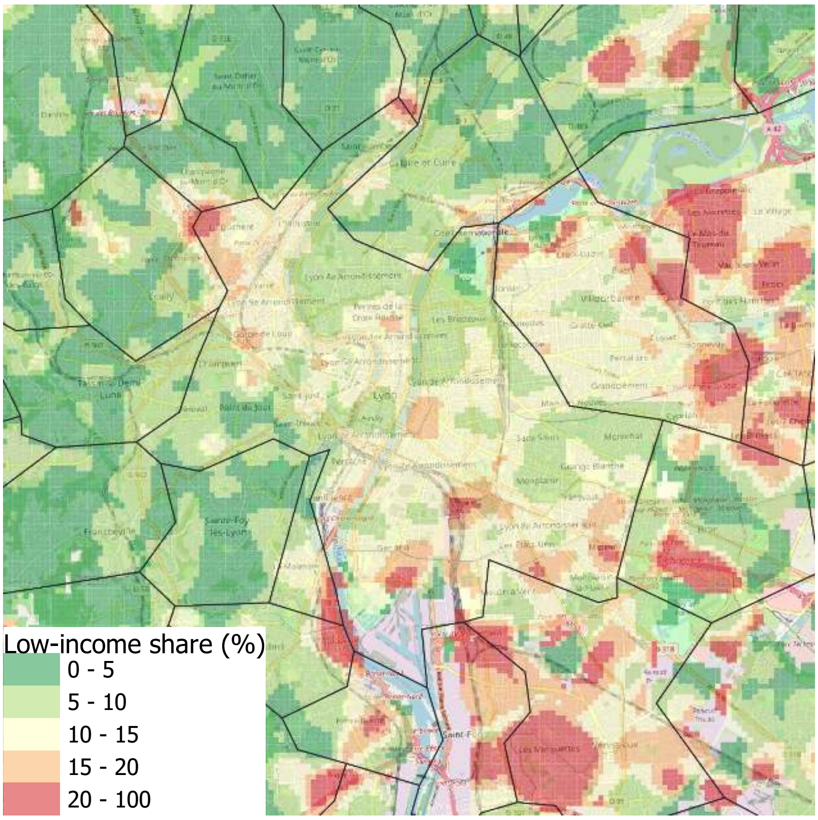


(a) Low-income population: p_i^{D1}

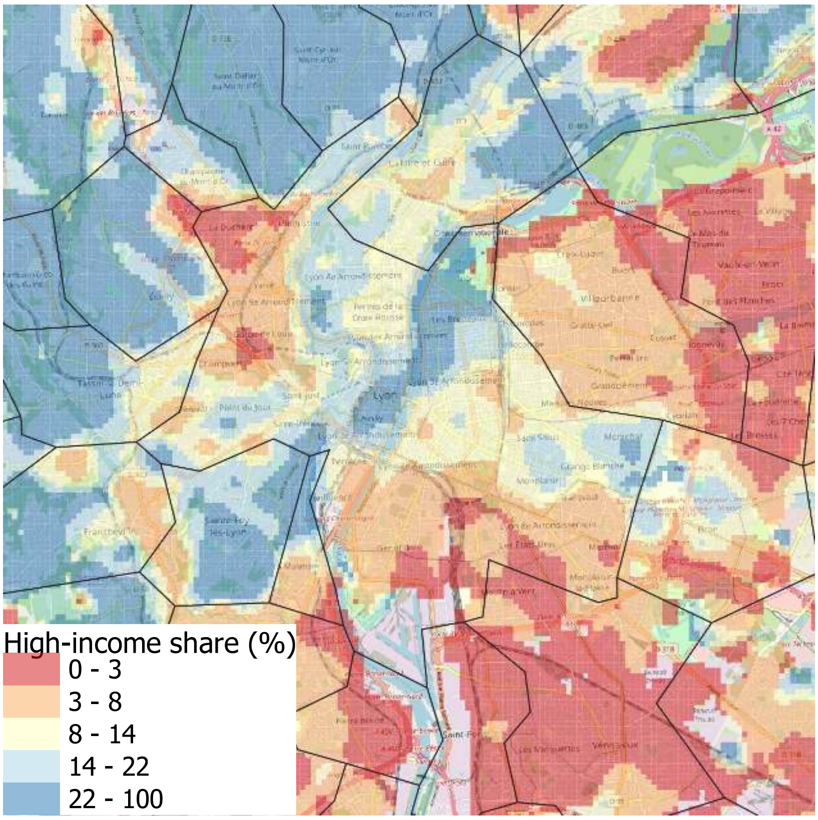


(b) High-income population: p_i^{D9}

Figure 2 – Residential segregation in Paris

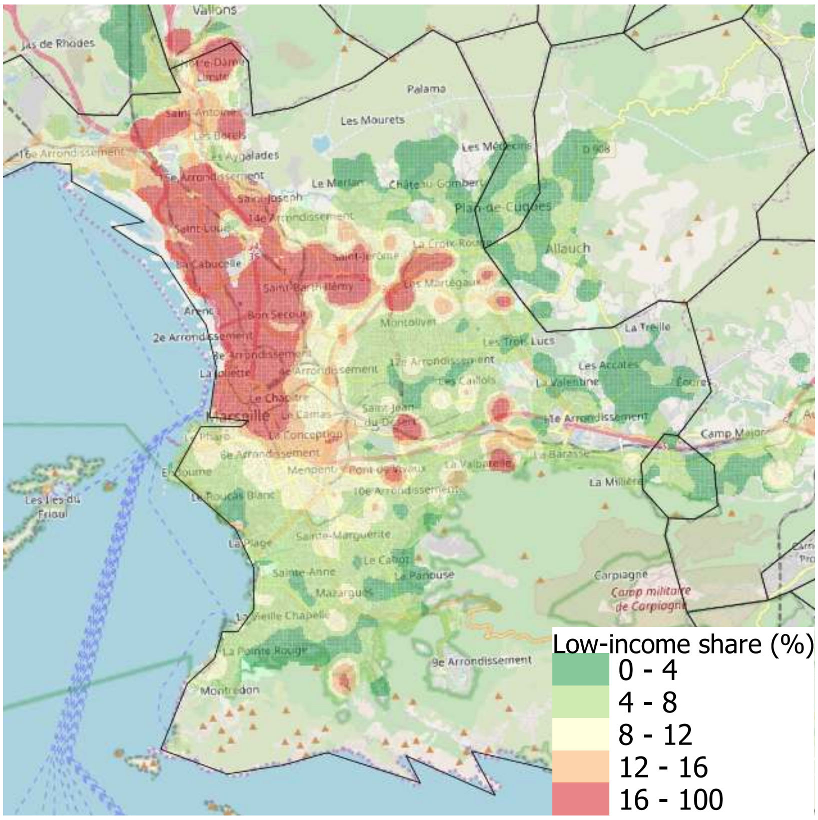


(a) Low-income population: p_i^{D1}

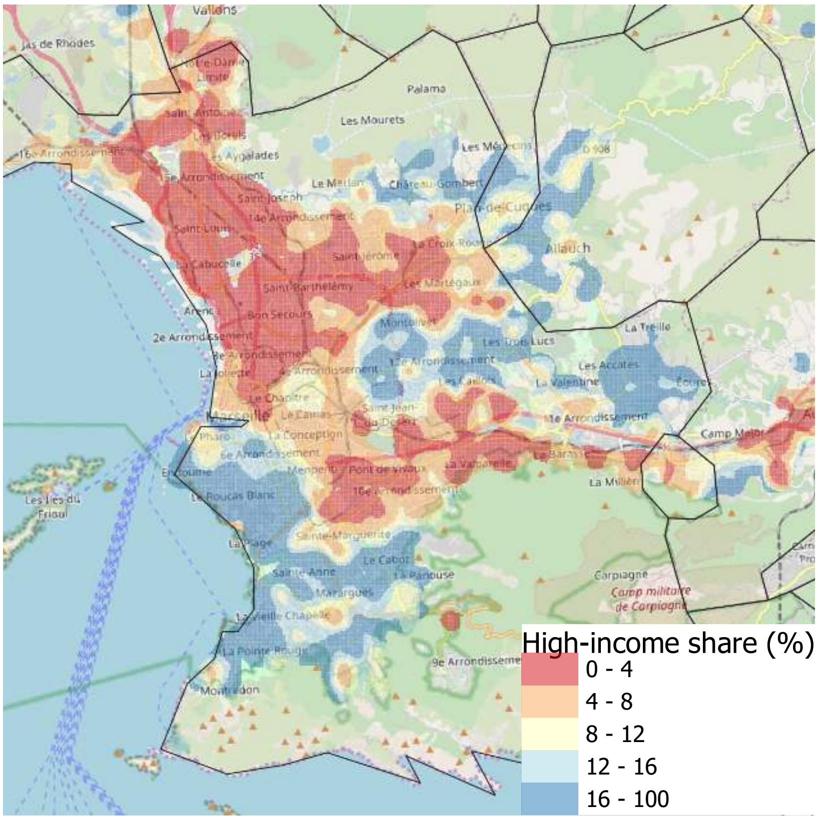


(b) High-income population: p_i^{D9}

Figure 3 – Residential segregation in Lyon



(a) Low-income population: (p_i^{D1})



(b) High-income population: (p_i^{D9})

Figure 4 – Residential segregation in Marseille

tive.[9] Cities are defined using urban cluster concept.[10] In a given urban cluster, both city centers and suburban areas are included. A direct comparison of areas using cell-level median income would be reductive, even with our fine spatial granularity (500 meters cells). For instance, within-cell income variance represents 80% of total income variance in Paris when considering 500x500 meters cells. In that case, median income spatial distribution will hide potentially strong local inequalities. It is thus more interesting to focus on the concentration of some income groups. We will focus on income based segregation for two groups: high- and low-income households. Rather than using a multimodal index, we study those two extremes of income distribution separately and build dissimilarity indices by decomposing population into low-income vs others and high-income vs others. High- and low-income groups are defined with respect to city-level deciles: households belonging to the high income group are people belonging to the tenth decile (income greater than P90) while low-income people are those belonging to the first decile (income lower than P10). Formally, for an empirical income distribution function F_n, we define city-level income thresholds as quantiles, i.e. μ^{D1} as $F_n(\mu^{D1}) = 0.1$ and μ^{D9} as $1 - F_n(\mu^{D9}) = 0.1$. At the spatial unit level, described in Section 4, we compute the share of people belonging to an income group. This is an estimator for the probability of having low-income (or high-income) people in the neighborhood. In other words, for a spatial unit i, we compute two quantities

$$p_c^{D1} = \mathbb{P}(y_x < \mu^{D1}) = \mathbb{E}(\mathbf{1}_{\{y_x < \mu^{D1}\}}) = \frac{1}{n_c}\sum_{x=1}^{n_c} \mathbf{1}_{\{y_x < \mu^{D1}\}} \tag{6}$$

$$p_c^{D9} = \mathbb{P}(y_x > \mu^{D9}) = \mathbb{E}(\mathbf{1}_{\{y_x > \mu^{D9}\}}) = \frac{1}{n_c}\sum_{x=1}^{n_c} \mathbf{1}_{\{y_x > \mu^{D9}\}} \tag{7}$$

where n_c is the number of people living inside the spatial unit as measured from tax files and the x index refers to individual level. A spatially uniform income distribution would lead to $p_c^{D1} = p_c^{D9} = 0.1$ in any spatial unit c. A value greater (resp. lower) than this benchmark means an over-representation (resp. under-representation) of the income group in the spatial unit.

3.2 Mobile phone data

We use a pseudonymized call details record (CDR) dataset that gives access to the activity of more than 18 millions phone users during the September 2007 period.[11] As shown in Table 1, the CDR contain information on time, location (cell tower used), as well as caller and call receiver ids. The nature of the interaction (voice or text message) is also available. The CDR report antenna-level position letting the exact position of phone user unknown. For segregation studies, two dimensions of information recorded in CDRs are interesting: interpersonal interactions and geographical positions. We focus on the spatial dimension. Interactions might be used for a future study on social segregation.

The dataset is exhaustive on mobile phone activities of Orange customers in Metropolitan France for 30 days in September 2007. We do not have additional information on phone users characteristics. In particular, the residence location of an individual is unknown. Since residence is key to combine phone data with spatially aggregated tax data, we use a phone user's track to estimate her residence neighborhood. The main bias that can arise from this simulation approach is known as the ecological

[9] Geocoded tax information is available for years before 2014. However, datasets are incomplete up to 2014. The discrepancy between phone data (2007) and tax (2014) collection year would only be problematic if the relative ranking of neighborhoods inside cities had dramatically changed. This is unlikely to happen since those changes generally take more than seven years to happen.

[10] A urban cluster is an unit offering at least 10,000 jobs and which is not situated in the suburban rim of another urban cluster.

[11] Call Details Records (CDR) are billing data. They are automatically collected for calls and text messages, sent or received by the user together with the interlocutors' identifiers and the identifier (position) of the cell that dispatched the communication. See Table 1 for an example.

Timestamp	Caller	Callee	Event	Duration(sec.)	Area_id	Cell_id
2007/10/01 12:09:55	HJT22RR1	R482GH001	VO	365	1548	530012
2007/10/01 12:10:32	TR001BB	25GG2477	SMS	12	32110	255337

Table 1 – Call Details Record: example

	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 2 – Septembre 2007 Call Details Record: summary statistics

fallacy, which is the incorrect imputation of income at the individual level from aggregated data. We propose in Section 4 alternative combinations between sources to measure the sensitivity of segregation indices to the imputation methodology. Both datasets are anonymous and the combination we perform with tax data is always based on spatial aggregates, not on individual income data.

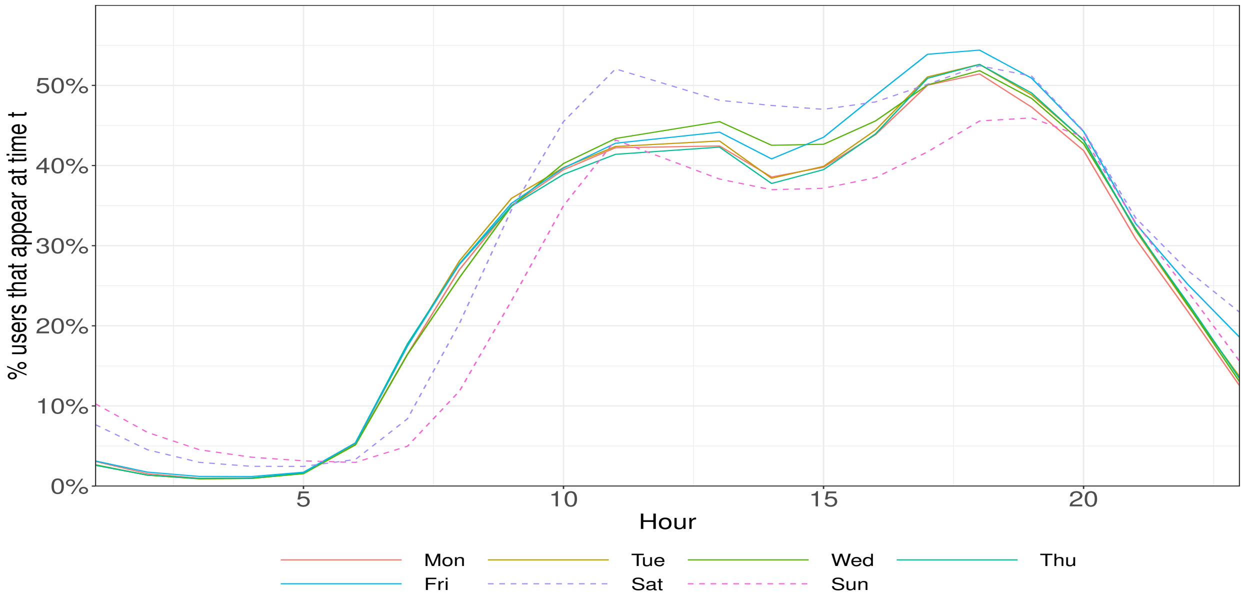
This database presents typical features of CDR datasets that are challenging for their automated analysis. One of the greatest challenges is the temporal scarcity at the individual level (Table 2).[12] An average individual uses his phone approximately 4 times a day. However, phone users are repetitively measured along the month since the average user appears every two days. For both frequencies, we are faced with strong phone usages heterogeneity. At the aggregate level, Figure 5 shows how many of our 18.5 millions French phone users appear at least once for each hour by day. Because individual recording depends of a phone user activity, the number of observed phone users drops significantly during the night. We reduce the week to 48 time windows to ensure that a minimum number of users appear at all times. 24 windows represent weekdays hours and 24 windows represent weekend hours.

Antennas are mostly concentrated in urban areas and along some major communication channels (railways, highways, etc.) as shown in Figure 10. For this reason, except when estimating phone users home, we will only consider dense urban areas. Appendix A proposes a methodology to probabilize phone users presence with a more homogeneous spatial granularity. Since phone users income simulation is performed from spatial aggregates, it is important to use a spatial partition that is less heterogeneous than the voronoi cells presented in Appendix A.

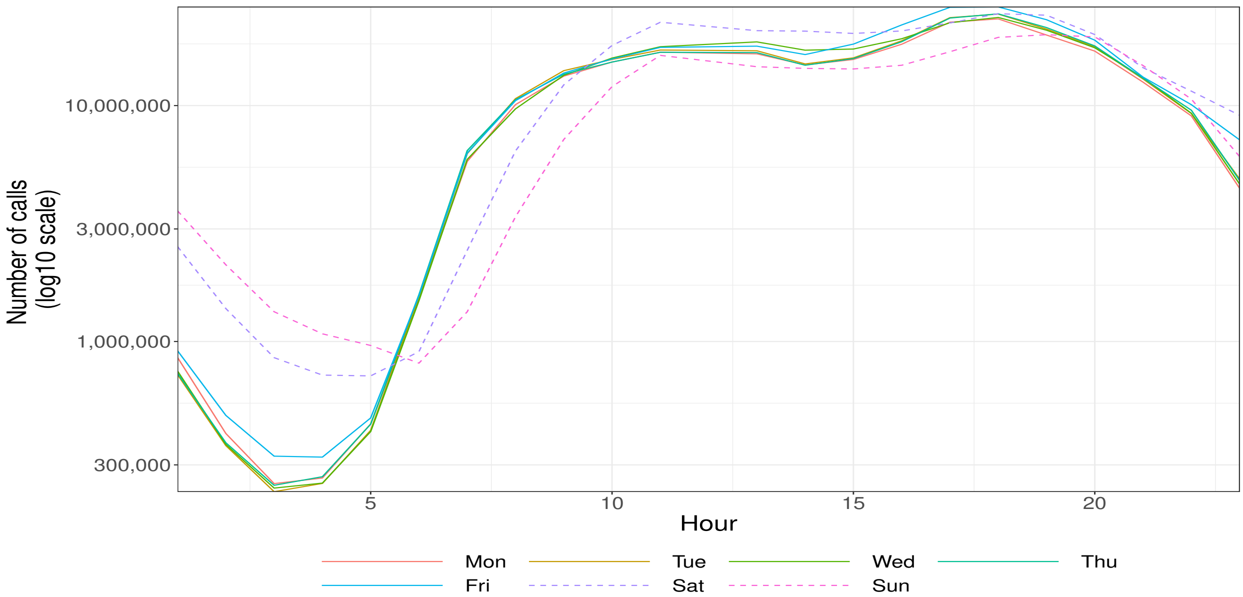
4 Methodology

Based on 500x500 meter cells, segregation is measured at the city level with a Theil index. We propose an infra-day picture of segregation where mobility can bring together people otherwise separated in residential space. Time windows aggregate several days together to construct a typical hour. Thus, within a given time window, an individual can be measured with uncertainty in several places. That requires to generalize segregation indices in a probabilistic framework (see Appendix B.3 for other

[12] This typical feature of CDR is exacerbated by the fact that unlimited packages were not frequent in 2007.



(a) Number of unique users



(b) Number of calls

Figure 5 – Number of calls and users by hour

segregation indices).[13]

Phone users economic status is simulated in two steps. They are presented with more details in Appendix A. This procedure is a Monte-Carlo simulation of phone users status with mixture distributions for cell home (ν_x^{home} in eq. 19) and income (income distribution moments from eq. 6).

Home estimation First, for a given phone user x, we estimate the distribution of her home cell ν_x^{home} (eq. 19). Because our knowledge on residential space density is a useful information to evaluate the likelihood that some cells host people, the home estimation method is based on Bayesian estimation (see eq. 18). Residential density derived from tax data is used as a prior that upweights densely populated cells. As shown by Figure 13, using a prior makes the population distribution more consistent with tax data. Once the distribution of x's home cells ν_x^{home} is estimated, we draw home cell (denoted c_x^{home}) from ν_x^{home} distribution. Starting from there, we restrict the sample to people whose simulated home cell belongs to Paris, Lyon or Marseille.

Income group simulation From tax data, in each cell, we know the complete income distribution. We extract, for each cell, the frequency of people belonging to first and last income deciles (eq. 6). Given c_x^{home}, we use Bernoulli simulations (with weights $p^{D1}_{c_x^{home}}$ and $p^{D9}_{c_x^{home}}$) to simulate x phone user economic status. This method gives a value of 0/1 for belonging, for instance, to the high-income group. Simulations for low-income and high-income indices are run independently. For instance, assume someone is estimated to live in a cell where $p^{D9}_{c_x^{home}} = 0.4$. To simulate his group, a Bernoulli distribution with a 40% probability of success will be drawn.

Segregation indices at city level In general, Theil index (2) measures the deviation of a cell entropy from the city level entropy. The generalized version we propose, using cell presence probability, is constructed by analogy with the standard definition. Instead of having observations, we have probability at cell level c for time t. We denote $\mathbb{P}_{x,t}(c)$ the probability that phone user x has to be in cell c at time t. Let's denote $\mathcal{X}$ the subset of users whose position is observed at least once during time window t. Assume we focus on income group g. The frequencies used in (2) can be rewritten

$$p_{c,t} = \frac{\sum_{x\in\mathcal{X}} \mathbb{P}_{x,t}(c)\mathbf{1}_{x\in g}}{\sum_{x\in\mathcal{X}} \mathbb{P}_{x,t}(c)} \tag{8a}$$

$$p_t^{\text{city}} = \frac{\sum_{x\in\mathcal{X}} \mathbf{1}_{x\in g}}{\text{card}(\mathcal{X})} \tag{8b}$$

$p_{c,t}$ denotes, for instance, the average proportion of phone users in a cell around the Eiffel tower at time t that belong to income group g. p_t^{city} denotes the same proportion at city-level. In eq. (8a), we consider the share of people observed in cell c at time t that comes from income group g. In eq. (8b), we define the share of people observed at time t that comes from income group g. The generalization of (2) writes as

$$H_t = \sum_{c=1}^{C} \frac{\sum_{x\in\mathcal{X}} \mathbb{P}_{x,t}(c)}{\text{card}(\mathcal{X})} \frac{E(p_{c,t}) - E(p_t^{\text{city}})}{E(p_t^{\text{city}})} \tag{9}$$

[13]Eq. (2) can be seen as an expectation. In that case, our probabilistic framework is just a case where observations are not sampled from a degenerated probability distribution (observation inside unit i in eq. (2) can be seen as presence with probability 1) but where presence location follows a non-degenerated probability measure. In other words, the index we propose is a Theil index with a more complicated measure.

where E is the entropy transformation (1). The equivalent generalization of other segregation indices is presented in Appendix B.3. Indices are computed for the 48 time windows defined previously. It is implicitly assumed that the sample of phone users belonging to $\mathcal{X}$ is representative of the population as a whole. In other words, it is assumed that there are no income-related characteristics that could cause users to appear or disappear over time.

As it is emphasized in Appendix A, to account for the downward bias with respect to tax data, it is necessary to assume people not observed during nighttime are at home. Appendix B proposes robustness checks in voluntarily maximizing (or minimizing) the likelihood that individuals belong to the income group of interest. The conclusion of those tests is that the bias that could arise from ecological fallacy should be of limited importance. These alternative methods show that the methodology adopted is the one that best reproduces residential segregation. Regardless of the method, a similar segregation dynamic is observed, with a significant decrease in indices during the day.

A note on the computational challenge

The call details record volume is 1.5TB. This represents more than 3 billions observations, geocoded at antenna level. We transform the individual traces at antenna level data into individual traces at 500 meters level. This operation is performed for the whole French sample (18.5 million individuals). The probabilization of an individual presence, described by Appendix B.1, creates a more voluminous object than her initial antenna level presence. Only after estimating residence, we restrict ourselves to the sample of people that lives within Paris, Lyon and Marseille. Estimating residence requires to draw a home cell for 18.5 million phone users using a rejection method based on observed nighttime probabilities. Phone data are processed within Orange facilities using `pySpark` framework with an Hadoop Distributed File System (HDFS) infrastructure. The typical `Spark` configuration we use requires 80 slave executors.

The gravity model applied in Section 6 is also very computationally demanding. The model is estimated on origin-destination flows using 500 meter cells. In Paris, the whole dataset represents more than 80 million observations. The non-linear optimization required to find the maximum likelihood estimator has, as far as we know, never been applied on such large scale data. We chose to use a random 5% sample of the whole dataset (see Section 6) to speed up computations and reduce memory needs. Results are very similar with those on the whole dataset (See Appendix C). We used `R` software to find the maximum likelihood estimator. Because the existing implementations were not adapted to large scale data, even with the 5% sample, we optimized some steps with C++. Some investigations are still needed to speed up computations.

5 Results

5.1 Validation with tax data

Nighttime segregation derived from phone data should be comparable to levels based on residential data. Indices derived from positional data by Athey et al. (2019) are generally lower than those based on residential data. Our results are similar on that aspect (Table 3). We estimate the levels of segregation for the high-income earners better than for the low-income group. For the latter, since the indices can be low in residential data, the error percentages are larger.

To understand better the mechanisms at stake, we propose to explore dynamics by considering

	Paris		Lyon		Marseille	
	Low-income	High-income	Low-income	High-income	Low-income	High-income
			Weekdays			
Max difference with tax data	0.017	0.023	0.029	0.018	0.016	0.003
Relative difference with tax data (%)	21.59	11.99	30.96	12.86	12.66	1.73
			Weekend			
Max difference with tax data	0.018	0.024	0.03	0.019	0.017	0.004
Relative difference with tax data (%)	22.07	12.59	31.79	13.81	13.42	2.47
Theil index in tax data	0.08	0.191	0.093	0.14	0.124	0.146
Max. Theil index in phone data	0.062	0.168	0.064	0.122	0.108	0.144

Max difference with tax data: $H^{\text{tax}} - H^{\text{max}}(t)$
Relative difference with tax data (%): $1 - H^{\text{tax}}/H^{\text{max}}(t)$

Table 3 – Divergence between maximum segregation in mobile phone and tax data

	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
			Weekend			
Max amplitude	0.05	0.12	0.05	0.1	0.08	0.11
Relative amplitude (%)	75.3	69.99	86.18	81.95	79.05	77.6
Within night (19h-9h) relative amplitude (%)	50.65	45.59	55.91	50.89	52.5	50.02

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

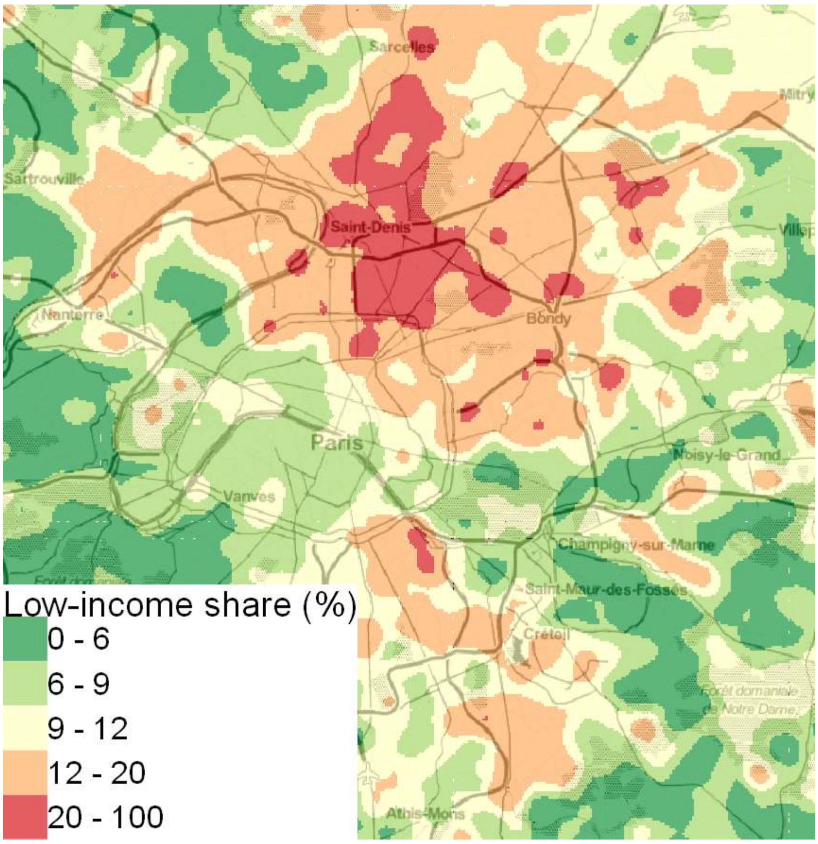
Table 4 – Maximum variations in segregation along the day

neighborhoods evolution inside city. Neighborhoods are often characterized using residential income data (see for instance Floch, 2017). It is interesting to use both spatial and temporal dimensions in mobile phone data to characterize places with more dimensions.

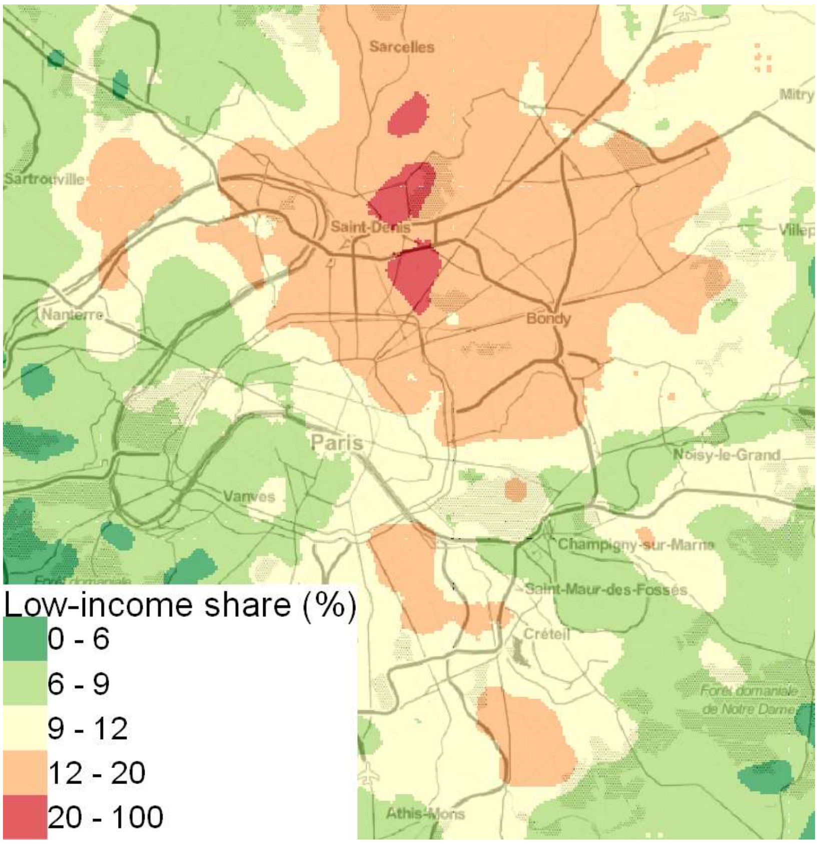
5.2 Segregation dynamics

An illustration of the evolution of low-income concentration during the day in Paris is presented on Figures 6. In both Figures, one can see that low-income people tend to be more concentrated in the north-east during nighttime. This is consistent with tax data (Figure 2). Poorer households are still more concentrated in the north-east and south-east, the two nighttime hotspots. However, they also spread in the city-center and the west of the urban cluster, a region from which they are excluded in residential space (Figure 2).

The infraday evolution of segregation is reported on Figure 7. Indices are bootstrapped (50 iterations). For both low- and high-income groups, segregation reaches its acme at night and goes down to a relatively stable level during the day. Segregation does not seem to be different during weekends, both in terms of level and within-day variation. The effect of mobility on spatial income distribution is not negligible: all segregation indices lead to the conclusion than city are less segregated during daytime than nighttime. Magnitudes are more important with a Theil index (around 75% decrease) than with a dissimilarity index (around 50%). Income groups distribution is less uneven during the day. However, because only spatial segregation is considered, nothing can be said about interactions that occur when

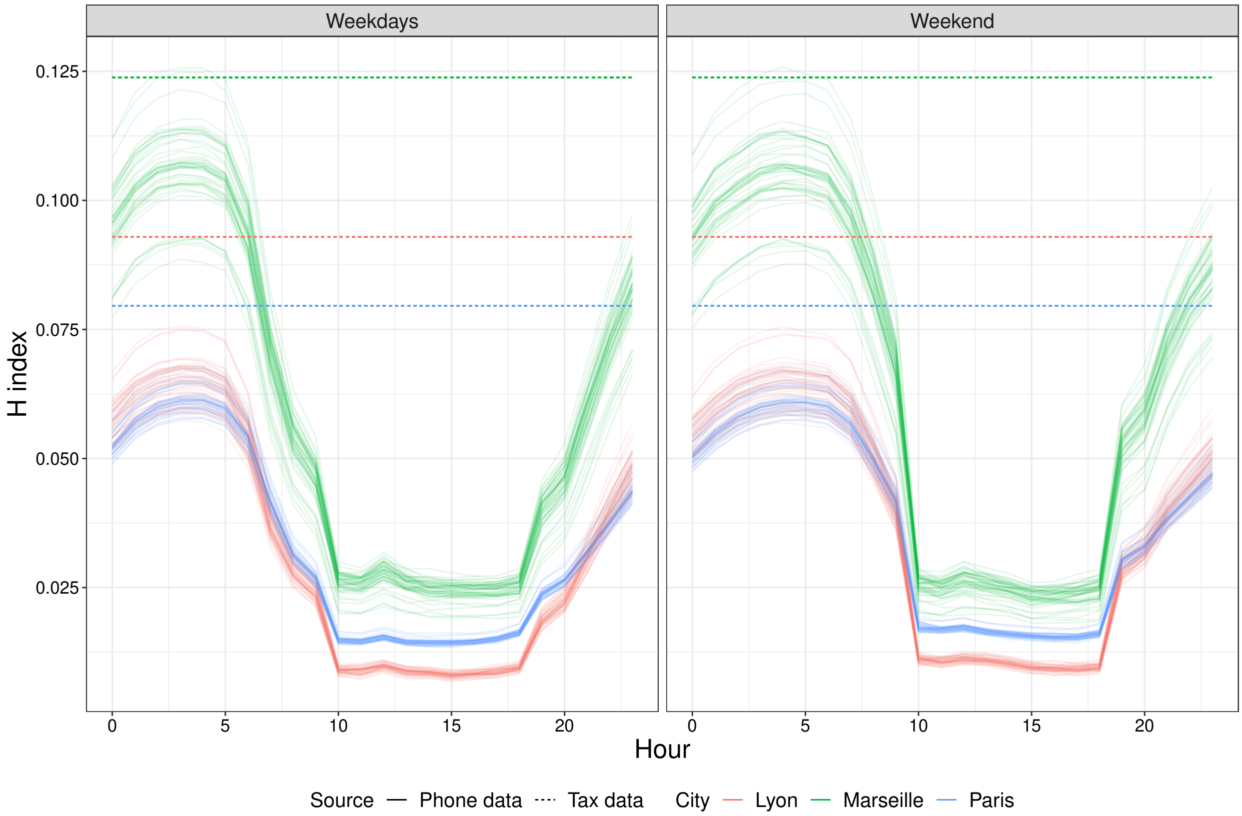


(a) Low-income distribution at 6am

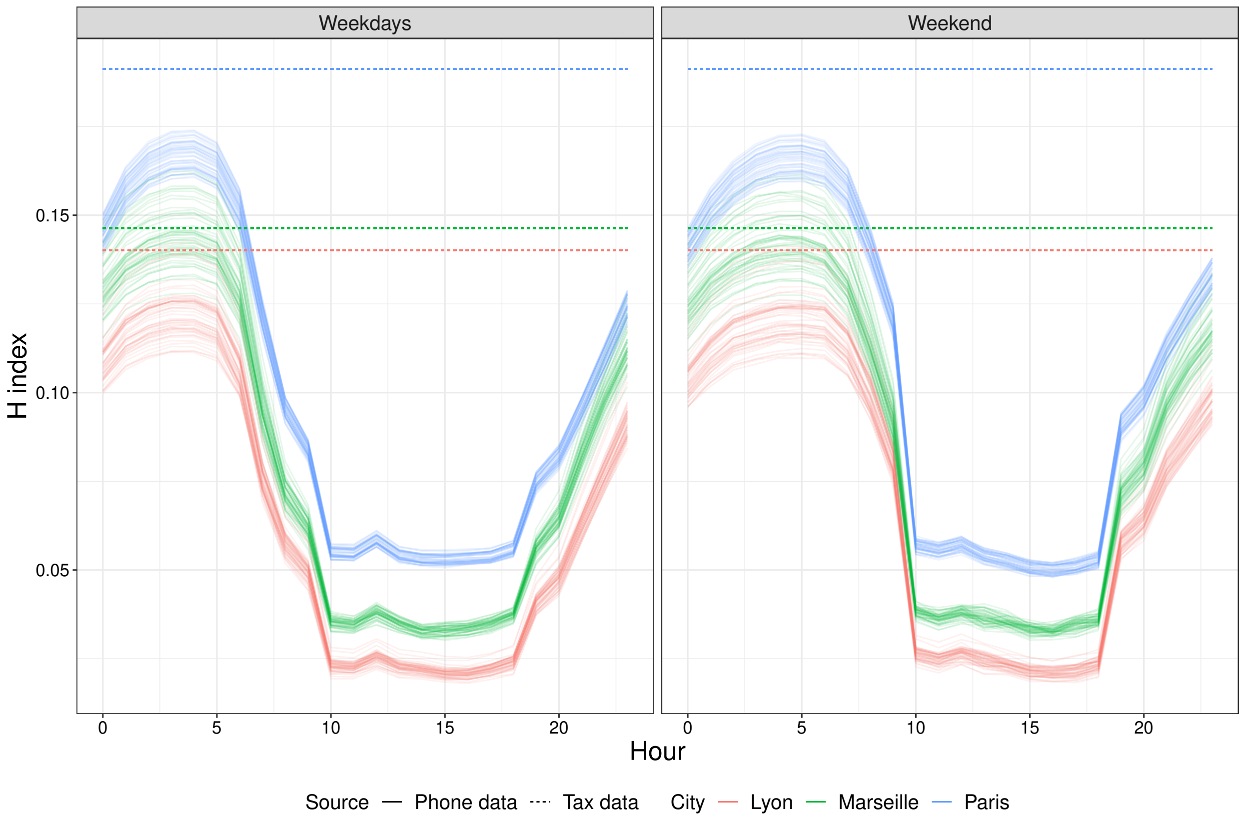


(b) Low-income distribution at 4pm

Figure 6 – Low-income concentration at 6am and 4pm during weekdays in Paris urban cluster



(a) Low-income Theil index



(b) High-income Theil index

Figure 7 – Segregation evolution during the day in the three main French cities

people cross the same public space (here defined at 500 meters cells level). Daytime level for Theil H index is around 0.05 for high-income people in Paris, two times higher than low-income index.

When using the same segregation index than Davis et al. (2019), i.e. the dissimilarity index, we find a similar order of magnitude for the decrease of segregation index. This magnitude is not far from the one derived by Athey et al. (2019) using isolation index. The magnitudes we get are greater than the 15-30% daily variation measured by Le Roux, Vallée, and Commenges (2017) from Paris transport survey data.

Variation is significant within the night. Between the acme (around 4-5am) and its lowest point (generally 7-8am), the dissimilarity index has known a 30 to 40% decrease. At 7am, segregation is close from its daytime level. Conclusions regarding segregation evolution are not affected by the income estimation method as pinpointed in Appendix B. In all simulation schemes, daily segregation is stabilized after 10 am (commuting time). Segregation takes more time to return to its maximum level than it takes to fall. Athey et al. (2019) results are similar. A slight difference in mornings is visible between weekend and weekdays. Flows that bring people out of their living environment take place later on the weekend compared to weekdays.

From Figure 7, we can see that segregation falls relatively more, for both high- and low-income populations, in Lyon than in Paris and Marseille. For the low-income group, the hierarchy between the cities that can be observed during nighttime changes during daytime. Segregation starts to be lower in Lyon than in Paris around 6am and up to 9pm. For the high-income group, there is no change in the hierarchy of segregated cities during daytime. Paris stays the city where segregation at the top is the highest.

5.3 Characterizing places by their population pattern

Clustering consists in grouping together places with similar characteristics. We propose to identify a few types of homogeneous spaces based on the evolution in the concentration of low- and high-income people inside the cell. The goal is to find some places that share common patterns in terms of socioeconomic composition. Neighborhoods could also be characterized according to the facilities available and phone data could be used to study the differentiated use of urban space between social groups. Nevertheless, despite the precision of our grid, 500 meters uncertainty regarding people presence remains too important for the conclusions of such an approach to be reliable. Therefore, we favour the use of population dynamics at the cell level, controlling for the socioeconomic environment as well as equipment data to characterize clusters.

Our variable of interest is the sequence of low-income and high-income concentrations in our cells. In other words, we gather places using $(p_{it}^{\mathrm{D1}})_{i,0\leq t\leq 47}$ and $(p_{it}^{\mathrm{D9}})_{i,0\leq t\leq 47}$ series. We resort to k-means classification algorithm, an unsupervised machine learning technique, to group cells with similar patterns. Our goal is to find a partition, among all possible K-partitions, that minimizes the Euclidean distance between points and cluster center. We partition cities in $K = 4$ types of places. We characterize them from the dynamics of segregation. The choice of 4 clusters is motivated by the trade-off between variety and parsimony. A larger number of clusters results in only a moderate decrease in the sum of squared errors. Having four clusters still allows to deduce interesting intuitions about the spatial partition derived.

The mean profile of the clusters as well as the map of the partition are represented in Figure 8. Cluster 1 correspond to high-income neighborhoods. During nighttime, high-income people represent around 40% of the cell population. On the contrary, the proportion of low-income people during the night is very low (less than 5% of the cell). The day brings a bit of diversity to these spaces, but they remain unequal. Conversely, the north-east of Paris concentrates the majority of the cells belonging to cluster 2, which is the one of low-income cells. High income people remain less present during daytime in these places than would be expected with a spatially uniform distribution. It is natural to find that the position of clusters 1 and 2 is consistent with the heatmap derived from tax data. For both income groups, clusters 1 and 2 correspond to places where low-income or high-income are 2 to 4 times more concentrated than expected by a uniform prior. Cluster 3 represents cells with a dynamics similar to cluster 1. However, these cells concentrate more middle class people since the proportion of low-income is similar to cluster 1 while the concentration of high-income is lower. Except in the north-east where we find some cells belonging to cluster 2, cluster 3 completes cluster 1 to cover the center of Paris. Finally, cluster 4 represents middle-class cells where both low and high-income concentration are close to the uniform distribution prior.

A multinomial logistic model is estimated to evaluate the importance of some environmental variables on the likelihood of belonging to a given cluster. The residential composition of neighborhoods plays a key role in explaining clusters, which is confirmed by Table 5 for Paris. Cluster 4, the one close to the constant uniform pattern, is taken as the reference[14]. Reported estimates represent log of the relative risk ratios between categories. Population concentration in tax data affects the type of cluster a neighborhood belongs to. For instance, the higher high-income concentration in residential data will be, the more likely a neighborhood will belong to the cluster 1. However, surprisingly, the high-income variable is not significant for cluster 3. Regarding low-income concentration, the sign follows the intuition for cluster 1 and 3. However, from Figure 8, we would expect the coefficient to be positive for cluster 2, which is not the case. Cluster 1 covers the neighborhoods where jobs are the most concentrated, especially *La Défense* central business discrict or Paris center. This is natural to find a positive and significant relationship between the number of jobs and the likelihood to belong to cluster 1. The higher the number of shops around a cell will be, the less likely it will belong to cluster 2 or 3. A high number of shops around a cell could reflect two types of spaces: city center places with many shops (most likely in cluster 1) or commercial areas in the suburbs (most likely in cluster 4). The contribution of each cluster to the overall segregation is presented in Figure 9[15].

Given the data available, we cannot go further to link cell-level social composition with some of the amenities of the cell, as Athey et al. (2019) propose. More recent data, giving access to a more complete information on equipment and the possibility to use a lower-level spatial granularity, might be of crucial importance for public policies. For instance, that could help to determine and the amenities that favor social mixing.

Phone data (Figure 8) as well as tax data (Figures 2, 4 and 3) exhibit a gravity structure. Even during daytime, we can still identify places where more people from a given income group are concentrated: they are generally not far from places where income group is concentrated in tax data. Hotspots in tax data are still hotspots in phone data even if daytime flows reduce inequalities between districts. This suggests the existence of spatial frictions that limit the homogenization of cities and explains why segregation does not disappear completely during the day.

[14]This means the general form of the model is, for $k = 1, 2, 3$,

$$\mathbb{P}(\text{cluster} = k) = \frac{\exp(\beta_k X)}{1 + \sum_{k=1}^{K-1} \exp(\beta_k X)}$$

β coefficients are directly reported on Table 5. When exponentiated, these values are equal to relative risk ratios.

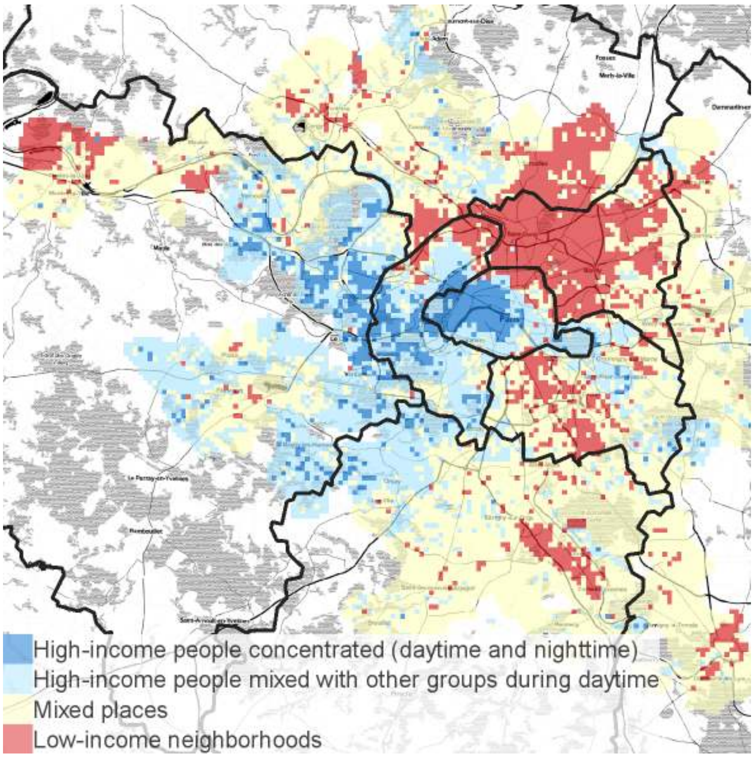
[15]Cells that were not assigned to some cluster because of missing values have been associated into an additional cluster. As one can see in Figure 9, these cells have a marginal influence on overall segregation.

Table 5 – Explaining clusters in Paris with multinomial logistic regression models

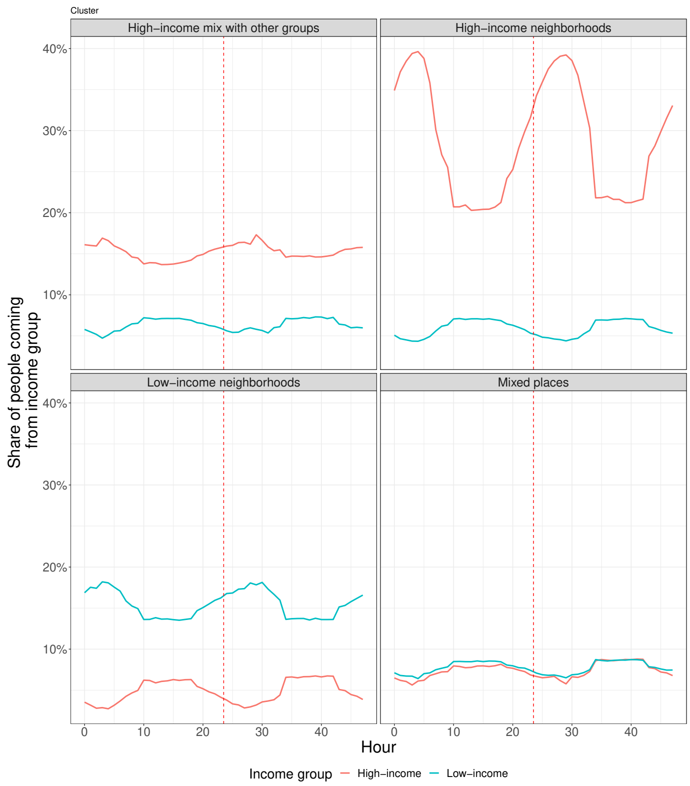
	Dependent variable:			
	High-income neighborhood (Cluster 1)	Low-income neighborhood (Cluster 2)	High-income with middle-class (Cluster 3)	(Reference)
(Intercept)	-2.637^{***}	-0.382	0.368	(Ref)
	(0.457)	(0.388)	(0.752)	
Low-income concentration in tax data (%) (1)	-0.100^{***}	-0.083^{***}	-0.055^{**}	(Ref)
	(0.026)	(0.024)	(0.026)	
High-income concentration in tax data (%) (1)	0.026^{***}	0.059^{***}	0.010	(Ref)
	(0.009)	(0.008)	(0.013)	
Employment in cell (log number of people) (2)	0.229^{***}	0.041	-0.009	(Ref)
	(0.072)	(0.056)	(0.069)	
Number of shops in cell (log) (3)	-0.150	-0.015	-0.240	(Ref)
	(0.116)	(0.113)	(0.153)	
Number of shops in adjacent cell (log) (3)	0.041	-0.364^{***}	-0.451^{***}	(Ref)
	(0.132)	(0.111)	(0.139)	
Number of schools in cell (4)	-0.092	-0.087	0.020	(Ref)
	(0.059)	(0.077)	(0.074)	
Number of schools in adjacent cell (4)	-0.074^{***}	-0.182^{***}	-0.018	(Ref)
	(0.024)	(0.037)	(0.030)	
Number of sport and entertainment facilities in cell (4)	-0.033	-0.104	0.009	(Ref)
	(0.122)	(0.247)	(0.196)	
Number of sport and entertainment facilities in adjacent cells (4)	-0.276^{**}	-0.178	-0.182	(Ref)
	(0.133)	(0.156)	(0.247)	
Number of subway and train lines in cell (5)	-0.174	-0.053	-0.849	(Ref)
	(0.315)	(0.364)	(0.747)	
Number of subway and train lines in adjacent cell (5)	-0.155	-0.133	-1.056	(Ref)
	(0.251)	(0.315)	(0.701)	
Observations		2,254		
Log likelihood (by obs.)		-0.429		
Bayesian information criterion		2,213		

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

(1) Low and high-income concentration (p^{D1} and p^{D9}) defined by eq. (6)
(2) Data: CLAP 2014 (*Conaissances Locales de l'Appareil Productif*)
(3) Data: Corporate data 2014 (*Cotisation foncière des entreprises* combined with *Sirene* data)
(4) Data: Equipement data 2007 (*Base publique des Equipements*)
(5) Data: Public transport lines dataset



(a) Map



(b) Clusters mean profile

Figure 8 – Places classification in Paris urban cluster

6 A gravity model for infra-day population flows

6.1 Specification

Gravity models are a common approach to estimate flows between two regions: see, for instance, Anderson and Van Wincoop (2003) for a gravity model with trade flows. The phone data literature, e.g. Krings et al. (2009), refers generally to Zipf's law to model these flows. According to that law, flows decrease at rate $1/d^2$ where d is the distance between origin and destination. This power 2 rate is generally taken for granted in the mobile phone as well as in many urban geography papers. However, this gold standard of a log2 relationship between distance and interactions poorly fits short distances (under 10km in Krings et al., 2009) which are relevant in urban areas. A correct estimate for travel costs at small geographic scale, i.e. urban level, would contribute to grasp better human dynamics in places where population is denser. In a city where spatial frictions are strong, people from distant neighborhoods will have limited interactions. The higher spatial frictions will be, the closer experienced segregation and residential segregation might be.

We propose to estimate travel costs to evaluate spatial frictions. In particular, we quantify the difference in transportation costs for low- and high-income people. We address a second limitation in the gravity models proposed in the mobile phone literature, namely the lack of control variables, and discuss the specification of an econometric model to measure flows determinants. The estimation of travel costs might be biased due to selection issues that need to be accounted for.

We address several challenges regarding inference from gravity models with estimation procedures that have first been developed by the international trade literature. Building a gravity model at 500m granularity scale with a population sample of nearly one third of a large city is a very ambitious task. The zero-inflated models we use make the estimation more complicated because they are computationally very demanding.[16] However, this complexity is necessary to derive robust estimates of travel costs, especially when looking at heterogeneity across several dimensions including income.

Our observed flows are probabilities to see people living in cell i and belonging to income group g go to cell j. Forgetting about the income group dimension, we define $p_{i\to j}$ as the sum of probabilities of people being in cell c_j between 2 and 6pm and living in cell c_i. Think for instance of people living in Saint-Denis and going to the Eiffel Tower in the afternoon. $p_{i\to j}$ parameters form therefore an origin-destination matrix where the origin point is the living neighborhood.

In migration studies, the framework of the random utility model is often used to derive a gravity model that takes the following form (Beine, Bertoli, and Fernández-Huertas Moraga, 2016):

$$p_{i\to j} = a\frac{M_i^{\beta_1}M_j^{\beta_2}}{D_{ij}^{\beta_3}} \tag{10}$$

where M_i and M_j are origin and destination characteristics (reminiscent of the mass variables in Newton equation). The coefficients β_1, β_2 and β_3 determine the effect of each variable on the strength of the link between cells i and j. β_3 is often denoted as the distance-decay parameter. A first approach consists in deriving the empirical counterpart of equation (10) after log-linearization:

$$\text{(OLS)} \qquad \log(p_{i\to j}) = \beta_0 + \beta_1\log(M_i) + \beta_2\log(M_j) - \beta_3\log(D_{ij}) + \epsilon_{ij} \tag{11}$$

[16] Routines available in `R` to fit negative binomial models (`MASS` and `pscl` packages) require too much RAM to be implementable. We thus created an `R` package based on `C++` to reduce memory requirements and speed up computations. That package will be published on `Github`. However, even with that implementation, fitting a maximum likelihood model with large scale data is still challenging.

where ϵ_{ij} is an error term. Assuming $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, the model can be estimated by OLS. However, this approach has several drawbacks. In particular, it does not account for heteroscedasticity. This may matter since the variance of flows might be related to the characteristics of origin and destination regions, especially their distance.

Santos Silva and Tenreyro (2006) and Santos Silva and Tenreyro (2011) advocate in favor of count models to estimate gravity equations because they account for heteroscedasticity. Santos Silva and Tenreyro (2006) and Head and Mayer (2014) also discuss the robustness of count data models to measurement errors. The two most common count data regression techniques used in this framework are based on Poisson and negative binomial distributions.[17] In both cases, the model writes

$$\lambda_{ij} = \mathbb{E}_{f,\theta}(p_{ij}|M_i, M_j, D_{ij}) = \exp\left(\beta_0 + \beta_1 M_i + \beta_2 M_j - \beta_3 D_{ij}\right) \tag{12a}$$

$$\text{with} \quad \begin{array}{ll} \text{(Poisson)} & \mathbb{P}(p_{ij} = p) = \frac{\exp(-\lambda_{ij})\lambda_{ij}^p}{p!} \\ \text{(NB)} & \mathbb{P}(p_{ij} = p) = \frac{\Gamma(p+1/\alpha)}{\Gamma(p+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\lambda_i}\right)^{1/\alpha} \left(\frac{\alpha\lambda_{ij}}{1+\alpha\lambda_{ij}}\right)^p \end{array} \tag{12b}$$

α is known to be a dispersion parameter in the negative binomial model. The two distributions are based on the same conditional mean model (12). Conditional expectation is underscored by f and θ to highlight the fact that the model prediction depends on distributional assumptions (a density probability function f with relevant parameters θ that will either belong to Poisson or negative binomial family). Assumption regarding family f will lead to favor Poisson or negative binomial model (12a). This will have implications regarding the conditional variance $\mathbb{V}_{f,\theta}(p_{ij}|M_i, M_j, D_{ij})$. Under a Poisson model, $\mathbb{V}_{f,\theta}(p_{ij}|M_i, M_j, D_{ij}) = \mathbb{E}_{f,\theta}(p_{ij}|M_i, M_j, D_{ij}) = \lambda_i$. With a negative binomial model, $\mathbb{V}(Y_i|X_i) = \lambda_i(1 + \alpha\lambda_{ij})$. The conditional variance depends on the dispersion parameter α. Thus, a negative binomial model will be used to model overdispersion or underdispersion.[18]

A more critical problem arises when using a log-linearized gravity model. This approach implies to drop out flows where $p_{i\to j} = 0$, i.e. situations where no flows are observed. Non-linear transformations of eq. (11) of the dependent variable should not be used to solve this problem (Santos Silva and Tenreyro, 2006). This sample selection problem is reminding of Heckman (1979): OLS ignore information regarding flows determinants and are likely to convey some estimation bias. Santos Silva and Tenreyro (2006) and Santos Silva and Tenreyro (2011) argue that count data models are well suited to handle zero-valued variables. For instance, with Poisson distribution, the likelihood of having a zero value is $\exp(-\lambda_{ij}) = \exp(-\beta X_{ij})$. However, fitting such models on a large frequency of zeros might bring some parameters values to be measured with bias, especially when zeros are not random with respect to observables. A solution to solve that problem is to use the standard Heckman (1979) correction for sample selection. In the standard framework, covariates that determine selection are introduced using the inverse Mills ratio. Helpman, Melitz, and Rubinstein (2008) address the biases generated by the unobserved country pairs flows with that approach. Zero-inflated models are the equivalent Heckman (1979) models in count data framework. They consist in extending the count data process with a selection model that determines the likelihood of observing a zero (Lambert, 1992).

Zero-flows are very common given the fine spatial granularity we adopt. For instance, for low-income in Paris, we observe around 4.5 millions cell level flows. However, the total number of flows that could arise is far more important (around 85 millions)[19]. Given the sparsity of the selected flows

[17] Poisson and negative binomial are discrete distributions. They are, originally, designed to handle counts. However, they can be used to model continuous valued dependent variables

[18] By overdispersion (resp. underdispersion), we mean situations such that the variance of the data is far too big (resp. small) to make the underlying distributional assumption in Poisson model (mean and variance are equal) suitable. See Appendix C for more details.

[19] The complete origin-destination matrix has a dimension of approximately 115 millions cells. However, empty cells cannot be considered as an origin point in phone data since nobody is supposed to live there, whatever the data source is. Thus, we should not consider empty cells has being a potential origin point (but they can be a destination point). This reduces the total number

within all potential flows, zero-inflated count data models seem more appropriate. They offer a generalization of Heckman (1979) procedure in a count data framework. With zero-inflated models, observed data are a mixture of two distributions. With a probability π_{ij}, count is zero. This is the selection process. Selection depends on some characteristics Z. With probability $1-\pi_{ij}$, counts (potentially including some zeros) are generated according to a count data model. Variables that explain the observed outcome are denoted X. The probability density function writes as

$$\mathbb{P}(p_{i\to j}=p)=\pi(Z)\mathbf{1}_{p=0}+(1-\pi(Z))f_{\text{count}}(p|X) \tag{13}$$

with $f_{\text{count}}(p|X)$ the probability density function of the distribution used in the outcome model. The choice of the selection model, which may be a logit or probit model, is justified by theory or by a statistical performance criterion. The choice for the count data distribution depends on the structure of the conditional variance for the outcome model. Even if they are far more computationally demanding, zero-inflated models are fundamental to estimate better spatial frictions.

Among possible specification choices, relying on the BIC criterion, we chose to apply a zero-inflated negative binomial model for all cities and all income groups. This specification is consistent with the data structure that exhibits a large share of zeros with respect to observed flows and a high variance for observed flows. As regards the parametric assumption in the selection process equation, our specification is based on the logit model. The gravity model we estimate takes the following form:

$$\text{(selection)} \qquad \mathbb{P}(p_{i\to j}>0)=1-\pi_{ij}=\frac{\exp(Z_{ij}\gamma)}{1+\exp(Z_{ij}\gamma)} \tag{14}$$

$$\text{(outcome)} \qquad \lambda_i(X_{ij})=\mathbb{E}_{f,\theta}(p_{i\to j}|X_{ij})=\exp(X_{ij}\beta) \tag{15}$$

where γ and β are sets of parameters that need to be estimated. The model is estimated by Maximum Likelihood. The log-likelihood for both Poisson and negative binomial are presented in Appendix C. The results are reported in Tables 6, 7 and 8 while the results related to the other specifications are reported in Appendix C. We assume that the same set of variables explains the selection and the outcome. This means that $Z=X$. Identification relies on the functional form of the likelihood function (see eq. 32 and 33). With a model estimated by Maximum Likelihood, theoretically, this does not create an identification issue under the hypothesis that the functional form is correct (see Appendix C and Papadopoulos and Santos Silva, 2008). However, as for the Heckman sample correction, an exclusion restriction would make the identification more credible. That exclusion restriction would take the form of a variable affecting the likelihood of observing zero-flows but not the strength of flows between the two regions. Finding such valid instrument is not trivial in our context. We thus choose to use the same covariates in selection and outcome equations.

We follow the literature by using population in home and destination cells to control for mass terms (M_i and M_j variables in eq. 10). Because we are interested in daytime flows, we control for employment in both cells. We also introduce controls for the cell composition in tax data because they are likely to affect cell attraction. Distance-related terms are decomposed between city center and suburbs because, given differences in transportation infrastructures or transportation modes between urban and rural areas, flows between cells travel costs are likely to be heterogeneous between city centers and suburbs. To help interpretation, we also introduce, when estimating the model for Paris, the cluster (derived in Section 5.3) the origin and destination cells belong to.

Table 6 – Marseille: gravity model (zero-inflated negative binomial specification)

	Dependent variable:			
	LOW-INCOME		HIGH-INCOME	
	Selection	Outcome	Selection	Outcome
Population in home cell (log)	1.456^{***}	0.728^{***}	1.306^{***}	0.723^{***}
	(0.024)	(0.015)	(0.022)	(0.015)
Population in destination cell (log)	0.092^{***}	-0.016^{***}	0.077^{***}	-0.004
	(0.009)	(0.006)	(0.009)	(0.006)
Employment in home cell (log)	0.243^{***}	0.153^{***}	0.253^{***}	0.142^{***}
	(0.011)	(0.007)	(0.010)	(0.007)
Employment in destination cell (log)	0.391^{***}	0.111^{***}	0.364^{***}	0.084^{***}
	(0.009)	(0.005)	(0.008)	(0.005)
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	
Count distribution	Negative Binomial		Negative Binomial	
Selection distribution	Logit		Logit	
Observations	1,368,224		1,333,298	
Log likelihood (by obs.)	-0.1		-0.1	
Bayesian information criterion	183,804		182,764	

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

Table 7 – Lyon: gravity model (zero-inflated negative binomial specification)

	Dependent variable:			
	LOW-INCOME		HIGH-INCOME	
	Selection	Outcome	Selection	Outcome
Population in home cell (log)	1.352^{***}	0.766^{***}	1.500^{***}	0.896^{***}
	(0.024)	(0.014)	(0.028)	(0.013)
Population in destination cell (log)	0.072^{***}	-0.034^{***}	0.097^{***}	-0.033^{***}
	(0.009)	(0.005)	(0.009)	(0.005)
Employment in home cell (log)	0.157^{***}	0.098^{***}	0.187^{***}	0.081^{***}
	(0.011)	(0.006)	(0.012)	(0.006)
Employment in destination cell (log)	0.271^{***}	0.087^{***}	0.288^{***}	0.097^{***}
	(0.007)	(0.004)	(0.008)	(0.004)
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	
Count distribution	Negative Binomial		Negative Binomial	
Selection distribution	Logit		Logit	
Observations	860,362		849,734	
Log likelihood (by obs.)	-0.1		-0.1	
Bayesian information criterion	218,929		219,668	

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

Table 8 – Paris: gravity model (zero-inflated negative binomial specification)

	Dependent variable:			
	LOW-INCOME		HIGH-INCOME	
	Selection	Outcome	Selection	Outcome
Population in home cell (log)	1.000***	0.708***	0.978***	0.579***
	(0.008)	(0.005)	(0.007)	(0.005)
Population in destination cell (log)	0.043***	−0.079***	0.017***	−0.068***
	(0.003)	(0.002)	(0.003)	(0.002)
Employment in home cell (log)	0.151***	0.061***	0.161***	0.061***
	(0.004)	(0.002)	(0.004)	(0.002)
Employment in destination cell (log)	0.380***	0.099***	0.350***	0.117***
	(0.003)	(0.002)	(0.002)	(0.002)
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 → suburbs)	−2.623***	−1.649***	−2.567***	−1.665***
	(0.009)	(0.004)	(0.008)	(0.004)
Distance (center → suburbs)	−2.398***	−1.377***	−2.299***	−1.371***
	(0.009)	(0.004)	(0.009)	(0.004)
Distance (suburbs → center)	−1.682***	−1.430***	−1.629***	−1.438***
	(0.009)	(0.004)	(0.009)	(0.004)
Distance (center → center)	−1.995***	−1.115***	−1.959***	−1.195***
	(0.014)	(0.007)	(0.015)	(0.007)
α (dispersion)	1.6		1.5	
Count distribution	Negative Binomial		Negative Binomial	
Selection distribution	Logit		Logit	
Observations	8,430,820		8,426,330	
Log likelihood (by obs.)	−0.1		−0.1	
Bayesian information criterion	1,976,421		1,975,188	

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

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

6.2 Results

Estimations for zero-inflated negative binomial models are reported on Tables 6, 7 and 8. They are based on a random 5% sample of the observed as well as non-observed flows. This approach is used to reduce time and memory needed to estimate the model. Appendix C presents results based on the whole sample for Marseille and Lyon.[20] Statistical theory ensures that favouring a sample with exhaustive data has a low cost (confidence intervals decrease at the speed of $\sqrt{n}$) compared to the gains in terms of speed of implementation (algorithmic complexity of regression depends on n).

Both selection and outcome equations are presented. Other specifications are reported in Appendix C. In terms of performance, the difference between the model explaining low-income flows is comparable with the model concerning high-income flows. The model fits better for Marseille than for Paris and Lyon.

Population in home cell plays the expected role both in selection and outcome equations. As regards the selection model, flows are more likely to occur when origin cells are dense. Population in destination cell plays the same role, which is expected. Population should play the same role on the outcome model: the higher population is at the origin or destination, the stronger flows should be. We find such evidence regarding home population in home cell but not in destination cell. Employment always play the expected role: the higher employment is, the more likely flows will be observed and the stronger flows will be. The relationship between flows and income groups concentration in tax data is less easy to interpret. In the three cities, the higher the residential concentration of low-income residents is, the higher the probability of not observing a flow to that neighborhood in the afternoon will be (selection equation). Conversely, neighborhoods with a strong concentration of high-income residents attract more flows. In the outcome equation, the estimates vary between cities. The difference may come from the fact that flows are directed to city centers during the day, which are places where most high-income people live in Paris and Lyon while, in Marseille, low-income people are more concentrated in city center.

The clustering performed, for Paris, in Section 5.3 helps to understand better the structure of urban flows. In the cells belonging to the high-income cluster, we are more likely to observe incoming flows, which tend to be stronger, other things being equal. This is consistent with the central position of these places in Paris' map (Figure 8a). Regarding the other cells where high-income people are over-represented (cluster 3), we also observe stronger flows. However, this time, the effect on selection is going the other way around: cells in this cluster are less likely to observe flows. Cluster 3 is more heterogeneous than cluster 1, which means that we observe purely residential cells in cluster 3 from which population outflows during the day. Finally, neighborhoods belonging to cluster 2 are less likely to experience inflows (selection model) and flows tend to be weaker (outcome model). This result illustrates the difference in attractivity during the day between Paris western high-income areas and north-eastern low-income areas.

Spatial frictions are measured in two ways. The more costly distance is, the less likely flows between two cells happen. In terms of selection, higher coefficients account for spatial frictions. In the outcome equation, which measures the strength of a tie between two places, high estimates account for high travel costs. Hence, expected signs are negative in the selection equation and negative in the outcome model, which we find in all models. In Marseille, the difference in the coefficients of the

of cells to 85 millions.

[20]For Paris, maximum likelihood model takes around 36 hours to converge on the whole sample and requires around 180Gb of RAM. This is not desirable to have such memory and time consuming models. Only the main specification (zero-inflated negative binomial specification) has been estimated on the whole sample of observed and non-observed flows. Point-estimates are almost identical with Table 8. The main difference with the model estimated on a 5% sample is in the size of the confidence interval, which is a second order difference.

outcome equation for low-income and high-income people is small. The difference between these two populations takes place in the selection model. The results suggest that the further away a space is, the less people from high-income neighborhood go there. This might be related to the ability of the well-off to locate their homes closer to their frequent destinations. In Lyon, the difference between income groups is found in the selection model but also in some coefficients of the outcome model, notably those concerning flows between suburbs. In these flows, it is for poor households that distance has the greatest impact on the probability of observing flows. On the other hand, there is almost no difference between the coefficients for flows within the city center. This reflects the dense network of public transport in the centre of Lyon. In Paris, the main difference between high and low-income people is found in the selection model. Distance has a much more negative effect on the probability of observing flows among people from poor neighborhoods than among rich ones. On the other hand, the differences in coefficients between the outcome equations of rich and poor populations are more marginal. In general, income has more influence on the heterogeneity of the coefficients in the selection model than on the outcome model. The conclusions regarding the difference in spatial frictions between suburbs and urban centers are robust to the specification adopted (Appendix C). Zero-inflated models coefficients are comparable between specifications for a given city and income group. The main difference is between traditional models (OLS and count data models) and zero-inflated models. The main specification, based on zero-inflated negative binomial models, improves significantly the quality of the fit.

6.3 Discussion and limitations

The three limitations of using GPS data acknowledged by Athey et al. (2019) also apply to phone data. The first source of measurement error might come from the imputation, at the individual level, of characteristics based on the environment an individual comes from. This error is often denoted as ecological fallacy. Appendix B is devoted to that point. Using exhaustive tax data as ground truth for residential segregation is of particular help to evaluate the potential bias from this approach. Our approach brings nighttime indices relatively close to residential segregation levels, though slightly lower. The second potential source of measurement error stems from the fact that the exact position of phone user is unknown. Constructing a more satisfactory model for position than the one generally used in the mobile phone literature is the object of Appendix A. It takes place in a collective effort at European level to improve population density inference from cell phone data (Ricciato et al., 2015). It is hard to acknowledge the effect of ignoring the exact position of phone users on estimations, especially for the gravity model. However, count models have been shown to be less sensitive to measurement error (Santos Silva and Tenreyro, 2011). Combined with a zero-inflated approach, which takes the selection problem into account, a count model should yield more precise estimates for transportation cost than those derived from OLS. Robustness checks presented in Appendix C show that the conclusions are not sensitive to the functional form assumed in the econometric specification.

The second limit of positional data discussed by Athey et al. (2019) comes from the sparsity of measurement at the individual level within a day. Individual traces in call details records are incomplete. Access to individual signalling data on a large sample would help to better reconstruct individual traces that could be used, for instance, to derive the individual exposure to other socioeconomic groups. However, signalling data volume would make computational challenges even more binding.

Finally, we chose to only focus on spatial segregation. It is well-known that being close to another group does not ensure that interactions take place (Chamboredon and Lemaire, 1970). However, even if people do not always interact, ensuring that populations from diverse horizons have access to the same public space is crucial from a public policy perspective. Understanding the dynamics of segregation as well as the spatial and social heterogeneity in transportation costs can lead to better select policy tools to fight segregation. For instance, in the presence of large transportation costs for poor people in Marseille city center, more public transportation infrastructure could reduce the isolation of some of the poorest neighborhoods.

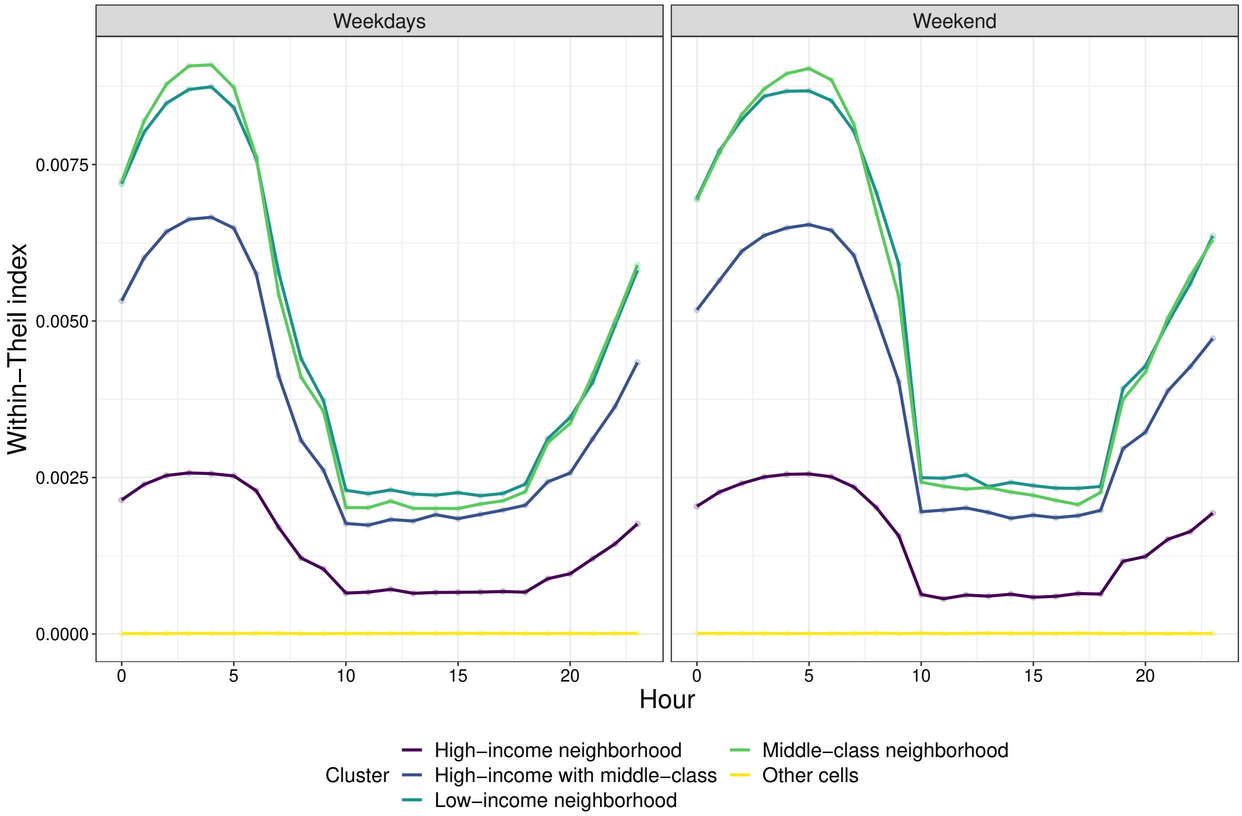
7 Conclusion

From the combination of phone and tax data, we study the evolution of segregation within typical days and the structure of spatial frictions in three French cities. We propose an innovative methodology to estimate phone users' income and derive segregation indices that account for mobility. Our method is well-suited to many anonymized dataset where individuals positions are recorded but their characteristics are unknown. Measures of segregation dynamics are robust to the home and income assignment methods adopted. We show that segregation goes down significantly during the day. In that sense, residential data lead to overestimate experienced segregation (Davis et al., 2019; Athey et al., 2019). However, residential segregation is still of interest. In the first place, living environment determines many outcomes in the life of an individual but also affects his descendants (Chetty and Hendren, 2018). In the other hand, since people spend around two thirds of their time at home (Büchel et al., 2019), residential surroundings still determine a large part of experienced segregation. Spatial frictions will play a key role in making experienced segregation and residential segregation differ. One should keep in mind that living in the same place does not guarantee that people interact together. Social segregation, which is more related to the interactions, should be the object of further research.

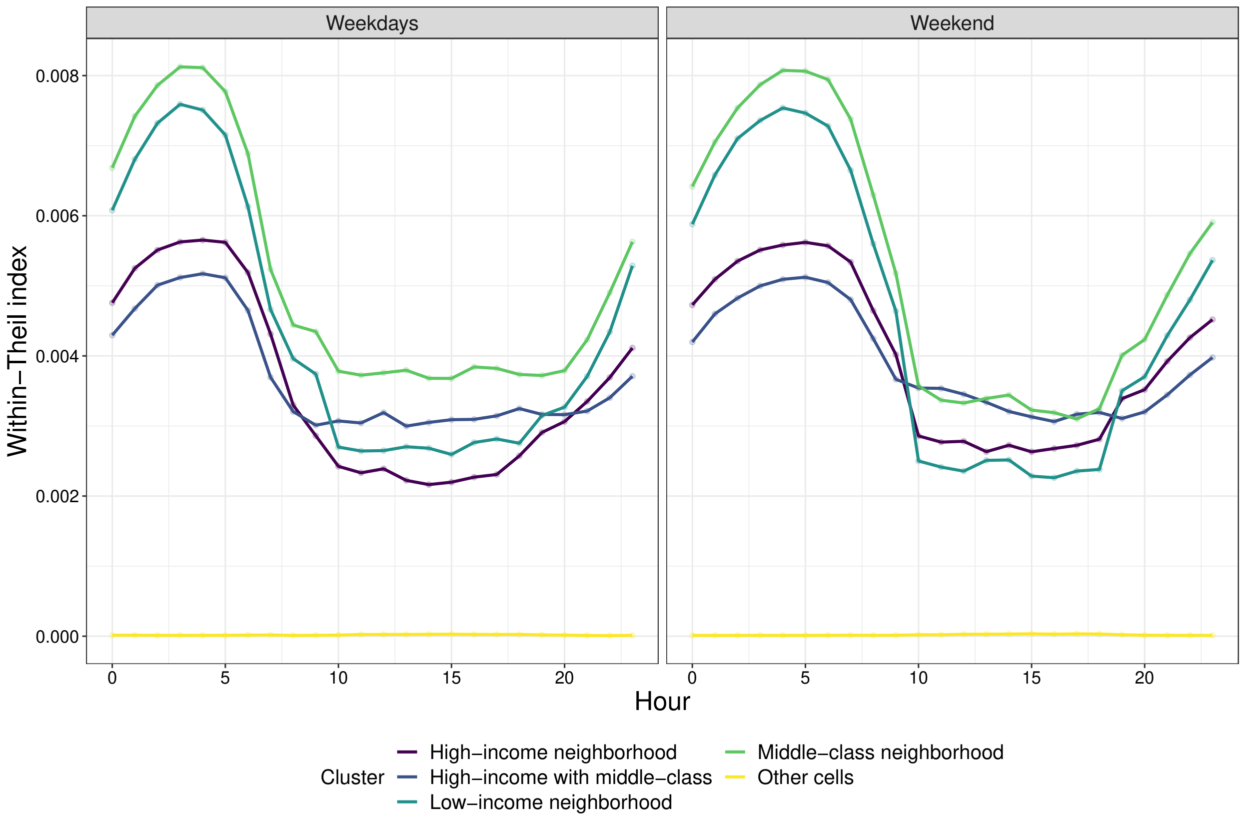
Measured with an entropy based index (Theil, 1972), we estimate that segregation is around 70% lower during daytime than during nighttime. With a dissimilarity index (Duncan and Duncan, 1955), variations are less extreme (around 50%) and comparable with Davis et al. (2019) and Athey et al. (2019) orders of magnitude. Most of the decrease in segregation happens early in the morning: at 9pm, segregation is already 40% lower than nighttime segregation at its acme. Segregation goes up when people start to go back home. Our empirical strategy to estimate the heterogeneity of spatial frictions across income groups and city areas is based on a gravity model. We use a zero-inflated count data model in order to account for censored flows and measurement error. The results help to understand better why some places are more segregated than others during daytime. For instance, in Paris, the spatial frictions are stronger for people living in suburban areas, where most low-income people live. The structure of mobility frictions differs across cities. In Marseille, a city with a large center but limited public transportation, people living in city centers face higher travel costs. In Lyon and Paris, two cities whose spatial and social structures are similar, flows between suburban areas are more affected by distance. Urban structure, more than income in both cities, is a strong determinant in transportation cost heterogeneity.

Some of the limits to our approach stem from the fact that phone use in 2007 was less frequent than in recent years. With more recent data, the understanding of segregation could be significantly improved. GPS data would also provide a better precision in the detection of individual coordinates and improve the possibility to zoom around some places. Athey et al. (2019), for instance, use 150m square cells, enabling to draw a very detailed analysis on segregation by public space type. Our 500 meters cells are still a large area that affects the ability to characterize cells based on amenities. However, zooming around some places with a focal length of this diameter is already a significant improvement in our knowledge of spatial inequalities.

A more complete vision of segregation helps calibrating the most appropriate instruments to reduce excessive spatial inequalities. It is crucial to ensure conclusions derived from mobile phone data are consistent with existing knowledge regarding urban segregation and produce information that could not be measured from classical sources. Exploring the infra-day dynamics of segregation while ensuring levels consistent with segregation measured from tax data is a way to satisfy both requirements. As automatically generated geocoded traces are expected to produce more and more knowledge for economists and social scientists, it is necessary to provide a rigorous framework for combining big data with complementary sources. Some of the challenges regarding mobile phone data combination with official statistics sources are addressed in this paper. Combined with tax data, mobile phone data help grasping better the dynamics of segregation at high temporal frequency and its interaction with spatial frictions at fine spatial granularity.



(a) Low-income



(b) High-income

Figure 9 – Contribution to overall segregation

References

Alonso, William (1964). *Location and Land Use: towards a General Theory of Land Rent.* Harvard University Press Cambridge, MA.

Anderson, James E and Eric Van Wincoop (2003). "Gravity with gravitas: a solution to the border puzzle". In: *American economic review* 93.1, pp. 170–192.

Athey, Susan et al. (2019). *Experienced Segregation*. Working Paper.

Aurenhammer, Franz (1991). "Voronoi diagrams—a survey of a fundamental geometric data structure". In: *ACM Computing Surveys (CSUR)* 23.3, pp. 345–405.

Banzhaf, H Spencer and Randall P Walsh (2013). "Segregation and Tiebout sorting: The link between place-based investments and neighborhood tipping". In: *Journal of Urban Economics* 74, pp. 83–98.

Beine, Michel, Simone Bertoli, and Jesús Fernández-Huertas Moraga (2016). "A practitioners' guide to gravity models of international migration". In: *The World Economy* 39.4, pp. 496–512.

Bell, Wendell (1954). "A Probability Model for the Measurement of Ecological Segregation". In: *Social Forces* 32.4, pp. 357–364. ISSN: 00377732, 15347605. URL: http://www.jstor.org/stable/2574118.

Black, Sandra E (1999). "Do better schools matter? Parental valuation of elementary education". In: *The Quarterly Journal of Economics* 114.2, pp. 577–599.

Büchel, Konstantin et al. (2019). "Calling from the outside: The role of networks in residential mobility". In:

Card, David, Alexandre Mas, and Jesse Rothstein (2008). "Tipping and the Dynamics of Segregation". In: *The Quarterly Journal of Economics* 123.1, pp. 177–218.

Chamboredon, Jean-Claude and Madeleine Lemaire (1970). "Proximité spatiale et distance sociale. Les grands ensembles et leur peuplement". In: *Revue française de sociologie*, pp. 3–33.

Chetty, Raj and Nathaniel Hendren (2018). "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*". In: *The Quarterly Journal of Economics* 133.3, pp. 1107–1162. ISSN: 0033-5533. DOI: 10.1093/qje/qjy007. eprint: https://academic.oup.com/qje/article-pdf/133/3/1107/25705047/qjy007.pdf. URL: https://doi.org/10.1093/qje/qjy007.

Davis, Donald R. et al. (2019). "How Segregated Is Urban Consumption?" In: *Journal of Political Economy* 127.4, pp. 1684–1738. DOI: 10.1086/701680. eprint: https://doi.org/10.1086/701680. URL: https://doi.org/10.1086/701680.

DiPasquale, Denise and William C Wheaton (1996). *Urban economics and real estate markets*. Vol. 23. 7. Prentice Hall Englewood Cliffs, NJ.

Dos Santos, Manon Domingues, Yannick L'Horty, and Elisabeth Tovar (2010). "Ségrégation urbaine et accès à l'emploi". In: *Revue d'économie régionale et urbaine* 1, p. 4.

Duncan, Otis Dudley and Beverly Duncan (1955). "A methodological analysis of segregation indexes". In: *American sociological review* 20.2, pp. 210–217.

Duranton, Giles and Diego Puga (2003). *Micro-Foundations of Urban Agglomeration Economies*. Working Paper 9931. National Bureau of Economic Research. DOI: 10.3386/w9931. URL: http://www.nber.org/papers/w9931.

Einav, Liran and Jonathan Levin (2014). "Economics in the age of big data". In: *Science* 346.6210, p. 1243089.

Fack, Gabrielle and Julien Grenet (2010). "When do better schools raise housing prices? Evidence from Paris public and private schools". In: *Journal of public Economics* 94.1-2, pp. 59–77.

Floch, Jean-Michel (2017). "Standards of living and segregation in twelve French metropolises". In: *Economie et Statistique* 497.497, pp. 73–96.

Glaeser, Edward L et al. (2018). "Big data and big cities: The promises and limitations of improved measures of urban life". In: *Economic Inquiry* 56.1, pp. 114–137.

Halbwachs, Maurice (1906). *Les expropriations et le prix des terrains à Paris (1860-1900)*.

Head, Keith and Thierry Mayer (2014). "Gravity equations: Workhorse, toolkit, and cookbook". In: *Handbook of international economics*. Vol. 4. Elsevier, pp. 131–195.

Heckman, James J (1979). "Sample selection bias as a specification error". In: *Econometrica: Journal of the econometric society*, pp. 153–161.

Helpman, Elhanan, Marc Melitz, and Yona Rubinstein (2008). "Estimating trade flows: Trading partners and trading volumes". In: *The quarterly journal of economics* 123.2, pp. 441–487.

Krings, Gautier et al. (2009). "Urban gravity: a model for inter-city telecommunication flows". In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.07, p. L07003.

Lambert, Diane (1992). “Zero-inflated Poisson regression, with an application to defects in manufacturing”. In: *Technometrics* 34.1, pp. 1–14.
Le Roux, Guillaume, Julie Vallée, and Hadrien Commenges (2017). “Social segregation around the clock in the Paris region (France)”. In: *Journal of Transport Geography* 59, pp. 134–145.
Massey, Douglas S and Nancy A Denton (1988). “The dimensions of residential segregation”. In: *Social forces* 67.2, pp. 281–315.
Oberti, Marco and Edmond Préteceille (2016). *La ségrégation urbaine*. La Découverte.
Openshaw, Stan (1984). “The modifiable areal unit problem”. In: Geo Abstracts University of East Anglia.
Papadopoulos, Georgios and J.M.C. Santos Silva (2012). “Identification issues in some double-index models for non-negative data”. In: *Economics Letters* 117.1, pp. 365–367. DOI: 10.1016/j.econlet.2012.06. URL: https://ideas.repec.org/a/eee/ecolet/v117y2012i1p365-367.html.
Papadopoulos, Georgios and Joao M C Santos Silva (2008). *Identification issues in models for underreported counts*. Economics Discussion Papers 3552. University of Essex, Department of Economics. URL: https://ideas.repec.org/p/esx/essedp/3552.html.
Park, Robert E (1926). “The urban community as a spatial pattern and a moral order”. In: *Urban social segregation*, pp. 21–31.
Pinçon, Michel and Monique Pinçon-Charlot (2009). *Paris. Quinze promenades sociologiques*. Payot.
— (2010). *Sociologie de paris*. La Découverte.
Reardon, Sean F and David O'Sullivan (2004). “Measures of spatial segregation”. In: *Sociological methodology* 34.1, pp. 121–162.
Reardon, Sean F, John T Yun, and Tamela McNulty Eitle (2000). “The changing structure of school segregation: Measurement and evidence of multiracial metropolitan-area school segregation, 1989–1995”. In: *Demography* 37.3, pp. 351–364.
Ricciato, Fabio et al. (2015). *Estimating population density distribution from network-based mobile phone data*. Publications Office of the European Union.
Sampson, Robert J., Jeffrey D. Morenoff, and Thomas Gannon-Rowley (2002). “Assessing “Neighborhood Effects”: Social Processes and New Directions in Research”. In: *Annual Review of Sociology* 28.1, pp. 443–478. DOI: 10.1146/annurev.soc.28.110601.141114. eprint: https://doi.org/10.1146/annurev.soc.28.110601.141114. URL: https://doi.org/10.1146/annurev.soc.28.110601.141114.
Santos Silva, JMC and Silvana Tenreyro (2006). “The log of gravity”. In: *The Review of Economics and statistics* 88.4, pp. 641–658.
— (2011). “Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator”. In: *Economics Letters* 112.2, pp. 220–222.
Schelling, Thomas C (1969). “Models of segregation”. In: *The American Economic Review* 59.2, pp. 488–493.
Staub, Kevin E and Rainer Winkelmann (2013). “Consistent estimation of zero-inflated count models”. In: *Health Economics* 22.6, pp. 673–686.
Theil, Henry (1972). *Statistical decomposition analysis; with applications in the social and administrative sciences*. Tech. rep.
Tiebout, Charles M (1956). “A pure theory of local expenditures”. In: *Journal of political economy* 64.5, pp. 416–424.
Verdugo, Gregory, Sorana Toma, et al. (2018). *Can Public Housing Decrease Segregation? Lessons and Challenges from Non-European Immigration in France*. Tech. rep. Observatoire Francais des Conjonctures Economiques (OFCE).
Voronoi, Georges (1908). “Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites.” In: *Journal für die reine und angewandte Mathematik* 133, pp. 97–178.
White, Michael J (1986). “Segregation and diversity measures in population distribution”. In: *Population index*, pp. 198–221.

	mean	s.d.	min	P10	P25	median	P75	P90	max
Size (km^2)	30.01	40.57	0.0022	0.18	1.25	10.9	47	85.87	424.51
Population in tax data (voronoi)	3427	2769	1	470	1295	2871	4882	7082	28,630
Number of cells per voronoi	143	179	1	4	12	63	225	392	1788
Number of voronoi per cell	1.2	0.46	1	1	1	1	1	2	28

Table 9 – Spatial units: summary statistics

A Methodology to combine sources

That technical appendix describes the methodology adopted to use the spatial dimension of mobile phone data.We will first explain how we probabilize phone users presence at 500 meters cell level before turning to the estimation of home and income status.

A.1 Spatial granularity

Precise phone user location is unknown because communications are reported at the antenna level. Detecting phone users location requires to map space with an antenna coverage. We follow the mobile phone literature by adopting voronoi tesselation (Voronoi, 1908; Aurenhammer, 1991) as our coverage model. Voronoi polygons map each point to the closest antenna.[21] Each antenna is associated with a convex polygon (hereafter called "voronoi") that represents its coverage area. Figure 11 shows a simple graphical representation of voronoi construction.

The main problem with such an approach is that the size of the voronois depends on the local density of antennas: in regions where the density is low (rural areas), voronois are large, which adds uncertainty to the phone user location (Figure 10). In other words, Voronoi tesselation produces highly heterogeneous spatial units which might result in problematic spatial aggregations of tax data, commonly referred as MAUP (*modifiable area unit problem*, Openshaw, 1984). Orders of magnitude that are reported on Table 9 show that voronoi polygons are very heterogeneous in both size and population. Since we combine mobile phone and tax data at the spatial unit level, it is important to choose units that are comparable in terms of size. While this is a fundamental aspect for the quality of the work carried out on mobile phone data, this issue is little explored in the literature. The framework presented in this paper ambitions to make the combination between phone and socioeconomic data more consistent by probabilizing phone users at the cell level and applying spatial aggregations based on the same level of spatial granularity.

A.2 Individual presence probabilization and home estimation

All along this paper, we probabilize phone users' locations by transforming events observed at voronoi level $(v)_{v \in \mathcal{V}}$ into cell level events $(c)_{c \in \mathcal{C}}$ ($\mathcal{V}$ and $\mathcal{C}$ are, respectively, the collection of voronoi and cells). Bayes' rule is used to that end. A stylized example is showed on Figure 12. The idea is to collect probabilities at cell level c knowing that a phone event occured into voronoi/antenna v.

Abstracting from time dimension, let's denote $\mathcal{S}(v)$ voronoi v surface and $\mathcal{S}(c \cap v)$ intersection area

[21] In real life, antennas coverage model is more complicated and depends on antenna characteristics, traffic use, meteorological conditions... The combination of these elements are hard to forecast and understand *ex-post*. Voronoi tesselation is generally seen as an agnostic model because it is equivalent in assuming that each point in space is covered by the closest antenna

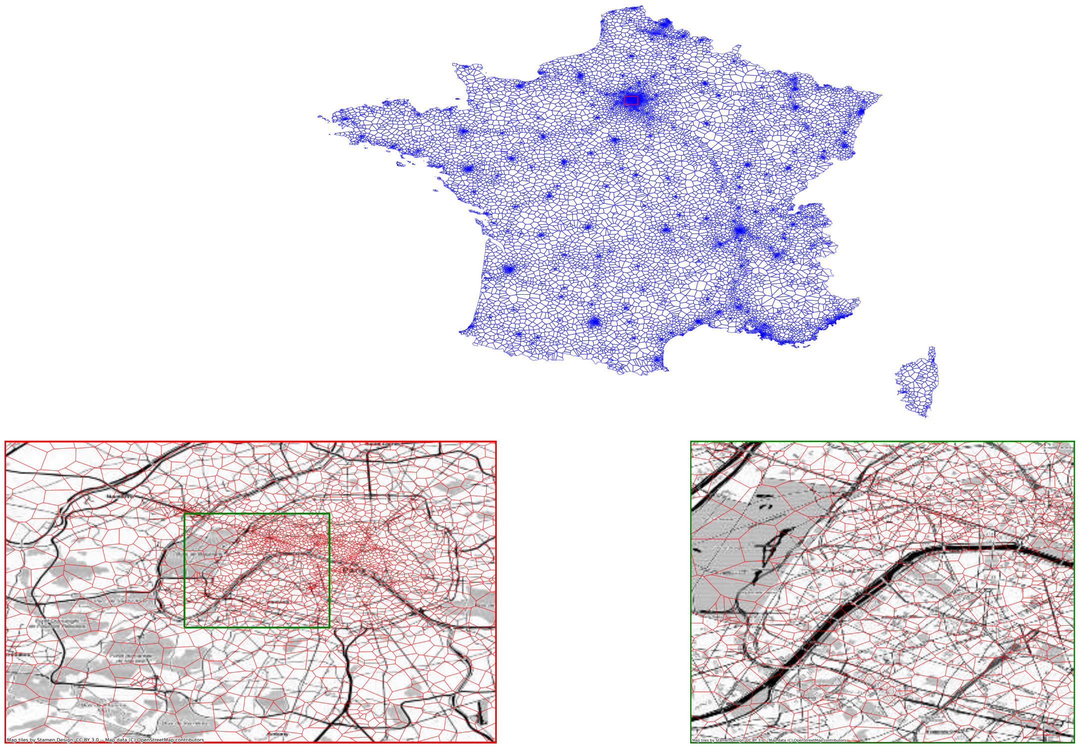


Figure 10 – Repartion of voronoi tesselation based on 2007 antennas location in France

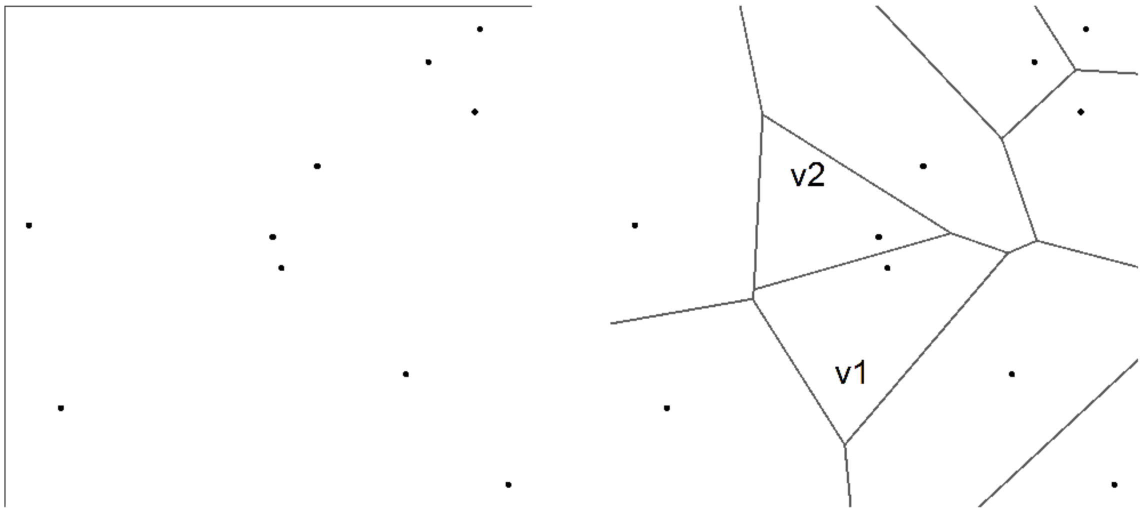


Figure 11 – Voronoi tesselation: principle

between cell c and voronoi v. The probability that an event observed by antenna v occurred in cell c is equal to

$$\mathbb{P}(c|v) = \frac{\mathbb{P}(c \cap v)}{\mathbb{P}(v)} = \frac{\mathcal{S}(c \cap v)}{\mathcal{S}(v)} \tag{16}$$

with an implicit assumption that every point within a voronoi has the same likelihood of being the origin of the call. Bayes' rule is then used to recollect presence at cell level

$$\forall c \in \mathcal{C}, \quad \mathbb{P}_{x,t}(c) = \sum_{v \in \mathcal{V}} \mathbb{P}\big(c\big|v\big)\mathbb{P}_{x,t}(v) \tag{17}$$

where $\mathcal{V}$ is the set of voronoi polygons and $\mathcal{C}$ the set of 500 meters cells. The probability $\mathbb{P}_{x,t}(v)$ is the frequency by which v appears in phone user trace at time t. The time dimension t has been reintroduced to pinpoint the fact that location in a given place is conditional to a time window.

A.3 Home estimation

Without prior knowledge on individual characteristics, we estimate phone users home cells from their nighttime trace (defined in the broad sense as any time between 7pm and 9am). We use our knowledge of population distribution from administrative data as a prior to improve the quality of home detection. Cells with high population density in administrative data should be more likely to host someone in phone data.

Using tax data, the population density is introduced as a prior to weight more cells within voronoi where residential density is higher. On the contrary, non dense areas are penalized by that prior. Residence probabilities reweighting at cell level is given by

$$\mathbb{P}_x\big(c^{\text{home}}|v\big) \propto \underbrace{\mathbb{P}\big(c^{\text{home}}\big)}_{\substack{\text{prior from}\\ \text{population density}}} \quad \underbrace{\mathbb{P}_x\big(v|c\big)}_{\substack{\text{ratio between areas}\\ \frac{s(v\cap c)}{s(c)}}} \tag{18}$$

where the superscript *home* has been added to emphasize the fact that this bayesian approach is valid for home detection only. The $\propto$ operator emphasizes that it is a proportionality relationship between terms, not a numerical equality[22]. Eq. (18) shows that the prior distribution is used to reweight information derived from a location using our knowledge on antennas coverage area and characteristics of cells inside that coverage area.

Phone user x home probabilities can be written as

$$\forall c_i \in \mathcal{C}, \quad \nu_x^{\text{home}}(c) = \frac{1}{\alpha_x} \sum_{v \in \mathcal{V}} \mathbb{P}_x\big(c^{\text{home}}|v\big)\mathbb{P}(v) \tag{19}$$

where $\mathbb{P}(v)$ is the frequency by which x appears into v_j in nighttime trace and α_x a normalization term that ensures the measure ν_x^{home} sums to one.

Introducing spatial prior makes sense when estimating phone user's residence. However, it does not when looking at presence probabilities. It is expected that population estimates based on phone data will coincide with those based on tax data. It justifies to weight more the cells containing more population. Reweighting presence is only valid at night, when people are supposed to be at home. On the other hand, when it comes to places frequented, it cannot be assumed that calls only take place in residential spaces. People may be in buildings, in forests or on the road. Thus, we make no assumption

[22] Eq. (18) can be derived the following way. From Bayes rule,

$$\mathbb{P}_x(c^{\text{home}}|v) = \frac{1}{\mathbb{P}_x(v)}\mathbb{P}_x(v|c^{\text{home}})\mathbb{P}_x(c^{\text{home}})$$

Prior from population density is the same for everybody, which makes possible to write $\mathbb{P}(c^{\text{home}})$ rather than $\mathbb{P}_x(c^{\text{home}})$ in eq. (18). The grid coverage of France being the same for home detection and presence detection, intersection areas do not differ. This enable to write $\mathbb{P}_x(v|c^{\text{home}}) = \mathbb{P}_x(v|c)$. Finally, at individual level, for a given v, $\mathbb{P}_x(v)$ is a constant and thus can be ignored when considering proportionality. Everything put together yields (18).

on where they have more chance to be located when considering presence probabilities ; population prior is used for home probabilities only. Figure 13 shows the substantial improvement of the quality of population estimates due to adding prior information on population distribution in the Bayesian model of localization.

Phone user status simulation is processed in two steps. First, we simulate phone users living place given the whole sequence of individual home probabilities. Then, we use cell level probabilities of belonging to group g to simulate phone users income groups. This is a Monte-Carlo simulation of phone users probability of belonging to low- or high-income group with mixtures of income distributions (the weights being given by the probability distribution of home locations ν_x^{home}).

A.4 Income simulation

The principle of income assignment is to simulate phone user's income status given local income indicators (eq. 6) and phone users' likely homes (eq. 18). Let's denote income groups $k = \{\text{D1}, \text{D9}\}$. We proceed in two steps, for a given phone user x:

1. Draw phone user x residence (c_x^{home}) from her likely home track ν_x^{home}. The phone user's home c_x^{home} will be used to simulate the income group an individual belongs to
2. Given cell residence c_x^{home}, we draw socioeconomic status (does phone user belongs to k?) from cell probabilities of belonging to low-income or high-income group given by eq. (6). In other words, for the income group of interest g, knowing phone user home cell is c_x^{home}

$$\mathbb{P}\big(x \in g \,|\, c_x^{\text{home}}\big) = \mu^g_{c_x^{\text{home}}} \tag{20}$$

where we denote by k the income group of interest. We perform high- and low-income simulations independently to compute bimodal segregation indices. Our approach is a Monte-Carlo simulation of phone user's income group in a context where both home and income are unknown. Indeed, for K iterations of the procedure we develop, the estimator at individual level for the probability of belonging to income group g would be

$$\widehat{\mathbb{P}_x}(x \in g) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{\{x\in g\}}$$

which is, an estimator for the probability of being in group g, mixture of two probability distributions (ν_x^{home} and $\mu^g_{c_x^{\text{home}}}$),

$$\mathbb{P}(x \in g) = \mathbb{P}(y_x \leq y^g) = \mathbb{E}\left(Y \leq y^g\right) = \int_{\mathcal{C}} \int_{-\infty}^{y^g} dF_c d\mu_x^{\text{home}}$$

where F_c is the income distribution inside cell c.

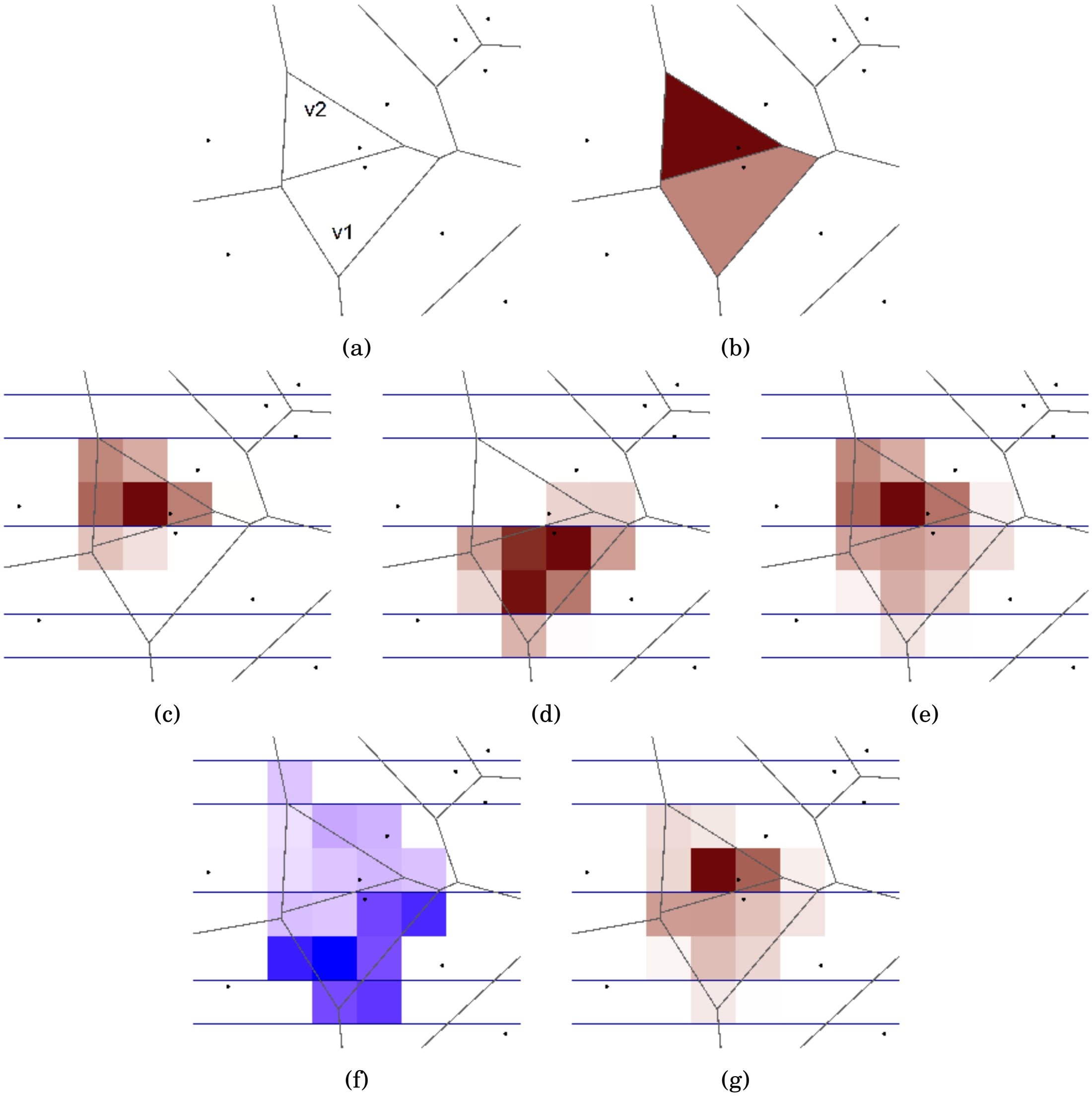
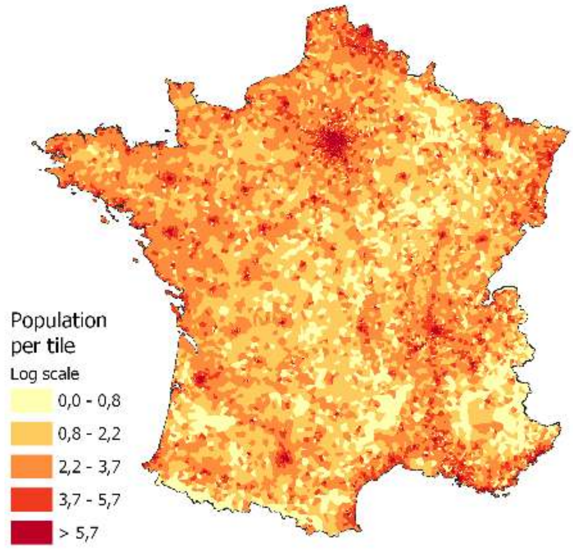


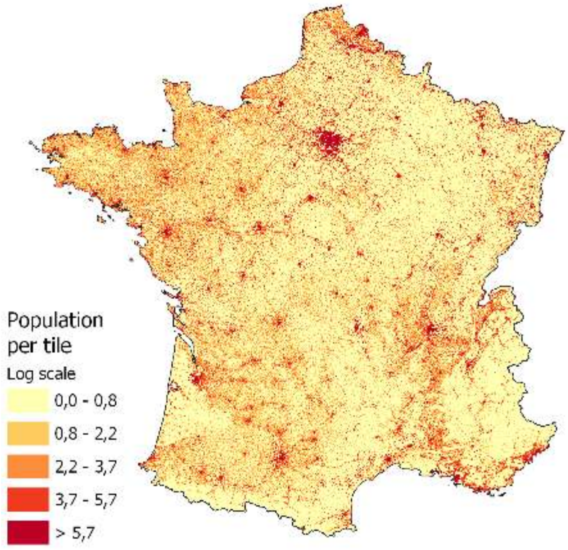
Figure 12 – Probabilization: illustration

Lecture : this is an example to illustrate the home detection process for a given phone user. We focus on two specific Voronoi polygons v_1 and v_2 (a). By night, over the month, 2/3 of the phone events of the user are located in v_2 - i.e. processed by the corresponding antenna - and 1/3 are located in v_1 (b). We allocate these events to the grid based on conditional probabilities $(\mathbb{P}(c_i|v_j))_{i,j}$ computed via Bayes' rule. If allocation is performed using only relative shares of tiles in the Voronoi polygon they intersect (see (c) and (d)) - i.e. we use an uninformative prior equal to 1 for all tiles - then probabilized home detection is given by (e). However, if we assume that population is actually denser in tiles that intersect v_1 (f), then probabilized home detection is modified when the prior is incorporated (g).

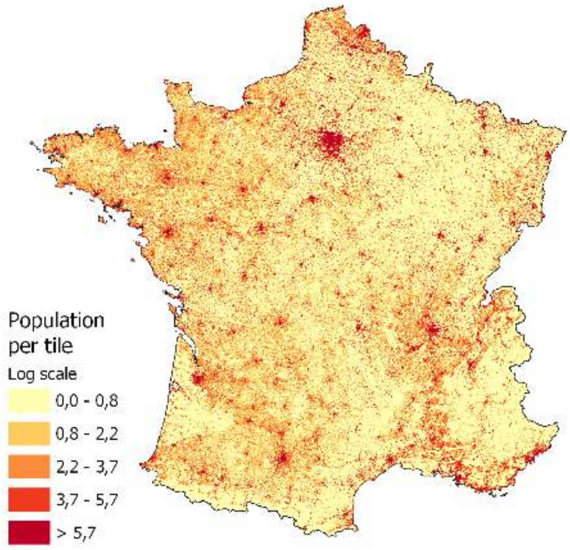
Figure 13 – Population counts from probabilistic home detection at France level



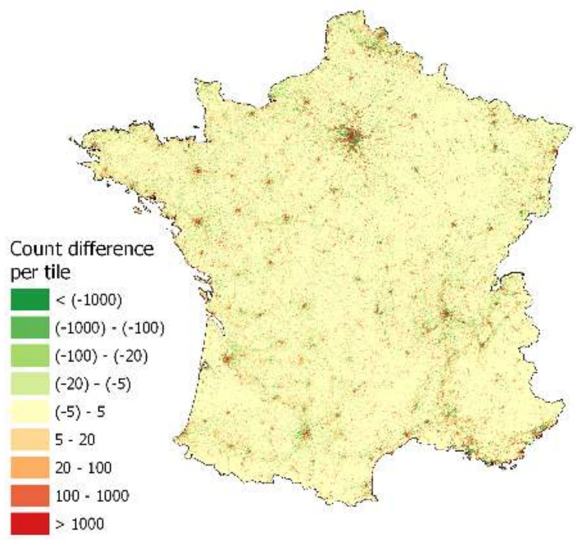
(a) Population distribution without prior



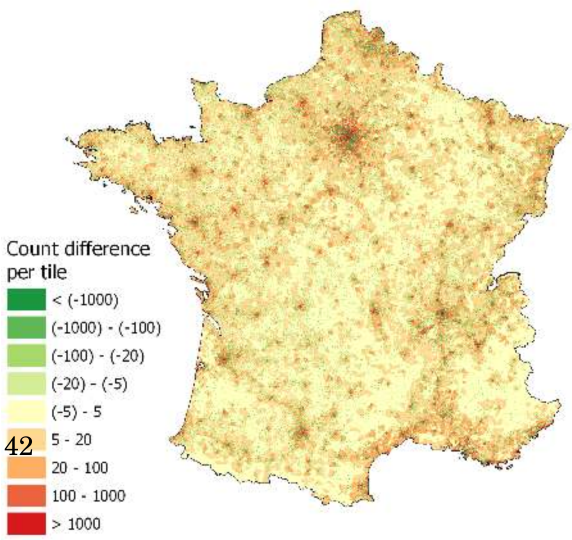
(b) Population distribution with prior



(c) Population distribution in tax data



(d) Difference between population distribution without prior and reference



(e) Difference between population distribution with prior and reference

B Robustness checks

B.1 Comparing segregation in tax and phone data

In order to compare segregation indicators derived from the different data sources, it is necessary to focus on nighttime hours. Figure 5a shows that only a fraction of our sample is observed during nighttime. This is natural because the main activity during nighttime is not making calls or writing text messages. A candid use of nighttime traces leads to segregation indices that are biased downward and do not produce enough spatial clustering. In the case of Paris, the segregation indices, without correction, would be 35 to 40% lower than those of the tax data. It is thus necessary to understand the reason behind the discrepancy between tax and phone data nighttime segregation.

Several factors could explain the discrepancy between sources. We investigate some of them in Table 10 by reporting dissimilarity indices in different tax data bootstrap settings. The three specifications we adopt are designed to evaluate the sensitivity of segregation indices to simulations. Confidence intervals at 95% are reported. The first bootstrap setting (specif. 1) is a standard bootstrap inside cells where tax data observations are sampled with replacement. The goal is to assess whether randomly drawing individuals rather than observing the population has any effect on segregation indices. The dissimilarity indices are not really affected by this randomization. Having a sample from exhaustive tax data - either one third of total population (specif. 2) or based on a home distribution density that matches the one determined from phone data (specif. 3)[23] - leads to the same conclusion. The simulation approach *per se* cannot explain the difference between segregation indices in phone and tax data.

Another candidate to explain the discrepancy between sources is the fact that only people observed at a given time t are used to compute dissimilarity indices from mobile phone data (eq. 21). Time aggregation to create a typical day is not enough to ensure every users are measured somewhere at any moment. As pointed in Table 2, our main problem is the sparse phone users trace. Two treatments have been explored to evaluate the effect of individual temporal sparsity on segregation indices. Our first approach was to create time windows regrouping more hours[24]. This allowed us to measure more users in each time window. At the slowest hour, we were measuring the presence of two thirds of the sample in this approach. Nevertheless, the levels for our night-time segregation estimators were still too low compared to those from residential data. One reason could be that taking larger time windows make phone users dispersed in more places, which can smooth differences in social composition between the cities areas. While that approach did not correct the downward bias for segregation, it had the cost of bringing less remarkable infra-day dynamics. We thus stick to the time window defined in Section 4 and investigate in another direction.

The solution applied here is consists in imputing missing users during nighttime into their home with probability 1. The effect on nighttime segregation (19:00 to 09:00) is reported in Figure 14. Within-night segregation follows the expected pattern with flatter profiles in the heart of the night. We still find that segregation increases sharply at night. Between the two extremes of nighttime segregation (8 am and 3 am), segregation is estimated to drop by 31% during weekdays and 26% during weekends.

A candid use of phone data, i.e. without any correction for users not observed during nighttime,

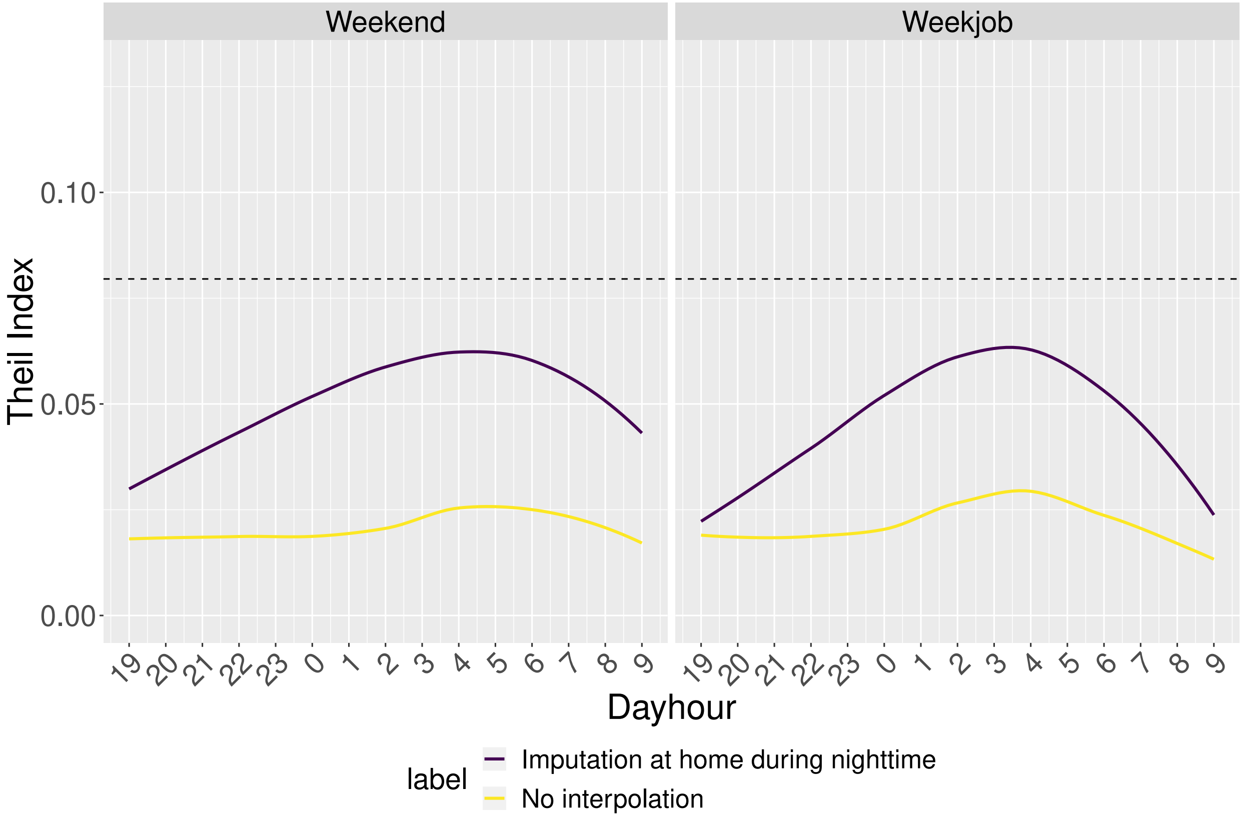
[23] The population distribution used in specification (3) corresponds to the home cell distribution that is used in phone data when performing `one stage simulation` method. This is not the one that is used in our main method to simulate home but rather the most likely home distribution we can infer from phone data.

[24] That alternative window was based on the following aggregations: 23:00-06:00, 06:00-10:00, 10:00-15:00, 15:00-19:00 and 19:00-23:00. Friday night was considered as a weekend night while Sunday night belongs to weekdays night

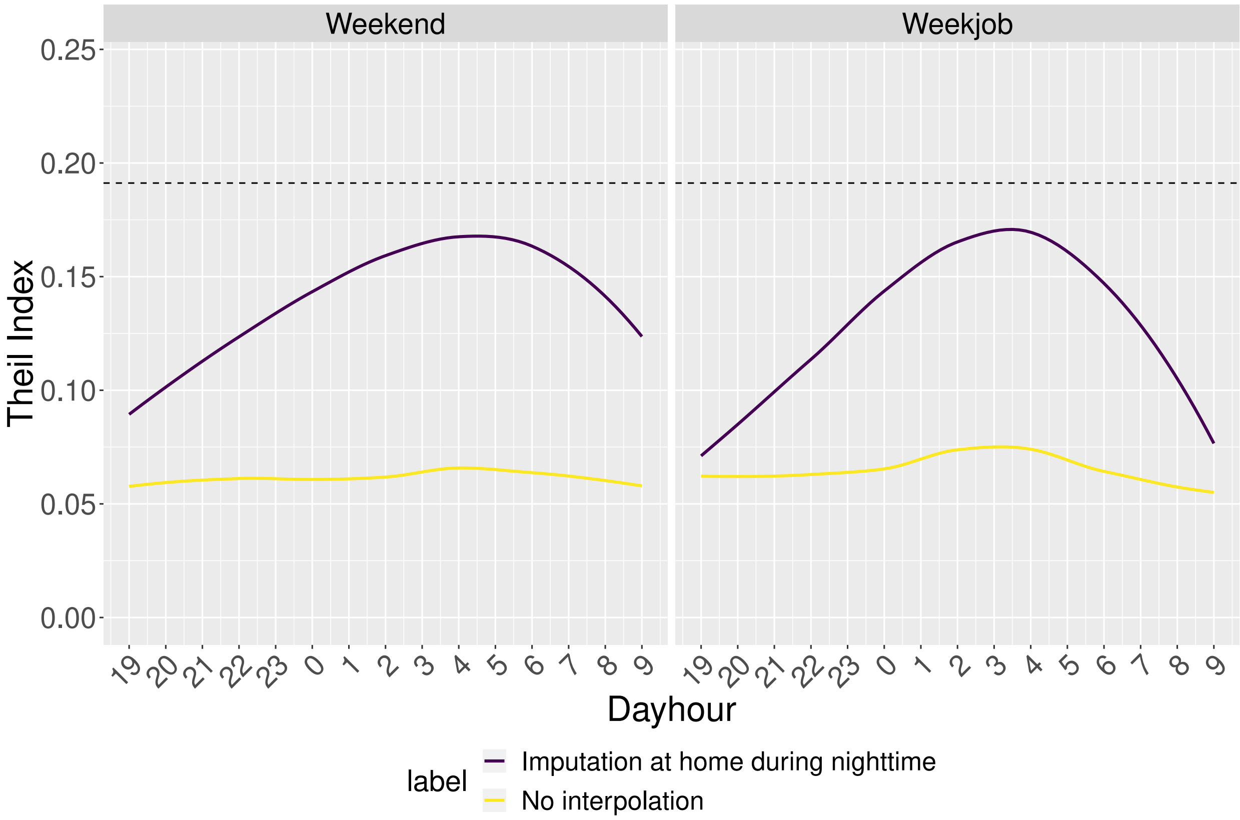
	DISSIMILARITY INDEX			
	Observed value	*Bootstrap design*		
		(1)	(2)	(3)
			LYON	
Low-income	0.295	0.298	0.300	0.298
		[0.298;0.299]	[0.299;0.302]	[0.296;0.299]
High-income	0.475	0.479	0.480	0.465
		[0.478;0.479]	[0.479;0.482]	[0.464;0.466]
			LYON	
Low-income	0.303	0.317	0.321	0.312
		[0.315;0.319]	[0.318;0.325]	[0.309;0.316]
High-income	0.379	0.397	0.403	0.403
		[0.395;0.399]	[0.399;0.405]	[0.400;0.407]
			LYON	
Low-income	0.354	0.373	0.378	0.387
		[0.371;0.375]	[0.374;0.381]	[0.382;0.391]
High-income	0.387	0.409	0.415	0.427
		[0.407;0.411]	[0.412;0.419]	[0.424;0.431]

Data:
2014 geocoded tax data over households (accounting for individuals) at France level
Notes:
Median Dissimilarity Index (eq. 2)over 100 iterations is reported. 95% confidence intervals are reported into brackets
(1): Bootstrap inside each cell with uniform weights (probability being chosen: $1/n_i$)
(2): Bootstrap inside each cell with uniform weights for 1/3 population (probability being chosen: $\frac{1}{3n_i}$)
(3): Bootstrap inside each cell with uniform weights for population from mobile phone data (probability being chosen: $\frac{1}{n_{\text{phone}}}$).

Table 10 – Effects of performing simulation in tax data



(a) Low-income



(b) High-income

Figure 14 – Paris: effect of imputing missing people into home during nighttime

would produce indices 35 to 40% lower than tax data levels. Re-scaling segregation results to get a more consistent estimates of residential segregation is thus possible by imputing position for non-observed users during nighttime. More continuous phone traces would help us in estimating more precisely residence and presence at any time. Individual passive phone data as well as recent Call Details Record would satisfy that requirement and lead to better estimates for segregation. More precise individual traces could also have the advantage of getting a finer spatial granularity that would lead to better estimates of income group by reducing the infra-cell income variance.

Assuming phone users are at home during the night can make sense given the residence definition we adopted (phone users September nighttime trace). However, it would be a strong hypothesis to also impute phone users to home during daytime. Although imputation would concern less phone users (people are more active during daytime), many phone users have no reason to be at home during daytime weekdays. Assuming they are at home might bias segregation estimators. In other words, it seems undesirable to impute missing users at home during daytime. The approach we used brings consistent orders of magnitude with recent estimates by Davis et al. (2019) or Athey et al. (2019).

B.2 Alternative home and income simulation methods

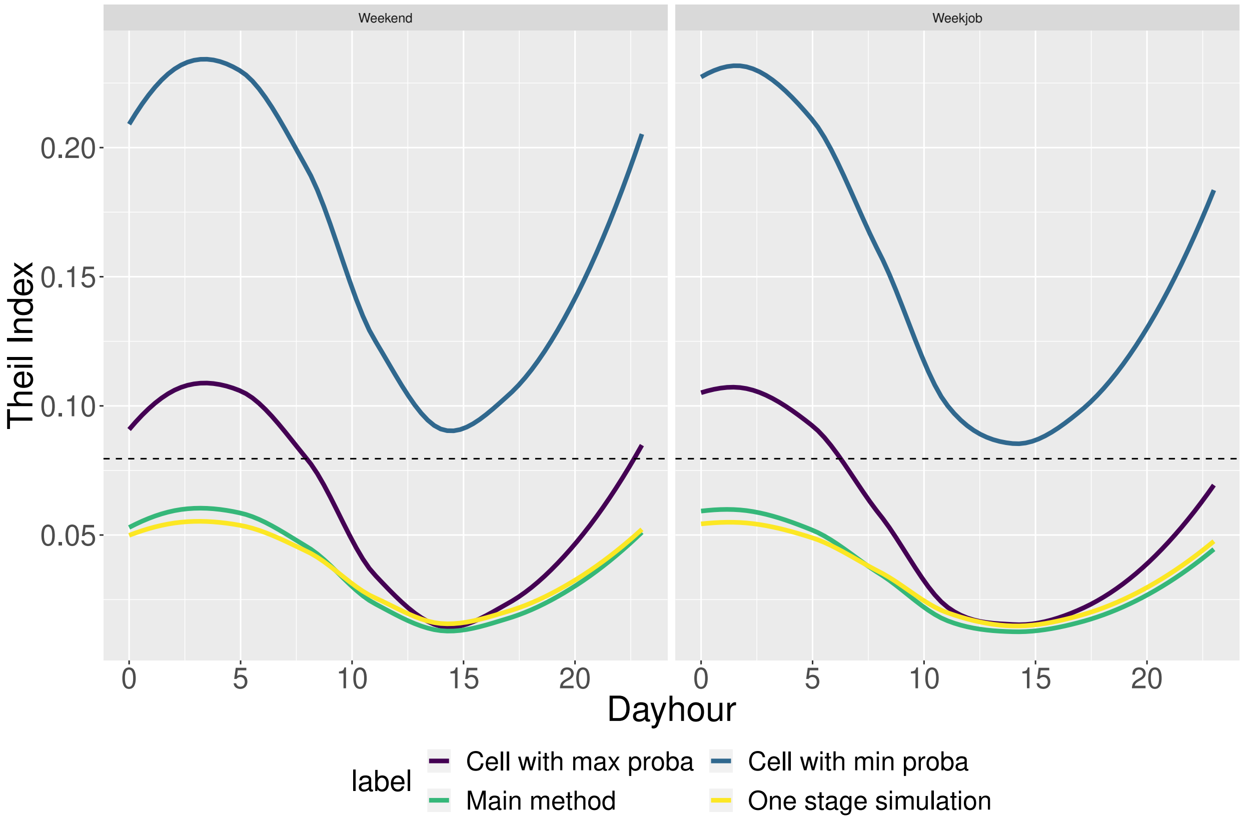
In this section, we propose robustness tests by comparing our method with other home or income estimation strategies. The last two methods of income simulation aim at evaluating whether segregation indices are sensitive to situations where the ecological fallacy, i.e. the error in income group assignment due to the error related to home detection, is maximum. The following `One stage simulation` approach enables us to investigate the stability of results in a simplified income assignment scheme.

1. `One stage simulation`: Phone user's home is not simulated from likely home track but considered as the cell where she has maximum probability of having the residence. This is a simplified version of the preceding approach where the first step, namely home simulation, is skipped. More precisely, instead of drawing home from ν_x^{home}, we take $c_x^{\text{home}} = \arg\max_{c_i} \nu_x^{\text{home}}(c_i)$. Income group assignment is unchanged.
2. `Cell with max probability`: Given a phone user likely home track, residence is assigned to the cell where the probability of being of income group k is the highest. This situation enables to have an estimates for segregation when we assign too many phone users to the sub-population of interest.
3. `Cell with min probability`: Given a phone user likely home track, residence is assigned to the cell where the probability of being of income group k is the lowest. This situation enables to have an estimate for segregation when we do not assign enough phone users to the sub-population of interest.

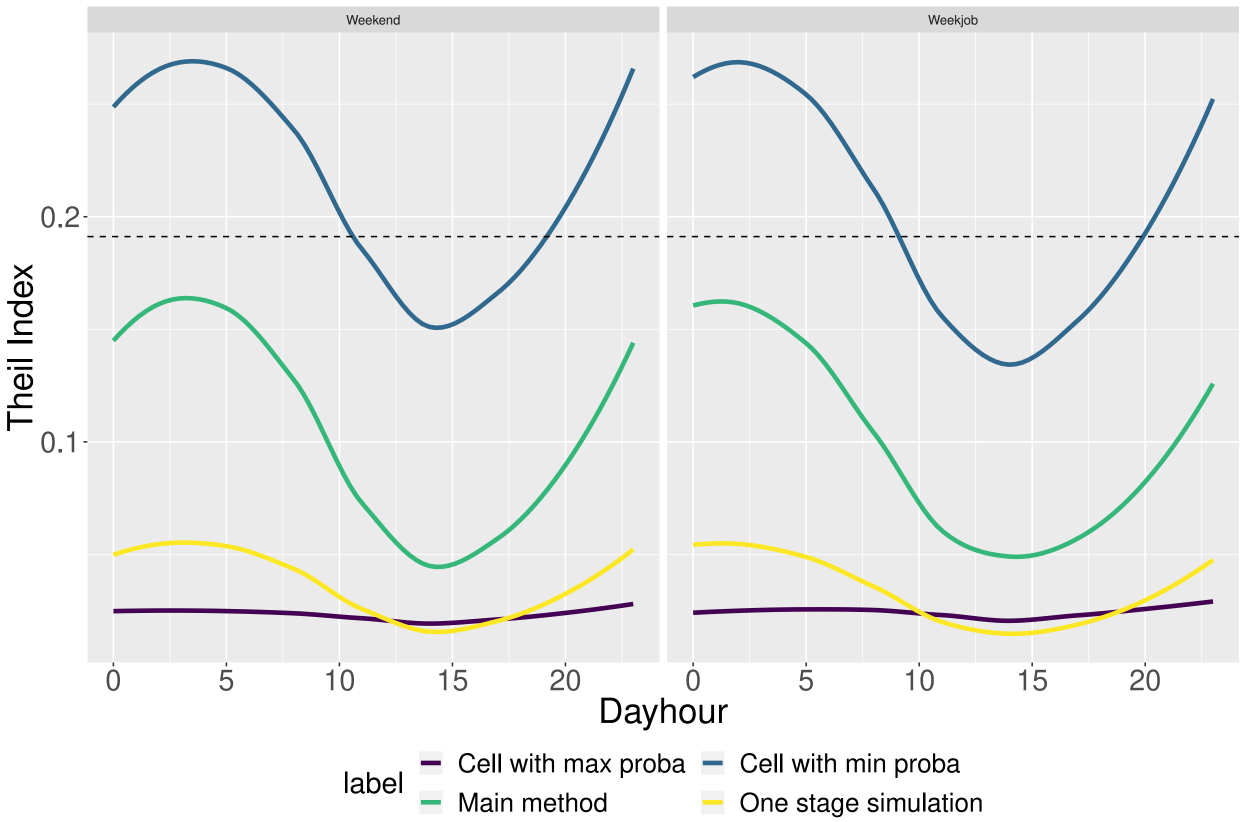
For low-income people (Figure 15a), segregation is closer from the lowest possible level than from the highest one. Simulation scheme changes the level of nighttime segregation but does not always shift up daytime segregation: in three out of four methods, daytime segregation levels are almost equal. For high-income segregation (Figure 15b), the segregation dynamics we obtain is between the two extreme methods that exaggerate ecological fallacy. Surprisingly, there is one method (`cell with max probability`) that produces a flat segregation profile. A more important point is the strong difference in level between our main methodology (two-step estimation procedure) and the one-step approach. That difference could come from the large nighttime window adopted (7pm-9am) that does not produce a good estimate of high-income people residence when selecting the place where most phone calls have been made. Taking into account the full path of nighttime trace produces a better estimate for people that living in high-income neighborhood. This point is worth to be mentioned because that residence choice criterion is still among the most used by the mobile phone literature. However, both methods are consistent from a dynamic viewpoint. For a given socioeconomic status simulation method, results are very stable. Regarding the consistency with segregation estimates from tax data, our nighttime levels are always closer to observed levels in residential data than other methods. For high-income people (Figure 15b), selecting only the most likely cell leads to strongly biased estimates of nighttime segregation. For low-income people (Figure 15a), the bias is less pronounced. These results argue in favour of the two-step approach proposed in Appendix A.

B.3 Alternative segregation indices

There is an extended literature regarding segregation measurement. Most notable contributions arise from Reardon and O'Sullivan (2004) and Massey and Denton (1988). We explore sensitivity to the segregation index by comparing a few standard segregation indices on tax data. The major indices are



(a) Low-income



(b) High-income

Figure 15 – Paris: robustness analysis

presented in Table 11. Reardon and O'Sullivan (2004) categorize segregation indices in two dimensions: exposure and evenness. The former refers to the extent that members of one group encounter members of another group (or their own group, in the case of spatial isolation) in their local spatial environment. The latter refers to the extent to which groups are similarly distributed in space. Each dimension calls for different indices. The most famous indices are the dissimilarity index (Duncan and Duncan, 1955) and the Theil (1972) index, also known as information theory index because it is based on entropy. They both measure the evenness dimension of segregation (Reardon and O'Sullivan, 2004).

Evenness measures The value of the Theil H index is given by eq. (9). The dissimilarity index measures the deviation of a cell entropy from the city-level entropy. The generalized version we propose uses cell presence probability. In the generalized version of propose, we have probability at cell level c for time t instead of observing people with probability one:

$$D_t^g = \frac{1}{2}\sum_{c\in\mathcal{C}}\left| \frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\in g}}{\underbrace{\sum_{x\in\mathcal{X}}\mathbf{1}_{x\in g}}_{\substack{\text{Number people of income group } g\\ \text{that are observed at time } t}}} - \frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\notin g}}{\underbrace{\sum_{x\in\mathcal{X}}\mathbf{1}_{x\notin g}}_{\substack{\text{Number people not in income group } g\\ \text{that are observed at time } t}}}\right| \tag{21}$$

The sums of probabilities at sub-populations level are present in the numerators (sub-population belonging to income group g are on the left hand-side, other sub-populations on the right hand-side). Sub-populations at city level observed at time t are used in the denominators. As in his standard formulation, the index can be interpreted as classically as the share of individuals in group g that would need to be moved in order to match a spatially uniform distribution pattern. With income quantiles, the uniform pattern is finding 10% of income group g in each cell. Indices are computed for the 48 time windows defined previously.

Exposure measures In this family, one can find a series of measures of segregation that are related. The first measure is the interaction index. It can be generalized as

$$_xP_y(t)^g = \sum_{c\in C}\frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\in g}}{\sum_{x\in\mathcal{X}}\mathbf{1}_{x\in g}}\frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\notin g}}{\text{card}(\mathcal{X})} \tag{22}$$

The isolation index can be derived from the the interaction index:

$$_xP_x(t)^g = \sum_{c\in C}\frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\in g}}{\sum_{x\in\mathcal{X}}\mathbf{1}_{x\in g}}\frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\in g}}{\text{card}(\mathcal{X})} = 1 -_x P_y(t)^g \tag{23}$$

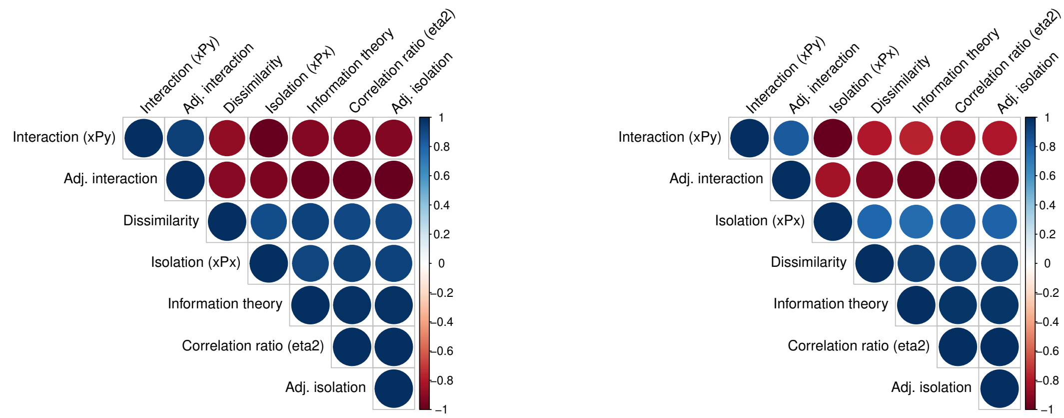
Isolation and interaction indices can also be re-normalized by the proportion of minority group in the city to get an adjusted isolation (resp. interaction) index. Another index measuring exposure is the η^2 measure:

$$\eta_t^g = \frac{_xP_x(t)^g - \frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\in g}}{\sum_{x\in\mathcal{X}}\mathbf{1}_{x\in g}}}{1 - \frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\in g}}{\sum_{x\in\mathcal{X}}\mathbf{1}_{x\in g}}} \tag{24}$$

Figure 16 shows the relationship (with a correlation matrix) between all these measures in tax data at the national level for the 125 biggest French cities. Correlation matrix is ordered using hierarchical clustering in order to present the strength of the relationship between indices. Lower correlation is equal to 0.88 in absolute value. The Theil (1972) index is highly correlated with the dissimilarity index (correlation: 0.92). The relationship between indices follows the expected pattern (the interaction index should be low when others are high, and vice-versa).

Index	Dimension	Reference
Dissimilarity Index	Evenness	Duncan and Duncan (1955)
Information theory index	Evenness	Theil (1972)
Isolation index (${}_xP_x$)	Exposure	Bell (1954)
Correlation ratio (η^2)	Exposure	White (1986)
Adjusted isolation index (adj ${}_xP_x$)	Exposure	
Interaction index (${}_xP_y$)	Exposure	Bell (1954)
Adjusted interaction index (adj ${}_xP_y$)	Exposure	

Table 11 – Segregation indexes compared

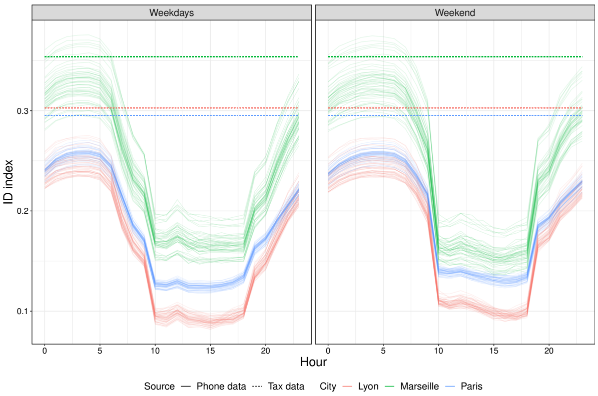


(a) Low-income

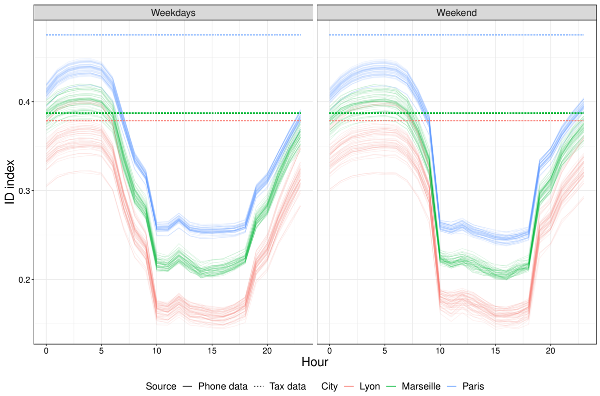
(b) High-income

Figure 16 – Correlation matrix between segregation indices

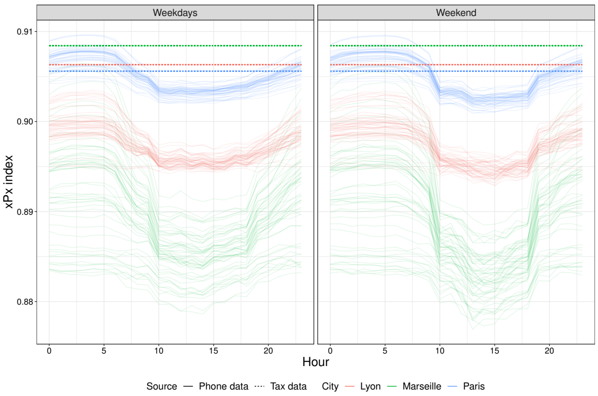
Figure 17 presents the evolution of segregation indices using phone data (see Figure 7 for Theil H index). The patterns are similar on all these figures. Measures related to interaction index should be interpreted in the opposite direction of the others. As for Theil H index, segregation level goes down during daytime. The magnitude of the infraday variation depends on the index. In most cases, for Marseille, nighttime segregation level is close to the one computed using tax data. In Paris and Lyon, we generally underestimate residential segregation for low-income people with tax data. For high-income people, levels of nighttime segregation obtained with η^2, adjusted interaction and isolation indices are very close to observed levels with tax data. The nighttime level of the dissimilarity index derived from phone data is close to the one from tax data in Lyon and Marseille. However, for Paris, nighttime segregation is still slightly underestimated with the dissimilarity index.



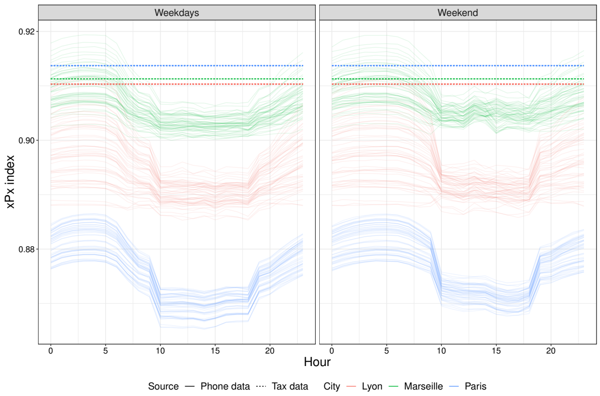
(a) Dissimilarity index: Low-income



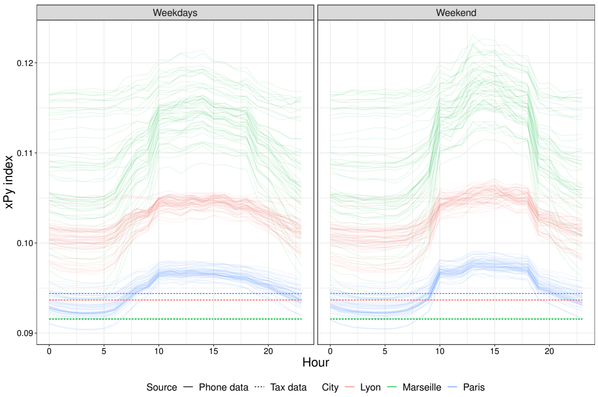
(b) Dissimilarity index: High-income



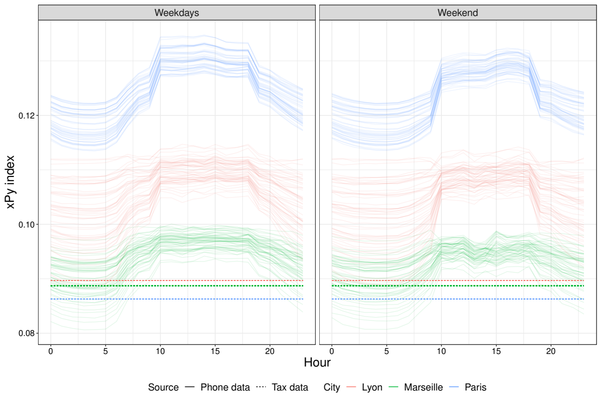
(c) Isolation index: Low-income



(d) Isolation index: High-income

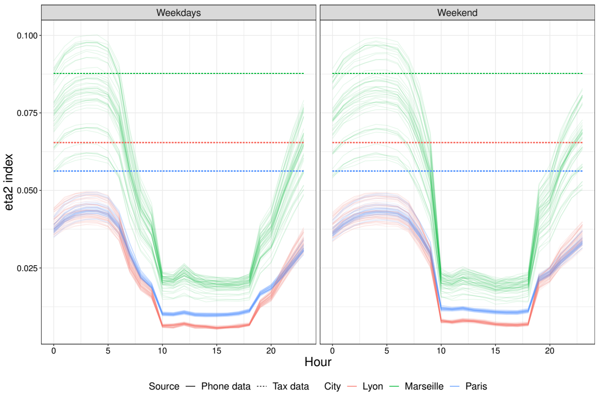


(e) Interaction index: Low-income

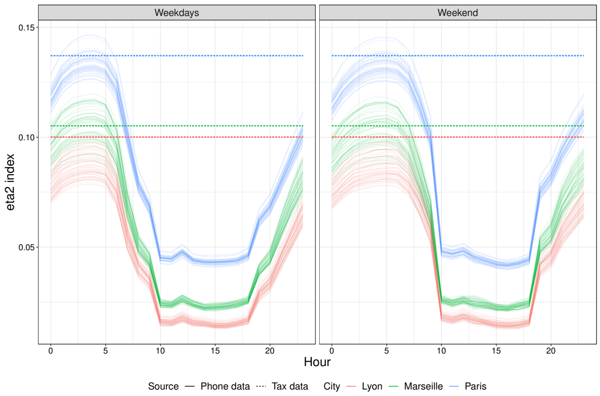


(f) Interaction index: High-income

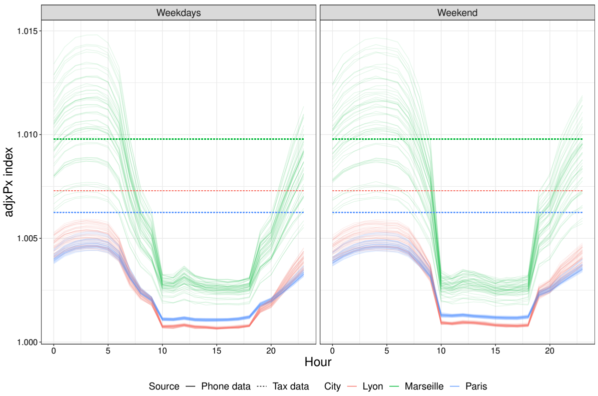
Figure 17 – Segregation indices with phone data



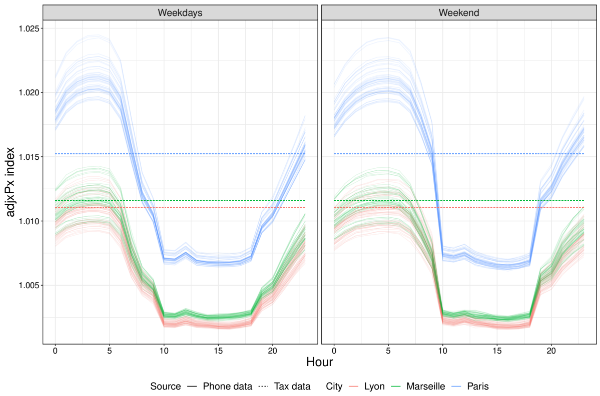
(g) η^2 index: Low-income



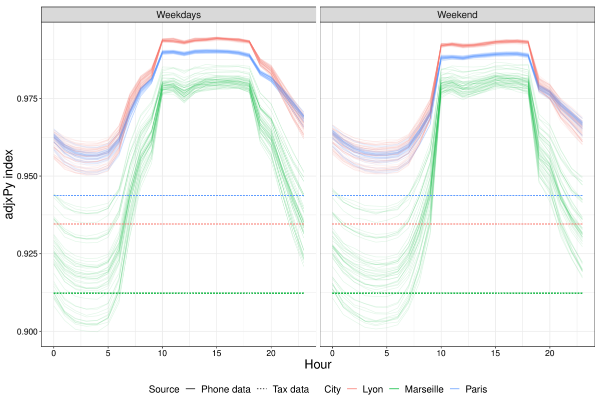
(h) η^2 index: High-income



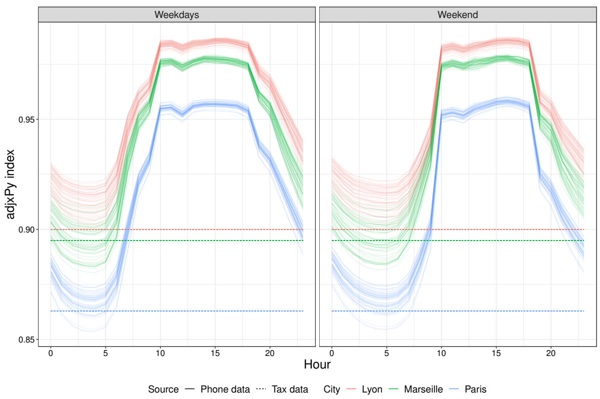
(i) Adjusted isolation index: Low-income



(j) Adjusted isolation index: High-income



(k) Adjusted interaction index: Low-income



(l) Adjusted interaction index: High-income

Figure 17 – Segregation indices with phone data (following)

C Appendix on the gravity model

C.1 Zero-inflated models

Count data models When modelling count data, researchers resort mainly to two distributions: Poisson and negative binomial distributions. The Poisson probability density function (pdf) writes:

$$f(y_i) = \mathbb{P}(Y_i = y_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \tag{25}$$

In a Poisson regression, the link function gives the relationship between endogenous and exogenous variables as

$$\mathbb{E}_{f,\theta}(Y_i|X_i) = \exp(X_i\beta) = \mu_i$$

With such model form, Poisson model is not restricted to discrete valued data and can be used to model the relationship between a continuous dependent variable and some covariates. Under some Poisson data generating process, $\mathbb{E}(Y_i|X_i) = \mathbb{V}(Y_i|X_i) = \mu_i$. Poisson model therefore yields the restriction that conditional mean and variance are equal. By definition, heteroskedasticity is accounted for in a Poisson model.

Negative binomial distributions are often used to model data when the assumption of equal mean and variance is unsuitable. A negative binomial distribution can be modelled as a Gamma-Poisson mixture. With a negative binomial distribution,

- $Y_i|\lambda_i$ is a Poisson distribution with conditional mean $\mathbb{E}(Y_i|\lambda_i) = \lambda_i$
- λ_i is determined by a Gamma distribution with mean $\mathbb{E}(\lambda_i) = \mu_i$ and variance $\mathbb{V}(\lambda_i) = \mu_i^2/\theta$

A new parameter θ is introduced to model over- or under-dispersion. The density of a negative binomial distribution is

$$\begin{aligned} g(y_i) = \mathbb{P}(Y_i = y_i) &= \int \mathbb{P}(Y_i = y_i|\lambda_i) f(\lambda_i) d\lambda_i \\ &= \frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu_i}\right)^\theta \left(\frac{\mu_i}{\theta+\mu_i}\right)^{y_i} \end{aligned}$$

One has $\mathbb{E}(Y_i|X_i) = \mu_i$ and $\mathbb{V}(Y_i|X_i) = \mu_i + \mu_i^2/\theta$. The link function writes $\mathbb{E}_{f,\theta}(Y_i|X_i) = \exp(X_i\beta) = \mu_i$. This model is sometimes called Negative Binomial I. We define $\alpha = 1/\theta$. The conditional variance rewrites as $\mathbb{V}(Y_i|X_i) = \mu_i(1+\alpha\mu_i)$. The pdf function rewrites as

$$\mathbb{P}(Y_i = y_i) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \tag{26}$$

For continuous data, (25) and (26) can be generalized using Euler's Γ function.

Both models can be estimated by maximum likelihood (ML). Hence the estimator is consistent and asymptotically normal (CAN). The Central limit theorem for negative binomial ensures in this setting that

$$\sqrt{n}\left((\widehat{\beta}, \widehat{\alpha}) - (\beta, \alpha)\right) \to \mathcal{N}(0, \mathbb{I}(\beta, \alpha))^{-1}$$

where $\mathbb{I}$ is the Fisher information matrix.

Both Poisson and negative binomial models accomodate with zero values. For instance, under a Poisson distribution, the likelihood of having a zero value is $\exp(-\lambda_i) = \exp(-\beta X_i)$. Numerous zeroes might severely bias parameter estimates. Zero-inflated models have therefore been introduced when zeros are too frequent for count data models to fit consistently observed data (Lambert, 1992).

Zero-inflated models Under zero-inflated models, a mixture of two distributions is assumed to fit observed data. With some probability π_i, the observation is zero. However, with probability $1 - \pi_i$, counts (including zeros) are generated according to a count data model.

It is often convenient to bring this model to the data by introducing two different sets of explanatory variables, some affecting the likelihood of having a zero (selection process) while others will determine the observed value y (outcome equation). The first set of variables refers typically to excluded instruments and is denoted by Z. The second set of variables (covariates) is denoted X. Z and X may or may not include terms in common.

The pdf of a zero-inflated Poisson (ZIP) model will take the following form

$$(ZIP) \qquad \mathbb{P}(Y_i = y_i|\pi_i, \mu_i) = \pi_i \mathbf{1}_{y=0} + (1-\pi_i)\mathcal{P}_{\mu_i}(y_i) \tag{27a}$$

$$= \begin{cases} \pi_i + (1-\pi_i)\exp(-\mu_i) & \text{if } y_i = 0 \\ (1-\pi_i)\frac{\mu_i^{y_i}\exp(-\mu_i)}{y_i!} & \text{if } y_i > 0 \end{cases} \tag{27b}$$

The first term in (27b) is derived from the fact that zeroes arise with probability π_i from the selection process and with probability $(1-\pi_i)f(0) = (1-\pi_i)\exp(-\mu_i)$ from the outcome process with f the Poisson pdf given by (25). The second term is just general expression of a Poisson pdf reweighted by the probability $1 - \pi_i$.

For a negative binomial (NB) distribution in the count process we know from (26) that $g(0) = (1 + \alpha\mu_i)^{1/\alpha}$. Following the same logic as for the ZIP, the zero-inflated Negative Binomial (ZINB) writes:

$$\mathbb{P}(Y_i = y_i|\pi_i, \mu_i, \alpha) = \pi_i \mathbf{1}_{y=0} + (1-\pi_i)\mathcal{NB}_{\mu_i,\alpha} \tag{28a}$$

$$(ZINB) \qquad = \begin{cases} \pi_i + (1-\pi_i)(1+\alpha\mu_i)^{-1/\alpha} & \text{if } y_i = 0 \\ (1-\pi_i)\frac{\Gamma(y_i+1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} & \text{if } y_i > 0 \end{cases} \tag{28b}$$

The model behind the selection equation can be a logit or probit. With a logit model, it is assumed that

$$\pi_i = \frac{\exp(Z_i\gamma)}{1 + \exp(Z_i\gamma)} \tag{29}$$

where γ is a set of parameters that needs to be estimated. Alternatively, for a probit model,

$$\pi_i = \Phi(Z_i\gamma) \tag{30}$$

The expression for the conditional mean is the same in both Poisson and NB models:

$$\mu_i = \exp(X_i\beta) \tag{31}$$

with β another set of parameters to be estimated. Both models will differ in the conditional variance $\mathbb{V}(Y_i|X_i)$ structure.

Log-likelihood Under a ZIP with a logit selection model, using (29) and (31), the log-likelihood can be written

$$
\begin{aligned}
(ZIP) \qquad l(\beta,\gamma|Y_i,X_i,Z_i) = &\sum_{i=1}^{n} \mathbf{1_{Y_i=0}} \log\Big(\exp(Z_i\gamma) + \exp(-\exp(X_i\beta))\Big) \\
&+ \sum_{i=1}^{n} \mathbf{1}_{Y_i>0}\Big[y_i X_i\beta - \exp(X_i\beta) - \log(Y_i!)\Big] \\
&- \sum_{i=1}^{n} \log(1+\exp(Z_i\gamma))
\end{aligned}
\tag{32}
$$

And the likelihood for ZINB writes as

$$
\begin{aligned}
(ZINB) \qquad l(\beta,\gamma,\alpha|Y_i,X_i,Z_i) = &\sum_{i=1}^{n} \mathbf{1}_{Y_i=0} \log\Big(\exp(Z_i\gamma) + (1+\alpha\mu_i)^{-1/\alpha}\Big) \\
&+ \sum_{i=1}^{n} \mathbf{1}_{Y_i>0} \sum_{j=1}^{Y_i-1} \log(j+1/\alpha) \\
&+ \sum_{i=1}^{n} \mathbf{1}_{Y_i>0}\Big[Y_i\Big(\log(\alpha)+X_i\beta\Big) \\
&\qquad - (Y_i+\alpha^{-1})\log\big(1+\alpha\exp(X_i\beta)\big)\Big] \\
&- \sum_{i=1}^{n} \log(1+\exp(Z_i\gamma))
\end{aligned}
\tag{33}
$$

The values that maximize log-likelihood are found using numerical methods. We have a CAN estimator in both cases. The second model is less parsimonious since it includes the additional dispersion parameter α.

Identification The identification of zero-inflated models is not frequently discussed. Among the notable exceptions, see Papadopoulos and Santos Silva (2012) and Staub and Winkelmann (2013). A zero-inflated model can be reformulated as a moment condition:

$$
\mathbb{E}(Y|X,Z) = \big(1-\pi(Z)\big)\lambda(X) = \frac{\exp(X\beta)}{1+\exp(Z\gamma)}
$$

β and γ (and possibly α in a ZINB model) can be estimated by a method of moments, avoiding distributional assumptions. However, as pinpointed by Papadopoulos and Santos Silva (2012), this expression raises identification issues. Assume variables entering the selection and outcome processes are the same, i.e. $X = Z$. In that case,

$$
\mathbb{E}(Y|X,Z) = \big(1-\pi(Z)\big)\lambda(X) = \frac{\exp(X\beta)}{1+\exp(X\gamma)}
\tag{34}
$$

It is possible to find two set of parameters, namely (γ,β) and $(-\gamma,\beta+\gamma)$, that yield the same value for the conditional mean. When taking an objective function derived from (34), the model cannot be identified without exclusion restrictions. Staub and Winkelmann (2013) propose sufficient conditions to restore identification as an exclusion restriction in the selection process (at least one excluded instrument) or a sign restriction on at least one parameter. The advantage of using the likelihood function (eq. (32) or (33)) is that the identification relies on the distributional assumptions (Papadopoulos and Santos Silva, 2008; Papadopoulos and Santos Silva, 2012). This is visible in the log-likelihood where the parameters β and γ enter separately.

C.2 Alternative specifications

Robustness checks are reported on Tables 12 to 17. Only the outcome equation is reported for zero-inflated models.

Zero-inflated models perform better if we consider log-likelihood criterion. However, ZIP models that are less parsimonious are strongly penalized by the BIC criterion compared to non-zero inflated models. The estimated value of the dispersion parameter in negative binomial models without selection suggests model drift (e.g. Table 15). The parameter for dispersion in the ZINB models appear more realistic.

For distance-decay parameters, introducing a selection model reduces the scale of the coefficients. Some of the spatial frictions are taken into account in the selection model by the inflated zero models. Moreover, the latter allow to account for the heterogeneity of travel costs that models without selection fail to explain. As expected, not accounting for zeros with a model estimated by OLS leads to biased estimates of spatial frictions. However, ignoring selection also leads to biased coefficients. ZIP and ZINB models yield travel costs that have approximately the same order of magnitudes.

Table 12 – Marseille ; Low-income: gravity model (robustness analysis, only outcome model reported)

	Dependent variable:						
	Non Zero Inflated			Zero Inflated			
	(OLS)	(Poisson)	(Negative Binomial)	(ZIP 1)	(ZIP 2)	(ZINB 1)	(Baseline)
Population in home cell (log)	0.540***	0.975***	1.416***	0.730***	0.722***	0.930***	0.728***
	(0.013)	(0.003)	(0.012)	(0.003)	(0.003)	(0.015)	(0.015)
Population in destination cell (log)	−0.009	−0.004***	0.010**	−0.012***	−0.024***	−0.022***	−0.016***
	(0.006)	(0.001)	(0.006)	(0.001)	(0.001)	(0.006)	(0.006)
Employment in home cell (log)	0.116***	0.220***	0.299***	0.190***	0.199***	0.131***	0.153***
	(0.006)	(0.001)	(0.007)	(0.001)	(0.001)	(0.006)	(0.007)
Employment in destination cell (log)	0.086***	0.145***	0.318***	0.118***	0.117***	0.099***	0.111***
	(0.005)	(0.001)	(0.005)	(0.001)	(0.001)	(0.005)	(0.005)
p_j^{D1} in destination cell (tax data)	0.075	0.096***	1.163***	0.239***	0.409***	0.347**	0.428***
	(0.125)	(0.023)	(0.138)	(0.023)	(0.023)	(0.145)	(0.148)
p_j^{D9} in destination cell (tax data)	−0.012	0.851***	−0.877***	0.795***	0.790***	0.343***	0.204**
	(0.088)	(0.017)	(0.091)	(0.017)	(0.017)	(0.096)	(0.094)
Distance (suburbs → suburbs)	−0.805***	−1.896***	−2.369***	−1.268***	−1.249***	−1.019***	−1.070***
	(0.017)	(0.004)	(0.014)	(0.005)	(0.005)	(0.018)	(0.017)
Distance (center → suburbs)	−0.904***	−1.941***	−2.497***	−1.132***	−1.109***	−1.172***	−1.121***
	(0.013)	(0.004)	(0.012)	(0.004)	(0.004)	(0.014)	(0.014)
Distance (suburbs → center)	−0.813***	−1.770***	−2.205***	−1.033***	−1.008***	−1.113***	−1.105***
	(0.014)	(0.006)	(0.013)	(0.007)	(0.007)	(0.016)	(0.016)
Distance (center → center)	−1.263***	−1.902***	−2.776***	−1.724***	−1.710***	−1.713***	−1.614***
	(0.014)	(0.003)	(0.017)	(0.003)	(0.003)	(0.016)	(0.016)
α (dispersion)			5.1			1.7	1.6
Count distribution	Gaussian	Poisson	Negative Binomial	Poisson	Poisson	Negative Binomial	Negative Binomial
Selection distribution				Probit	Logit	Probit	Logit
Observations	19,011	1,368,224	1,368,224	1,368,224	1,368,224	1,368,224	1,368,224
Log likelihood (by obs.)	−1.6	−0.2	−0.1	−0.2	−0.2	−0.1	−0.1
Bayesian information criterion	61,970	581,713	200,780	471,850	471,845	184,293	183,804

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

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

Table 13 – Marseille ; High-income: gravity model (robustness analysis, only outcome model reported)

	Dependent variable:						
	Non Zero Inflated			Zero Inflated			
	(OLS)	(Poisson)	(Negative Binomial)	(ZIP 1)	(ZIP 2)	(ZINB 1)	(Baseline)
Population in home cell (log)	0.557***	0.989***	1.455***	0.747***	0.747***	0.755***	0.723***
	(0.013)	(0.003)	(0.018)	(0.003)	(0.003)	(0.015)	(0.015)
Population in destination cell (log)	−0.002	0.007***	0.032***	−0.004***	−0.001	−0.004	−0.004
	(0.006)	(0.001)	(0.008)	(0.001)	(0.001)	(0.006)	(0.006)
Employment in home cell (log)	0.114***	0.210***	0.303***	0.175***	0.172***	0.145***	0.142***
	(0.006)	(0.001)	(0.010)	(0.001)	(0.001)	(0.007)	(0.007)
Employment in destination cell (log)	0.068***	0.138***	0.318***	0.104***	0.105***	0.083***	0.084***
	(0.005)	(0.001)	(0.007)	(0.001)	(0.001)	(0.005)	(0.005)
p_j^{D1} in destination cell (tax data)	0.336***	−0.087***	1.396***	0.071***	0.104***	0.311**	0.322**
	(0.129)	(0.023)	(0.209)	(0.024)	(0.024)	(0.145)	(0.145)
p_j^{D9} in destination cell (tax data)	−0.007	0.470***	−0.889***	0.337***	0.376***	0.001	−0.029
	(0.087)	(0.017)	(0.131)	(0.017)	(0.017)	(0.090)	(0.090)
Distance (suburbs $\rightarrow$ suburbs)	−0.800***	−1.784***	−2.467***	−1.081***	−1.074***	−1.054***	−1.023***
	(0.017)	(0.004)	(0.020)	(0.005)	(0.005)	(0.018)	(0.018)
Distance (center $\rightarrow$ suburbs)	−0.887***	−1.933***	−2.577***	−1.105***	−1.102***	−1.098***	−1.069***
	(0.013)	(0.004)	(0.018)	(0.004)	(0.004)	(0.014)	(0.013)
Distance (suburbs $\rightarrow$ center)	−0.802***	−1.758***	−2.285***	−1.005***	−0.999***	−1.058***	−1.033***
	(0.014)	(0.006)	(0.019)	(0.007)	(0.007)	(0.016)	(0.016)
Distance (center $\rightarrow$ center)	−1.270***	−1.876***	−2.855***	−1.679***	−1.672***	−1.660***	−1.611***
	(0.014)	(0.003)	(0.025)	(0.003)	(0.003)	(0.015)	(0.015)
α (dispersion)			10.4			1.5	1.5
Count distribution	Gaussian	Poisson	Negative Binomial	Poisson	Poisson	Negative Binomial	Negative Binomial
Selection distribution				Probit	Logit	Probit	Logit
Observations	18,844	1,333,298	1,333,298	1,333,298	1,333,298	1,333,298	1,333,298
Log likelihood (by obs.)	−1.6	−0.2	−0.1	−0.2	−0.2	−0.1	−0.1
Bayesian information criterion	61,426	557,776	204,068	439,997	439,470	183,183	182,764

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

Table 14 – Lyon ; Low-income: gravity model (robustness analysis, only outcome model reported)

	Dependent variable:						
	Non Zero Inflated			Zero Inflated			
	(OLS)	(Poisson)	(Negative Binomial)	(ZIP 1)	(ZIP 2)	(ZINB 1)	(Baseline)
Population in home cell (log)	0.547***	1.083***	1.299***	0.887***	0.869***	0.834***	0.766***
	(0.012)	(0.003)	(0.010)	(0.003)	(0.003)	(0.015)	(0.014)
Population in destination cell (log)	−0.004	−0.009***	−0.025***	−0.004***	−0.003***	−0.041***	−0.034***
	(0.005)	(0.001)	(0.004)	(0.001)	(0.001)	(0.005)	(0.005)
Employment in home cell (log)	0.090***	0.211***	0.172***	0.191***	0.193***	0.112***	0.098***
	(0.006)	(0.001)	(0.005)	(0.001)	(0.001)	(0.007)	(0.006)
Employment in destination cell (log)	0.057***	0.130***	0.167***	0.103***	0.105***	0.076***	0.087***
	(0.004)	(0.001)	(0.003)	(0.001)	(0.001)	(0.004)	(0.004)
p_j^{D1} in destination cell (tax data)	−0.465***	−0.075***	0.394***	−0.269***	−0.220***	−0.220	−0.479***
	(0.140)	(0.027)	(0.127)	(0.028)	(0.027)	(0.153)	(0.151)
p_j^{D9} in destination cell (tax data)	−0.006	0.532***	−0.030	0.518***	0.603***	0.709***	0.538***
	(0.079)	(0.016)	(0.068)	(0.016)	(0.016)	(0.090)	(0.087)
Distance (suburbs → suburbs)	−1.383***	−2.197***	−3.190***	−1.660***	−1.640***	−1.896***	−1.860***
	(0.016)	(0.003)	(0.012)	(0.004)	(0.004)	(0.020)	(0.020)
Distance (center → suburbs)	−1.209***	−2.023***	−2.609***	−1.709***	−1.700***	−1.548***	−1.525***
	(0.013)	(0.003)	(0.011)	(0.003)	(0.003)	(0.015)	(0.014)
Distance (suburbs → center)	−1.091***	−1.771***	−2.149***	−1.634***	−1.644***	−1.493***	−1.553***
	(0.013)	(0.004)	(0.011)	(0.005)	(0.005)	(0.017)	(0.016)
Distance (center → center)	−1.010***	−1.738***	−2.051***	−1.690***	−1.682***	−1.250***	−1.251***
	(0.017)	(0.003)	(0.015)	(0.003)	(0.003)	(0.016)	(0.017)
α (dispersion)			3.1			1.5	1.4
Count distribution	Gaussian	Poisson	Negative Binomial	Poisson	Poisson	Negative Binomial	Negative Binomial
Selection distribution				Probit	Logit	Probit	Logit
Observations	24,096	860,362	860,362	860,362	860,362	860,362	860,362
Log likelihood (by obs.)	−1.6	−0.4	−0.1	−0.3	−0.3	−0.1	−0.1
Bayesian information criterion	78,050	653,625	231,795	573,610	574,299	219,539	218,929

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

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

Table 15 – Lyon ; High-income: gravity model (robustness analysis, only outcome model reported)

	Dependent variable:						
	Non Zero Inflated			Zero Inflated			
	(OLS)	(Poisson)	(Negative Binomial)	(ZIP 1)	(ZIP 2)	(ZINB 1)	(Baseline)
Population in home cell (log)	0.576***	1.083***	1.642***	0.885***	0.908***	0.854***	0.896***
	(0.012)	(0.003)	(0.108)	(0.003)	(0.003)	(0.014)	(0.013)
Population in destination cell (log)	−0.007	−0.015***	0.071	−0.006***	−0.008***	−0.046***	−0.033***
	(0.005)	(0.001)	(0.064)	(0.001)	(0.001)	(0.005)	(0.005)
Employment in home cell (log)	0.085***	0.191***	0.050	0.173***	0.169***	0.095***	0.081***
	(0.006)	(0.001)	(0.065)	(0.001)	(0.001)	(0.006)	(0.006)
Employment in destination cell (log)	0.060***	0.152***	0.463***	0.124***	0.124***	0.091***	0.097***
	(0.004)	(0.001)	(0.054)	(0.001)	(0.001)	(0.004)	(0.004)
p_j^{D1} in destination cell (tax data)	−0.709***	−0.440***	5.876***	−0.687***	−0.765***	−0.777***	−0.392***
	(0.138)	(0.028)	(1.932)	(0.028)	(0.029)	(0.144)	(0.146)
p_j^{D9} in destination cell (tax data)	0.169**	0.824***	−0.952	0.849***	0.852***	0.747***	0.790***
	(0.079)	(0.016)	(1.082)	(0.016)	(0.016)	(0.087)	(0.088)
Distance (suburbs → suburbs)	−1.395***	−2.190***	−3.744***	−1.614***	−1.632***	−1.899***	−1.858***
	(0.016)	(0.003)	(0.163)	(0.004)	(0.004)	(0.020)	(0.020)
Distance (center → suburbs)	−1.227***	−1.976***	−2.860***	−1.653***	−1.663***	−1.569***	−1.530***
	(0.014)	(0.003)	(0.187)	(0.003)	(0.003)	(0.015)	(0.015)
Distance (suburbs → center)	−1.080***	−1.709***	−2.673***	−1.551***	−1.562***	−1.539***	−1.583***
	(0.013)	(0.004)	(0.178)	(0.005)	(0.005)	(0.016)	(0.018)
Distance (center → center)	−1.010***	−1.715***	−2.648***	−1.650***	−1.660***	−1.241***	−1.242***
	(0.017)	(0.003)	(0.332)	(0.003)	(0.003)	(0.017)	(0.017)
α (dispersion)			15117.2			1.5	1.5
Count distribution	Gaussian	Poisson	Negative Binomial	Poisson	Poisson	Negative Binomial	Negative Binomial
Selection distribution				Probit	Logit	Probit	Logit
Observations	24,308	849,734	849,734	849,734	849,734	849,734	849,734
Log likelihood (by obs.)	−1.6	−0.4	−0.3	−0.3	−0.3	−0.1	−0.1
Bayesian information criterion	78,779	634,583	551,979	548,534	547,806	219,861	219,668

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

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

Table 16 – Paris ; Low-income: gravity model (robustness analysis, only outcome model reported)

	Dependent variable:						
	Non Zero Inflated			Zero Inflated			
	(OLS)	(Poisson)	(Negative Binomial)	(ZIP 1)	(ZIP 2)	(ZINB 1)	(Baseline)
Population in home cell (log)	0.338***	0.923***	1.000***	0.806***	0.844***	0.764***	0.708***
	(0.003)	(0.001)	(0.003)	(0.001)	(0.001)	(0.005)	(0.005)
Population in destination cell (log)	−0.035***	−0.018***	−0.078***	−0.032***	−0.031***	−0.073***	−0.079***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in home cell (log)	0.032***	0.088***	0.126***	0.089***	0.084***	0.058***	0.061***
	(0.002)	(0.000)	(0.002)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in destination cell (log)	0.069***	0.125***	0.270***	0.093***	0.083***	0.120***	0.099***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.002)
p_j^{D1} in destination cell (tax data)	−0.762***	−0.830***	−0.363***	−1.082***	−1.054***	−1.029***	−1.038***
	(0.039)	(0.010)	(0.045)	(0.010)	(0.010)	(0.049)	(0.049)
p_j^{D9} in destination cell (tax data)	0.536***	0.955***	1.105***	0.948***	0.962***	0.815***	0.773***
	(0.023)	(0.005)	(0.025)	(0.005)	(0.005)	(0.031)	(0.031)
Destination cell belongs to cluster 1 (where high-income people are over-represented)	0.068***	0.171***	0.211***	0.202***	0.181***	0.087***	0.066***
	(0.007)	(0.002)	(0.009)	(0.002)	(0.002)	(0.009)	(0.009)
Destination cell belongs to cluster 2 (where low-income people are over-represented)	−0.118***	−0.085***	−0.299***	−0.064***	−0.085***	−0.226***	−0.224***
	(0.006)	(0.002)	(0.007)	(0.002)	(0.002)	(0.008)	(0.008)
Destination cell belongs to cluster 3 (where high-income mix with middle class)	0.134***	0.021***	−0.075***	0.074***	0.058***	0.066***	0.111***
	(0.008)	(0.002)	(0.008)	(0.002)	(0.002)	(0.010)	(0.010)
Distance (suburbs $\rightarrow$ suburbs)	−1.079***	−2.213***	−2.973***	−1.712***	−1.722***	−1.746***	−1.649***
	(0.003)	(0.001)	(0.004)	(0.001)	(0.001)	(0.004)	(0.004)
Distance (center $\rightarrow$ suburbs)	−0.956***	−1.934***	−2.430***	−1.643***	−1.653***	−1.461***	−1.377***
	(0.003)	(0.001)	(0.004)	(0.001)	(0.001)	(0.004)	(0.004)
Distance (suburbs $\rightarrow$ center)	−0.946***	−1.741***	−2.212***	−1.600***	−1.593***	−1.452***	−1.430***
	(0.003)	(0.001)	(0.004)	(0.001)	(0.001)	(0.004)	(0.004)
Distance (center $\rightarrow$ center)	−0.701***	−1.674***	−2.380***	−1.608***	−1.615***	−1.260***	−1.115***
	(0.006)	(0.001)	(0.008)	(0.001)	(0.001)	(0.007)	(0.007)
α (dispersion)			3.6			1.5	1.6
Count distribution	Gaussian	Poisson	Negative Binomial	Poisson	Poisson	Negative Binomial	Negative Binomial
Selection distribution				Probit	Logit	Probit	Logit
Observations	219,891	8,430,876	8,430,876	8,430,820	8,430,820	8,430,820	8,430,820
Log likelihood (by obs.)	−1.4	−0.3	−0.1	−0.3	−0.3	−0.1	−0.1
Bayesian information criterion	636,458	5,839,632	2,102,916	4,943,950	4,946,715	1,981,267	1,976,421

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

Table 17 – Paris ; High-income: gravity model (robustness analysis, only outcome model reported)

	Dependent variable:						
	Non Zero Inflated			Zero Inflated			
	(OLS)	(Poisson)	(Negative Binomial)	(ZIP 1)	(ZIP 2)	(ZINB 1)	(Baseline)
Population in home cell (log)	0.332***	0.902***	0.992***	0.795***	0.807***	0.751***	0.579***
	(0.003)	(0.001)	(0.004)	(0.001)	(0.001)	(0.005)	(0.005)
Population in destination cell (log)	−0.035***	−0.014***	−0.081***	−0.030***	−0.021***	−0.074***	−0.068***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in home cell (log)	0.031***	0.085***	0.124***	0.085***	0.088***	0.050***	0.061***
	(0.002)	(0.000)	(0.002)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in destination cell (log)	0.068***	0.119***	0.272***	0.082***	0.081***	0.116***	0.117***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.002)	(0.002)
p_j^{D1} in destination cell (tax data)	−0.760***	−0.953***	−0.428***	−1.114***	−1.143***	−1.058***	−1.101***
	(0.039)	(0.010)	(0.051)	(0.010)	(0.010)	(0.048)	(0.047)
p_j^{D9} in destination cell (tax data)	0.535***	0.899***	1.101***	0.880***	0.918***	0.783***	0.695***
	(0.023)	(0.005)	(0.028)	(0.005)	(0.005)	(0.031)	(0.030)
Destination cell belongs to cluster 1 (where high-income people are over-represented)	0.052***	0.160***	0.190***	0.170***	0.158***	0.074***	0.054***
	(0.007)	(0.002)	(0.010)	(0.002)	(0.002)	(0.009)	(0.009)
Destination cell belongs to cluster 2 (where low-income people are over-represented)	−0.123***	−0.066***	−0.302***	−0.040***	−0.050***	−0.237***	−0.270***
	(0.006)	(0.002)	(0.008)	(0.002)	(0.002)	(0.008)	(0.008)
Destination cell belongs to cluster 3 (where high-income mix with middle class)	0.116***	0.025***	−0.079***	0.082***	0.104***	0.067***	0.114***
	(0.008)	(0.002)	(0.009)	(0.002)	(0.002)	(0.010)	(0.010)
Distance (suburbs → suburbs)	−1.079***	−2.224***	−2.973***	−1.715***	−1.710***	−1.730***	−1.665***
	(0.003)	(0.001)	(0.004)	(0.001)	(0.001)	(0.004)	(0.004)
Distance (center → suburbs)	−0.961***	−1.922***	−2.439***	−1.629***	−1.638***	−1.465***	−1.371***
	(0.003)	(0.001)	(0.004)	(0.001)	(0.001)	(0.004)	(0.004)
Distance (suburbs → center)	−0.947***	−1.730***	−2.214***	−1.578***	−1.566***	−1.452***	−1.438***
	(0.003)	(0.001)	(0.004)	(0.001)	(0.001)	(0.004)	(0.004)
Distance (center → center)	−0.698***	−1.619***	−2.367***	−1.543***	−1.551***	−1.257***	−1.195***
	(0.006)	(0.001)	(0.009)	(0.001)	(0.001)	(0.007)	(0.007)
α (dispersion)			3.6			1.5	1.5
Count distribution	Gaussian	Poisson	Negative Binomial	Poisson	Poisson	Negative Binomial	Negative Binomial
Selection distribution				Probit	Logit	Probit	Logit
Observations	219,592	8,426,330	8,426,330	8,426,330	8,426,330	8,426,330	8,426,330
Log likelihood (by obs.)	−1.4	−0.3	−0.1	−0.3	−0.3	−0.1	−0.1
Bayesian information criterion	634,370	5,801,653	2,101,511	4,983,736	4,985,207	1,981,784	1,975,188

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

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

C.3 Regressions on the whole sample

Tables 18 and 19 (resp. Tables 20 and 21) present results on the whole sample for Marseille (resp. Lyon).

Distance related coefficients are robust to the choice of using a 5% sample rather than the whole sample. As expected, using a sample rather than all observations affects the confidence intervals that are larger. Using a 5% sample has no cost in terms of performance: log-likelihood by observation is the same when using the sample of the whole dataset. When using the whole sample, the negative binomial model that does not take into account zero-inflation diverges (dispersion parameter goes to an unrealistic level) while it converges with the 5% sample.

Table 18 – Marseille ; Low-income: gravity model on the whole (robustness analysis, only outcome model reported)

	Dependent variable:						
	Non Zero Inflated			Zero Inflated			
	(OLS)	(Poisson)	(Negative Binomial)	(ZIP 1)	(ZIP 2)	(ZINB 1)	(Baseline)
Population in home cell (log)	0.555***	1.004***	1.917***	0.754***	0.862***	0.863***	0.676***
	(0.004)	(0.001)	(0.155)	(0.001)	(0.001)	(0.005)	(0.005)
Population in destination cell (log)	0.011***	0.003***	0.373***	−0.012***	0.012***	0.015***	−0.019***
	(0.002)	(0.000)	(0.131)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in home cell (log)	0.121***	0.223***	0.196**	0.200***	0.164***	0.139***	0.176***
	(0.002)	(0.000)	(0.098)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in destination cell (log)	0.086***	0.146***	0.687***	0.125***	0.119***	0.100***	0.116***
	(0.002)	(0.000)	(0.097)	(0.000)	(0.000)	(0.002)	(0.002)
p_j^{D1} in destination cell (tax data)	0.044	−0.321***	1.003	−0.157***	−0.011	−0.077*	0.206***
	(0.039)	(0.007)	(2.987)	(0.007)	(0.007)	(0.041)	(0.043)
p_j^{D9} in destination cell (tax data)	0.071**	0.596***	−5.357***	0.593***	0.821***	0.314***	0.174***
	(0.028)	(0.005)	(1.552)	(0.005)	(0.005)	(0.028)	(0.028)
Distance (suburbs $\rightarrow$ suburbs)	−0.814***	−1.840***	−3.226***	−1.202***	−1.213***	−1.085***	−1.017***
	(0.005)	(0.001)	(0.267)	(0.002)	(0.002)	(0.006)	(0.006)
Distance (center $\rightarrow$ suburbs)	−0.907***	−1.918***	−3.229***	−1.145***	−1.191***	−1.171***	−1.126***
	(0.004)	(0.001)	(0.266)	(0.001)	(0.002)	(0.004)	(0.004)
Distance (suburbs $\rightarrow$ center)	−0.805***	−1.716***	−2.909***	−0.968***	−1.071***	−1.074***	−1.031***
	(0.005)	(0.002)	(0.258)	(0.002)	(0.003)	(0.005)	(0.005)
Distance (center $\rightarrow$ center)	−1.273***	−1.887***	−2.947***	−1.723***	−1.732***	−1.710***	−1.606***
	(0.005)	(0.001)	(0.295)	(0.001)	(0.001)	(0.005)	(0.005)
α (dispersion)			7e+05			1.5	1.5
Count distribution	Gaussian	Poisson	Negative Binomial	Poisson	Poisson	Negative Binomial	Negative Binomial
Selection distribution				Probit	Logit	Probit	Logit
Observations	177,600	11,158,754	11,158,754	11,158,754	11,158,754	11,503,616	11,503,616
Log likelihood (by obs.)	−1.6	−0.2	−0.2	−0.2	−0.2	−0.1	−0.1
Bayesian information criterion	576,271	5,302,083	5,418,424	4,345,402	4,360,021	1,703,947	1,699,324

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

Table 19 – Marseille ; High-income: gravity model on the whole (robustness analysis, only outcome model reported)

	Dependent variable:						
	Non Zero Inflated			Zero Inflated			
	(OLS)	(Poisson)	(Negative Binomial)	(ZIP 1)	(ZIP 2)	(ZINB 1)	(Baseline)
Population in home cell (log)	0.561***	1.011***	1.533***	0.789***	0.819***	0.897***	0.746***
	(0.004)	(0.001)	(0.037)	(0.001)	(0.001)	(0.005)	(0.005)
Population in destination cell (log)	0.010***	0.004***	−0.012	−0.021***	0.008***	0.006**	0.010***
	(0.002)	(0.000)	(0.023)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in home cell (log)	0.121***	0.222***	0.055***	0.196***	0.173***	0.134***	0.150***
	(0.002)	(0.000)	(0.023)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in destination cell (log)	0.084***	0.146***	0.492***	0.123***	0.118***	0.107***	0.096***
	(0.002)	(0.000)	(0.020)	(0.000)	(0.000)	(0.002)	(0.002)
p_j^{D1} in destination cell (tax data)	0.078**	−0.348***	6.687***	−0.124***	−0.171***	−0.112***	0.221***
	(0.039)	(0.007)	(0.712)	(0.007)	(0.007)	(0.042)	(0.043)
p_j^{D9} in destination cell (tax data)	0.046*	0.574***	1.887***	0.599***	0.613***	0.298***	0.112***
	(0.028)	(0.005)	(0.380)	(0.005)	(0.005)	(0.029)	(0.028)
Distance (suburbs $\rightarrow$ suburbs)	−0.813***	−1.844***	−3.794***	−1.218***	−1.188***	−1.080***	−1.071***
	(0.005)	(0.001)	(0.060)	(0.002)	(0.002)	(0.006)	(0.006)
Distance (center $\rightarrow$ suburbs)	−0.902***	−1.913***	−2.853***	−1.143***	−1.138***	−1.172***	−1.120***
	(0.004)	(0.001)	(0.069)	(0.001)	(0.001)	(0.004)	(0.004)
Distance (suburbs $\rightarrow$ center)	−0.803***	−1.712***	−2.848***	−0.992***	−1.043***	−1.074***	−1.052***
	(0.005)	(0.002)	(0.067)	(0.002)	(0.002)	(0.005)	(0.005)
Distance (center $\rightarrow$ center)	−1.275***	−1.892***	−2.621***	−1.773***	−1.737***	−1.704***	−1.599***
	(0.005)	(0.001)	(0.122)	(0.001)	(0.001)	(0.005)	(0.005)
α (dispersion)			25975.8			1.5	1.5
Count distribution	Gaussian	Poisson	Negative Binomial	Poisson	Poisson	Negative Binomial	Negative Binomial
Selection distribution				Probit	Logit	Probit	Logit
Observations	177,384	11,503,504	8,497,345	11,158,642	11,158,642	11,158,642	11,158,642
Log likelihood (by obs.)	−1.6	−0.2	−0.3	−0.2	−0.2	−0.1	−0.1
Bayesian information criterion	574,190	5,297,301	5,773,262	4,344,740	4,337,240	1,701,410	1,696,969

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

Table 20 – Lyon ; Low-income: gravity model on the whole (robustness analysis, only outcome model reported)

	Dependent variable:						
	Non Zero Inflated			Zero Inflated			
	(OLS)	(Poisson)	(Negative Binomial)	(ZIP 1)	(ZIP 2)	(ZINB 1)	(Baseline)
Population in home cell (log)	0.564***	1.085***	1.505***	0.897***	0.988***	0.888***	0.647***
	(0.004)	(0.001)	(0.037)	(0.001)	(0.001)	(0.005)	(0.004)
Population in destination cell (log)	−0.006***	−0.018***	−0.025	0.007***	−0.021***	−0.044***	−0.039***
	(0.002)	(0.000)	(0.023)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in home cell (log)	0.086***	0.204***	0.077***	0.199***	0.163***	0.093***	0.127***
	(0.002)	(0.000)	(0.023)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in destination cell (log)	0.061***	0.145***	0.511***	0.097***	0.118***	0.068***	0.094***
	(0.001)	(0.000)	(0.020)	(0.000)	(0.000)	(0.001)	(0.001)
p_j^{D1} in destination cell (tax data)	−0.564***	−0.226***	6.932***	−0.425***	−0.344***	−0.303***	−0.315***
	(0.044)	(0.009)	(0.710)	(0.009)	(0.009)	(0.048)	(0.047)
p_j^{D9} in destination cell (tax data)	0.056**	0.686***	2.112***	0.825***	0.777***	0.622***	0.520***
	(0.025)	(0.005)	(0.372)	(0.005)	(0.005)	(0.028)	(0.028)
Distance (suburbs $\rightarrow$ suburbs)	−1.383***	−2.190***	−3.803***	−1.600***	−1.638***	−1.886***	−1.804***
	(0.005)	(0.001)	(0.059)	(0.001)	(0.001)	(0.006)	(0.006)
Distance (center $\rightarrow$ suburbs)	−1.221***	−2.002***	−2.841***	−1.749***	−1.695***	−1.548***	−1.529***
	(0.004)	(0.001)	(0.068)	(0.001)	(0.001)	(0.005)	(0.004)
Distance (suburbs $\rightarrow$ center)	−1.091***	−1.762***	−2.852***	−1.616***	−1.619***	−1.547***	−1.608***
	(0.004)	(0.001)	(0.066)	(0.002)	(0.002)	(0.005)	(0.005)
Distance (center $\rightarrow$ center)	−1.001***	−1.747***	−2.598***	−1.661***	−1.707***	−1.228***	−1.227***
	(0.005)	(0.001)	(0.118)	(0.001)	(0.001)	(0.005)	(0.005)
α (dispersion)			26087.280			1.488	1.335
Count distribution	Gaussian	Poisson	Negative Binomial	Poisson	Poisson	Negative Binomial	Negative Binomial
Selection distribution				Probit	Logit	Probit	Logit
Observations	242,035	8,603,619	8,603,619	8,603,619	8,603,619	8,603,619	8,603,619
Log likelihood (by obs.)	−1.619	−0.373	−0.335	−0.327	−0.324	−0.128	−0.127
Bayesian information criterion	784,000	6,413,821	5,763,064	5,623,462	5,569,013	2,197,213	2,194,066

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

Table 21 – Lyon ; High-income: gravity model on the whole (robustness analysis, only outcome model reported)

	Dependent variable:						
	Non Zero Inflated			Zero Inflated			
	(OLS)	(Poisson)	(Negative Binomial)	(ZIP 1)	(ZIP 2)	(ZINB 1)	(Baseline)
Population in home cell (log)	0.567***	1.088***	1.533***	0.918***	0.988***	0.883***	0.795***
	(0.004)	(0.001)	(0.037)	(0.001)	(0.001)	(0.005)	(0.004)
Population in destination cell (log)	−0.003**	−0.017***	−0.012	−0.017***	−0.020***	−0.041***	−0.043***
	(0.002)	(0.000)	(0.023)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in home cell (log)	0.089***	0.207***	0.055***	0.182***	0.170***	0.087***	0.097***
	(0.002)	(0.000)	(0.023)	(0.000)	(0.000)	(0.002)	(0.002)
Employment in destination cell (log)	0.060***	0.144***	0.492***	0.121***	0.115***	0.091***	0.085***
	(0.001)	(0.000)	(0.020)	(0.000)	(0.000)	(0.001)	(0.001)
p_j^{D1} in destination cell (tax data)	−0.629***	−0.290***	6.687***	−0.486***	−0.422***	−0.371***	−0.358***
	(0.044)	(0.009)	(0.712)	(0.009)	(0.009)	(0.047)	(0.045)
p_j^{D9} in destination cell (tax data)	0.053**	0.666***	1.887***	0.785***	0.748***	0.598***	0.538***
	(0.025)	(0.005)	(0.380)	(0.005)	(0.005)	(0.027)	(0.027)
Distance (suburbs → suburbs)	−1.385***	−2.197***	−3.794***	−1.639***	−1.648***	−1.950***	−1.914***
	(0.005)	(0.001)	(0.060)	(0.001)	(0.001)	(0.006)	(0.006)
Distance (center → suburbs)	−1.222***	−2.006***	−2.853***	−1.712***	−1.703***	−1.590***	−1.553***
	(0.004)	(0.001)	(0.069)	(0.001)	(0.001)	(0.005)	(0.004)
Distance (suburbs → center)	−1.086***	−1.755***	−2.848***	−1.624***	−1.616***	−1.562***	−1.526***
	(0.004)	(0.001)	(0.067)	(0.002)	(0.002)	(0.005)	(0.005)
Distance (center → center)	−1.001***	−1.750***	−2.621***	−1.701***	−1.708***	−1.269***	−1.221***
	(0.005)	(0.001)	(0.122)	(0.001)	(0.001)	(0.005)	(0.005)
α (dispersion)			25975.805			1.467	1.337
Count distribution	Gaussian	Poisson	Negative Binomial	Poisson	Poisson	Negative Binomial	Negative Binomial
Selection distribution				Probit	Logit	Probit	Logit
Observations	242,621	8,497,345	8,497,345	8,497,345	8,497,345	8,497,345	8,497,345
Log likelihood (by obs.)	−1.621	−0.377	−0.340	−0.328	−0.328	−0.129	−0.129
Bayesian information criterion	786,926	6,413,224	5,773,262	5,574,688	5,578,000	2,199,086	2,195,426

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

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

Liste des documents de travail de la Direction des Études et Synthèses Économiques

G 9001	J. FAYOLLE et M. FLEURBAEY Accumulation, profitabilité et endettement des entreprises
G 9002	H. ROUSSE Détection et effets de la multicolinéarité dans les modèles linéaires ordinaires - Un prolongement de la réflexion de BELSLEY, KUH et WELSCH
G 9003	P. RALLE et J. TOUJAS-BERNATE Indexation des salaires : la rupture de 1983
G 9004	D. GUELLEC et P. RALLE Compétitivité, croissance et innovation de produit
G 9005	P. RALLE et J. TOUJAS-BERNATE Les conséquences de la désindexation. Analyse dans une maquette prix-salaires
G 9101	Équipe AMADEUS Le modèle AMADEUS - Première partie -Présentation générale
G 9102	J.L. BRILLET Le modèle AMADEUS - Deuxième partie -Propriétés variantielles
G 9103	D. GUELLEC et P. RALLE Endogenous growth and product innovation
G 9104	H. ROUSSE Le modèle AMADEUS - Troisième partie - Le commerce extérieur et l'environnement international
G 9105	H. ROUSSE Effets de demande et d'offre dans les résultats du commerce extérieur manufacturé de la France au cours des deux dernières décennies
G 9106	B. CREPON Innovation, taille et concentration : causalités et dynamiques
G 9107	B. AMABLE et D. GUELLEC Un panorama des théories de la croissance endogène
G 9108	M. GLAUDE et M. MOUTARDIER Une évaluation du coût direct de l'enfant de 1979 à 1989
G 9109	P. RALLE et alii France - Allemagne : performances économiques comparées
G 9110	J.L. BRILLET Micro-DMS ***NON PARU***
G 9111	A. MAGNIER Effets accélérateur et multiplicateur en France depuis 1970 : quelques résultats empiriques
G 9112	B. CREPON et G. DUREAU Investissement en recherche-développement : analyse de causalités dans un modèle d'accélérateur généralisé
G 9113	J.L. BRILLET, H. ERKEL-ROUSSE, J. TOUJAS-BERNATE "France-Allemagne Couplées" - Deux économies vues par une maquette macro-économétrique
G 9201	W.J. ADAMS, B. CREPON, D. ENCAOUA Choix technologiques et stratégies de dissuasion d'entrée
G 9202	J. OLIVEIRA-MARTINS, J. TOUJAS-BERNATE Macro-economic import functions with imperfect competition - An application to the E.C. Trade
G 9203	I. STAPIC Les échanges internationaux de services de la France dans le cadre des négociations multilatérales du GATT Juin 1992 (1ère version) Novembre 1992 (version finale)
G 9204	P. SEVESTRE L'économétrie sur données individuelles-temporelles. Une note introductive
G 9205	H. ERKEL-ROUSSE Le commerce extérieur et l'environnement international dans le modèle AMADEUS (réestimation 1992)
G 9206	N. GREENAN et D. GUELLEC Coordination within the firm and endogenous growth
G 9207	A. MAGNIER et J. TOUJAS-BERNATE Technology and trade: empirical evidences for the major five industrialized countries
G 9208	B. CREPON, E. DUGUET, D. ENCAOUA et P. MOHNEN Cooperative, non cooperative R & D and optimal patent life
G 9209	B. CREPON et E. DUGUET Research and development, competition and innovation: an application of pseudo maximum likelihood methods to Poisson models with heterogeneity
G 9301	J. TOUJAS-BERNATE Commerce international et concurrence imparfaite : développements récents et implications pour la politique commerciale
G 9302	Ch. CASES Durées de chômage et comportements d'offre de travail : une revue de la littérature
G 9303	H. ERKEL-ROUSSE Union économique et monétaire : le débat économique
G 9304	N. GREENAN - D. GUELLEC / G. BROUSSAUDIER - L. MIOTTI Innovation organisationnelle, dynamisme technologique et performances des entreprises
G 9305	P. JAILLARD Le traité de Maastricht : présentation juridique et historique
G 9306	J.L. BRILLET Micro-DMS : présentation et propriétés
G 9307	J.L. BRILLET Micro-DMS - variantes : les tableaux
G 9308	S. JACOBZONE Les grands réseaux publics français dans une perspective européenne
G 9309	L. BLOCH - B. CŒURÉ Profitabilité de l'investissement productif et transmission des chocs financiers
G 9310	J. BOURDIEU - B. COLIN-SEDILLOT Les théories sur la structure optimale du capital : quelques points de repère
G 9311	J. BOURDIEU - B. COLIN-SEDILLOT Les décisions de financement des entreprises françaises : une évaluation empirique des théories de la structure optimale du capital
G 9312	L. BLOCH - B. CŒURÉ Q de Tobin marginal et transmission des chocs financiers
G 9313	Équipes Amadeus (INSEE), Banque de France, Métric (DP) Présentation des propriétés des principaux modèles macroéconomiques du Service Public
G 9314	B. CREPON - E. DUGUET Research & Development, competition and innovation
G 9315	B. DORMONT Quelle est l'influence du coût du travail sur l'emploi ?
G 9316	D. BLANCHET - C. BROUSSE Deux études sur l'âge de la retraite
G 9317	D. BLANCHET Répartition du travail dans une population hétérogène : deux notes
G 9318	D. EYSSARTIER - N. PONTY AMADEUS - an annual macro-economic model for the medium and long term
G 9319	G. CETTE - Ph. CUNÉO - D. EYSSARTIER -J. GAUTIÉ Les effets sur l'emploi d'un abaissement du coût du travail des jeunes
G 9401	D. BLANCHET Les structures par âge importent-elles ?
G 9402	J. GAUTIÉ Le chômage des jeunes en France : problème de formation ou phénomène de file d'attente ? Quelques éléments du débat
G 9403	P. QUIRION Les déchets en France : éléments statistiques et économiques
G 9404	D. LADIRAY - M. GRUN-REHOMME Lissage par moyennes mobiles - Le problème des extrémités de série
G 9405	V. MAILLARD Théorie et pratique de la correction des effets de jours ouvrables
G 9406	F. ROSENWALD La décision d'investir
G 9407	S. JACOBZONE Les apports de l'économie industrielle pour définir la stratégie économique de l'hôpital public
G 9408	L. BLOCH, J. BOURDIEU, B. COLIN-SEDILLOT, G. LONGUEVILLE Du défaut de paiement au dépôt de bilan : les banquiers face aux PME en difficulté
G 9409	D. EYSSARTIER, P. MAIRE Impacts macro-économiques de mesures d'aide au logement - quelques éléments d'évaluation
G 9410	F. ROSENWALD Suivi conjoncturel de l'investissement
G 9411	C. DEFEUILLEY - Ph. QUIRION Les déchets d'emballages ménagers : une analyse économique des politiques française et allemande
G 9412	J. BOURDIEU - B. CŒURÉ - B. COLIN-SEDILLOT Investissement, incertitude et irréversibilité Quelques développements récents de la théorie de l'investissement
G 9413	B. DORMONT - M. PAUCHET L'évaluation de l'élasticité emploi-salaire dépend-elle des structures de qualification ?
G 9414	I. KABLA Le Choix de breveter une invention
G 9501	J. BOURDIEU - B. CŒURÉ - B. SEDILLOT Irreversible Investment and Uncertainty: When is there a Value of Waiting?
G 9502	L. BLOCH - B. CŒURÉ Imperfections du marché du crédit, investissement des entreprises et cycle économique
G 9503	D. GOUX - E. MAURIN Les transformations de la demande de travail par qualification en France Une étude sur la période 1970-1993
G 9504	N. GREENAN Technologie, changement organisationnel, qualifications et emploi : une étude empirique sur l'industrie manufacturière
G 9505	D. GOUX - E. MAURIN Persistance des hiérarchies sectorielles de salaires: un réexamen sur données françaises
G 9505 Bis	D. GOUX - E. MAURIN Persistence of inter-industry wages differentials: a reexamination on matched worker-firm panel data
G 9506	S. JACOBZONE Les liens entre RMI et chômage, une mise en perspective *NON PARU - article sorti dans Économie et Prévision n° 122 (1996) - pages 95 à 113*
G 9507	G. CETTE - S. MAHFOUZ Le partage primaire du revenu Constat descriptif sur longue période
G 9601	Banque de France - CEPREMAP - Direction de la Prévision - Érasme - INSEE - OFCE Structures et propriétés de cinq modèles macro-économiques français
G 9602	Rapport d'activité de la DESE de l'année 1995
G 9603	J. BOURDIEU - A. DRAZNIEKS L'octroi de crédit aux PME : une analyse à partir d'informations bancaires
G 9604	A. TOPIOL-BENSAÏD Les implantations japonaises en France
G 9605	P. GENIER - S. JACOBZONE Comportements de prévention, consommation d'alcool et tabagie : peut-on parler d'une gestion globale du capital santé ? *Une modélisation microéconométrique empirique*

G 9606 C. DOZ - F. LENGLART
Factor analysis and unobserved component models: an application to the study of French business surveys

G 9607 N. GREENAN - D. GUELLEC
La théorie coopérative de la firme

G 9608 N. GREENAN - D. GUELLEC
Technological innovation and employment reallocation

G 9609 Ph. COUR - F. RUPPRECHT
L'intégration asymétrique au sein du continent américain : un essai de modélisation

G 9610 S. DUCHENE - G. FORGEOT - A. JACQUOT
Analyse des évolutions récentes de la productivité apparente du travail

G 9611 X. BONNET - S. MAHFOUZ
The influence of different specifications of wages-prices spirals on the measure of the NAIRU: the case of France

G 9612 PH. COUR - E. DUBOIS, S. MAHFOUZ, J. PISANI-FERRY
The cost of fiscal retrenchment revisited: how strong is the evidence?

G 9613 A. JACQUOT
Les flexions des taux d'activité sont-elles seulement conjoncturelles ?

G 9614 ZHANG Yingxiang - SONG Xueqing
Lexique macroéconomique Français-Chinois

G 9701 J.L. SCHNEIDER
La taxe professionnelle : éléments de cadrage économique

G 9702 J.L. SCHNEIDER
Transition et stabilité politique d'un système redistributif

G 9703 D. GOUX - E. MAURIN
Train or Pay: Does it Reduce Inequalities to Encourage Firms to Train their Workers?

G 9704 P. GENIER
Deux contributions sur dépendance et équité

G 9705 E. DUGUET - N. IUNG
R & D Investment, Patent Life and Patent Value
An Econometric Analysis at the Firm Level

G 9706 M. HOUDEBINE - A. TOPIOL-BENSAÏD
Les entreprises internationales en France : une analyse à partir de données individuelles

G 9707 M. HOUDEBINE
Polarisation des activités et spécialisation des départements en France

G 9708 E. DUGUET - N. GREENAN
Le biais technologique : une analyse sur données individuelles

G 9709 J.L. BRILLET
Analyzing a small French ECM Model

G 9710 J.L. BRILLET
Formalizing the transition process: scenarios for capital accumulation

G 9711 G. FORGEOT - J. GAUTIÉ
Insertion professionnelle des jeunes et processus de déclassement

G 9712 E. DUBOIS
High Real Interest Rates: the Consequence of a Saving Investment Disequilibrium or of an insufficient Credibility of Monetary Authorities?

G 9713 Bilan des activités de la Direction des Études et Synthèses Économiques - 1996

G 9714 F. LEQUILLER
Does the French Consumer Price Index Overstate Inflation?

G 9715 X. BONNET
Peut-on mettre en évidence les rigidités à la baisse des salaires nominaux ?
Une étude sur quelques grands pays de l'OCDE

G 9716 N. IUNG - F. RUPPRECHT
Productivité de la recherche et rendements d'échelle dans le secteur pharmaceutique français

G 9717 E. DUGUET - I. KABLA
Appropriation strategy and the motivations to use the patent system in France - An econometric analysis at the firm level

G 9718 L.P. PELÉ - P. RALLE
Âge de la retraite : les aspects incitatifs du régime général

G 9719 ZHANG Yingxiang - SONG Xueqing
Lexique macroéconomique français-chinois, chinois-français

G 9720 M. HOUDEBINE - J.L. SCHNEIDER
Mesurer l'influence de la fiscalité sur la localisation des entreprises

G 9721 A. MOUROUGANE
Crédibilité, indépendance et politique monétaire
Une revue de la littérature

G 9722 P. AUGERAUD - L. BRIOT
Les données comptables d'entreprises
Le système intermédiaire d'entreprises
Passage des données individuelles aux données sectorielles

G 9723 P. AUGERAUD - J.E. CHAPRON
Using Business Accounts for Compiling National Accounts: the French Experience

G 9724 P. AUGERAUD
Les comptes d'entreprise par activités - Le passage aux comptes - De la comptabilité d'entreprise à la comptabilité nationale - *A paraître*

G 9801 H. MICHAUDON - C. PRIGENT
Présentation du modèle AMADEUS

G 9802 J. ACCARDO
Une étude de comptabilité générationnelle pour la France en 1996

G 9803 X. BONNET - S. DUCHÊNE
Apports et limites de la modélisation « Real Business Cycles »

G 9804 C. BARLET - C. DUGUET - D. ENCAOUA - J. PRADEL
The Commercial Success of Innovations
An econometric analysis at the firm level in French manufacturing

G 9805 P. CAHUC - Ch. GIANELLA - D. GOUX - A. ZILBERBERG
Equalizing Wage Differences and Bargaining Power - Evidence form a Panel of French Firms

G 9806 J. ACCARDO - M. JLASSI
La productivité globale des facteurs entre 1975 et 1996

G 9807 Bilan des activités de la Direction des Études et Synthèses Économiques - 1997

G 9808 A. MOUROUGANE
Can a Conservative Governor Conduct an Accomodative Monetary Policy?

G 9809 X. BONNET - E. DUBOIS - L. FAUVET
Asymétrie des inflations relatives et menus costs : tests sur l'inflation française

G 9810 E. DUGUET - N. IUNG
Sales and Advertising with Spillovers at the firm level: Estimation of a Dynamic Structural Model on Panel Data

G 9811 J.P. BERTHIER
Congestion urbaine : un modèle de trafic de pointe à courbe débit-vitesse et demande élastique

G 9812 C. PRIGENT
La part des salaires dans la valeur ajoutée : une approche macroéconomique

G 9813 A.Th. AERTS
L'évolution de la part des salaires dans la valeur ajoutée en France reflète-t-elle les évolutions individuelles sur la période 1979-1994 ?

G 9814 B. SALANIÉ
Guide pratique des séries non-stationnaires

G 9901 S. DUCHÊNE - A. JACQUOT
Une croissance plus riche en emplois depuis le début de la décennie ? Une analyse en comparaison internationale

G 9902 Ch. COLIN
Modélisation des carrières dans Destinie

G 9903 Ch. COLIN
Évolution de la dispersion des salaires : un essai de prospective par microsimulation

G 9904 B. CREPON - N. IUNG
Innovation, emploi et performances

G 9905 B. CREPON - Ch. GIANELLA
Wages inequalities in France 1969-1992
An application of quantile regression techniques

G 9906 C. BONNET - R. MAHIEU
Microsimulation techniques applied to intergenerational transfers - Pensions in a dynamic framework: the case of France

G 9907 F. ROSENWALD
L'impact des contraintes financières dans la décision d'investissement

G 9908 Bilan des activités de la DESE - 1998

G 9909 J.P. ZOYEM
Contrat d'insertion et sortie du RMI
Évaluation des effets d'une politique sociale

G 9910 Ch. COLIN - Fl. LEGROS - R. MAHIEU
Bilans contributifs comparés des régimes de retraite du secteur privé et de la fonction publique

G 9911 G. LAROQUE - B. SALANIÉ
Une décomposition du non-emploi en France

G 9912 B. SALANIÉ
Une maquette analytique de long terme du marché du travail

G 9912 Bis Ch. GIANELLA
Une estimation de l'élasticité de l'emploi peu qualifié à son coût

G 9913 Division « Redistribution et Politiques Sociales »
Le modèle de microsimulation dynamique DESTINIE

G 9914 E. DUGUET
Macro-commandes SAS pour l'économétrie des panels et des variables qualitatives

G 9915 R. DUHAUTOIS
Évolution des flux d'emplois en France entre 1990 et 1996 : une étude empirique à partir du fichier des bénéfices réels normaux (BRN)

G 9916 J.Y. FOURNIER
Extraction du cycle des affaires : la méthode de Baxter et King

G 9917 B. CRÉPON - R. DESPLATZ - J. MAIRESSE
Estimating price cost margins, scale economies and workers' bargaining power at the firm level

G 9918 Ch. GIANELLA - Ph. LAGARDE
Productivity of hours in the aggregate production function: an evaluation on a panel of French firms from the manufacturing sector

G 9919 S. AUDRIC - P. GIVORD - C. PROST
Évolution de l'emploi et des coûts par qualification entre 1982 et 1996

G 2000/01 R. MAHIEU
Les déterminants des dépenses de santé : une approche macroéconomique

G 2000/02 C. ALLARD-PRIGENT - H. GUILMEAU - A. QUINET
The real exchange rate as the relative price of nontrables in terms of tradables: theoretical investigation and empirical study on French data

G 2000/03 J.-Y. FOURNIER
L'approximation du filtre passe-bande proposée par Christiano et Fitzgerald

G 2000/04 Bilan des activités de la DESE - 1999

G 2000/05 B. CREPON - F. ROSENWALD
Investissement et contraintes de financement : le poids du cycle
Une estimation sur données françaises

G 2000/06 A. FLIPO
Les comportements matrimoniaux de fait

G 2000/07 R. MAHIEU - B. SÉDILLOT
Microsimulations of the retirement decision: a supply side approach

G 2000/08 C. AUDENIS - C. PROST
Déficit conjoncturel : une prise en compte des conjonctures passées

G 2000/09 R. MAHIEU - B. SÉDILLOT
Équivalent patrimonial de la rente et souscription de retraite complémentaire

G 2000/10 R. DUHAUTOIS
Ralentissement de l'investissement : petites ou grandes entreprises ? industrie ou tertiaire ?

G 2000/11 G. LAROQUE - B. SALANIÉ
Temps partiel féminin et incitations financières à l'emploi

G2000/12 Ch. GIANELLA
Local unemployment and wages

G2000/13 B. CREPON - Th. HECKEL
- Informatisation en France : une évaluation à partir de données individuelles
- Computerization in France: an evaluation based on individual company data

G2001/01 F. LEQUILLER
- La nouvelle économie et la mesure de la croissance du PIB
- The new economy and the measure ment of GDP growth

G2001/02 S. AUDRIC
La reprise de la croissance de l'emploi profite-t-elle aussi aux non-diplômés ?

G2001/03 I. BRAUN-LEMAIRE
Évolution et répartition du surplus de productivité

G2001/04 A. BEAUDU - Th. HECKEL
Le canal du crédit fonctionne-t-il en Europe ? Une étude de l'hétérogénéité des comportements d'investissement à partir de données de bilan agrégées

G2001/05 C. AUDENIS - P. BISCOURP - N. FOURCADE - O. LOISEL
Testing the augmented Solow growth model: An empirical reassessment using panel data

G2001/06 R. MAHIEU - B. SÉDILLOT
Départ à la retraite, irréversibilité et incertitude

G2001/07 Bilan des activités de la DESE - 2000

G2001/08 J. Ph. GAUDEMET
Les dispositifs d'acquisition à titre facultatif d'annuités viagères de retraite

G2001/09 B. CRÉPON - Ch. GIANELLA
Fiscalité, coût d'usage du capital et demande de facteurs : une analyse sur données individuelles

G2001/10 B. CRÉPON - R. DESPLATZ
Évaluation des effets des dispositifs d'allégements de charges sociales sur les bas salaires

G2001/11 J.-Y. FOURNIER
Comparaison des salaires des secteurs public et privé

G2001/12 J.-P. BERTHIER - C. JAULENT - R. CONVENEVOLE - S. PISANI
Une méthodologie de comparaison entre consommations intermédiaires de source fiscale et de comptabilité nationale

G2001/13 P. BISCOURP - Ch. GIANELLA
Substitution and complementarity between capital, skilled and less skilled workers: an analysis at the firm level in the French manufacturing industry

G2001/14 I. ROBERT-BOBEE
Modelling demographic behaviours in the French microsimulation model Destinie: An analysis of future change in completed fertility

G2001/15 J.-P. ZOYEM
Diagnostic sur la pauvreté et calendrier de revenus : le cas du "Panel européen des ménages »

G2001/16 J.-Y. FOURNIER - P. GIVORD
La réduction des taux d'activité aux âges extrêmes, une spécificité française ?

G2001/17 C. AUDENIS - P. BISCOURP - N. RIEDINGER
Existe-t-il une asymétrie dans la transmission du prix du brut aux prix des carburants ?

G2002/01 F. MAGNIEN - J.-L. TAVERNIER - D. THESMAR
Les statistiques internationales de PIB par habitant en standard de pouvoir d'achat : une analyse des résultats

G2002/02 Bilan des activités de la DESE - 2001

G2002/03 B. SÉDILLOT - E. WALRAET
La cessation d'activité au sein des couples : y a-t-il interdépendance des choix ?

G2002/04 G. BRILHAULT
- Rétropolation des séries de FBCF et calcul du capital fixe en SEC-95 dans les comptes nationaux français
- Retropolation of the investment series (GFCF) and estimation of fixed capital stocks on the ESA-95 basis for the French balance sheets

G2002/05 P. BISCOURP - B. CRÉPON - T. HECKEL - N. RIEDINGER
How do firms respond to cheaper computers? Microeconometric evidence for France based on a production function approach

G2002/06 C. AUDENIS - J. DEROYON - N. FOURCADE
L'impact des nouvelles technologies de l'information et de la communication sur l'économie française - un bouclage macro-économique

G2002/07 J. BARDAJI - B. SÉDILLOT - E. WALRAET
Évaluation de trois réformes du Régime Général d'assurance vieillesse à l'aide du modèle de microsimulation DESTINIE

G2002/08 J.-P. BERTHIER
Réflexions sur les différentes notions de volume dans les comptes nationaux : comptes aux prix d'une année fixe ou aux prix de l'année précédente, séries chaînées

G2002/09 F. HILD
Les soldes d'opinion résument-ils au mieux les réponses des entreprises aux enquêtes de conjoncture ?

G2002/10 I. ROBERT-BOBÉE
Les comportements démographiques dans le modèle de microsimulation Destinie - Une comparaison des estimations issues des enquêtes Jeunes et Carrières 1997 et Histoire Familiale 1999

G2002/11 J.-P. ZOYEM
La dynamique des bas revenus : une analyse des entrées-sorties de pauvreté

G2002/12 F. HILD
Prévisions d'inflation pour la France

G2002/13 M. LECLAIR
Réduction du temps de travail et tensions sur les facteurs de production

G2002/14 E. WALRAET - A. VINCENT
- Analyse de la redistribution intragénérationnelle dans le système de retraite des salariés du privé - Une approche par microsimulation
- Intragenerational distributional analysis in the french private sector pension scheme - A microsimulation approach

G2002/15 P. CHONE - D. LE BLANC - I. ROBERT-BOBEE
Offre de travail féminine et garde des jeunes enfants

G2002/16 F. MAUREL - S. GREGOIR
Les indices de compétitivité des pays : interprétation et limites

G2003/01 N. RIEDINGER - E.HAUVY
Le coût de dépollution atmosphérique pour les entreprises françaises : Une estimation à partir de données individuelles

G2003/02 P. BISCOURP et F. KRAMARZ
Création d'emplois, destruction d'emplois et internationalisation des entreprises industrielles françaises : une analyse sur la période 1986-1992

G2003/03 Bilan des activités de la DESE - 2002

G2003/04 P.-O. BEFFY - J. DEROYON - N. FOURCADE - S. GREGOIR - N. LAÏB - B. MONFORT
Évolutions démographiques et croissance : une projection macro-économique à l'horizon 2020

G2003/05 P. AUBERT
La situation des salariés de plus de cinquante ans dans le secteur privé

G2003/06 P. AUBERT - B. CRÉPON
Age, salaire et productivité
La productivité des salariés décline-t-elle en fin de carrière ?

G2003/07 H. BARON - P.O. BEFFY - N. FOURCADE - R. MAHIEU
Le ralentissement de la productivité du travail au cours des années 1990

G2003/08 P.-O. BEFFY - B. MONFORT
Patrimoine des ménages, dynamique d'allocation et comportement de consommation

G2003/09 P. BISCOURP - N. FOURCADE
Peut-on mettre en évidence l'existence de rigidités à la baisse des salaires à partir de données individuelles ? Le cas de la France à la fin des années 90

G2003/10 M. LECLAIR - P. PETIT
Présence syndicale dans les firmes : quel impact sur les inégalités salariales entre les hommes et les femmes ?

G2003/11 P.-O. BEFFY - X. BONNET - M. DARRACQ-PARIES - B. MONFORT
MZE: a small macro-model for the euro area

G2004/01 P. AUBERT - M. LECLAIR
La compétitivité exprimée dans les enquêtes trimestrielles sur la situation et les perspectives dans l'industrie

G2004/02 M. DUÉE - C. REBILLARD
La dépendance des personnes âgées : une projection à long terme

G2004/03 S. RASPILLER - N. RIEDINGER
Régulation environnementale et choix de localisation des groupes français

G2004/04 A. NABOULET - S. RASPILLER
Les déterminants de la décision d'investir : une approche par les perceptions subjectives des firmes

G2004/05 N. RAGACHE
La déclaration des enfants par les couples non mariés est-elle fiscalement optimale ?

G2004/06 M. DUÉE
L'impact du chômage des parents sur le devenir scolaire des enfants

G2004/07 P. AUBERT - E. CAROLI - M. ROGER
New Technologies, Workplace Organisation and the Age Structure of the Workforce: Firm-Level Evidence

G2004/08 E. DUGUET - C. LELARGE
Les brevets accroissent-ils les incitations privées à innover ? Un examen microéconométrique

G2004/09 S. RASPILLER - P. SILLARD
Affiliating versus Subcontracting: the Case of Multinationals

G2004/10 J. BOISSINOT - C. L'ANGEVIN - B. MONFORT
Public Debt Sustainability: Some Results on the French Case

G2004/11 S. ANANIAN - P. AUBERT
Travailleurs âgés, nouvelles technologies et changements organisationnels : un réexamen à partir de l'enquête « REPONSE »

G2004/12 X. BONNET - H. PONCET
Structures de revenus et propensions différentes à consommer - Vers une équation de consommation des ménages plus robuste en prévision pour la France

G2004/13 C. PICART
Évaluer la rentabilité des sociétés non financières

G2004/14 J. BARDAJI - B. SÉDILLOT - E. WALRAET
Les retraites du secteur public : projections à l'horizon 2040 à l'aide du modèle de microsimulation DESTINIE

G2005/01 S. BUFFETEAU - P. GODEFROY
Conditions de départ en retraite selon l'âge de fin d'études : analyse prospective pour les générations 1945 à1974

G2005/02 C. AFSA - S. BUFFETEAU
L'évolution de l'activité féminine en France : une approche par pseudo-panel

G2005/03 P. AUBERT - P. SILLARD
Délocalisations et réductions d'effectifs dans l'industrie française

G2005/04 M. LECLAIR - S. ROUX
Mesure et utilisation des emplois instables dans les entreprises

G2005/05 C. L'ANGEVIN - S. SERRAVALLE
Performances à l'exportation de la France

et de l'Allemagne - Une analyse par secteur et destination géographique

G2005/06 Bilan des activités de la Direction des Études et Synthèses Économiques - 2004

G2005/07 S. RASPILLER
La concurrence fiscale : principaux enseignements de l'analyse économique

G2005/08 C. L'ANGEVIN - N. LAÏB
Éducation et croissance en France et dans un panel de 21 pays de l'OCDE

G2005/09 N. FERRARI
Prévoir l'investissement des entreprises
Un indicateur des révisions dans l'enquête de conjoncture sur les investissements dans l'industrie.

G2005/10 P.-O. BEFFY - C. L'ANGEVIN
Chômage et boucle prix-salaires :
apport d'un modèle « qualifiés/peu qualifiés »

G2005/11 B. HEITZ
A two-states Markov-switching model of inflation in France and the USA: credible target VS inflation spiral

G2005/12 O. BIAU - H. ERKEL-ROUSSE - N. FERRARI
Réponses individuelles aux enquêtes de conjoncture et prévision macroéconomiques :
Exemple de la prévision de la production manufacturière

G2005/13 P. AUBERT - D. BLANCHET - D. BLAU
The labour market after age 50: some elements of a Franco-American comparison

G2005/14 D. BLANCHET - T. DEBRAND -
P. DOURGNON - P. POLLET
L'enquête SHARE : présentation et premiers résultats de l'édition française

G2005/15 M. DUÉE
La modélisation des comportements démographiques dans le modèle de microsimulation DESTINIE

G2005/16 H. RAOUI - S. ROUX
Étude de simulation sur la participation versée aux salariés par les entreprises

G2006/01 C. BONNET - S. BUFFETEAU - P. GODEFROY
Disparités de retraite de droit direct entre hommes et femmes : quelles évolutions ?

G2006/02 C. PICART
Les gazelles en France

G2006/03 P. AUBERT - B. CRÉPON -P. ZAMORA
Le rendement apparent de la formation continue dans les entreprises : effets sur la productivité et les salaires

G2006/04 J.-F. OUVRARD - R. RATHELOT
Demographic change and unemployment:
what do macroeconometric models predict?

G2006/05 D. BLANCHET - J.-F. OUVRARD
Indicateurs d'engagements implicites des systèmes de retraite : chiffrages, propriétés analytiques et réactions à des chocs démographiques types

G2006/06 G. BIAU - O. BIAU - L. ROUVIERE
Nonparametric Forecasting of the Manufacturing Output Growth with Firm-level Survey Data

G2006/07 C. AFSA - P. GIVORD
Le rôle des conditions de travail dans les absences pour maladie

G2006/08 P. SILLARD - C. L'ANGEVIN - S. SERRAVALLE
Performances comparées à l'exportation de la France et de ses principaux partenaires
Une analyse structurelle sur 12 ans

G2006/09 X. BOUTIN - S. QUANTIN
Une méthodologie d'évaluation comptable du coût du capital des entreprises françaises : 1984-2002

G2006/10 C. AFSA
L'estimation d'un coût implicite de la pénibilité du travail chez les travailleurs âgés

G2006/11 C. LELARGE
Les entreprises (industrielles) françaises sont-elles à la frontière technologique ?

G2006/12 O. BIAU - N. FERRARI
Théorie de l'opinion
Faut-il pondérer les réponses individuelles ?

G2006/13 A. KOUBI - S. ROUX
Une réinterprétation de la relation entre productivité et inégalités salariales dans les entreprises

G2006/14 R. RATHELOT - P. SILLARD
The impact of local taxes on plants location decision

G2006/15 L. GONZALEZ - C. PICART
Diversification, recentrage et poids des activités de support dans les groupes (1993-2000)

G2007/01 D. SRAER
Allègements de cotisations patronales et dynamique salariale

G2007/02 V. ALBOUY - L. LEQUIEN
Les rendements non monétaires de l'éducation : le cas de la santé

G2007/03 D. BLANCHET - T. DEBRAND
Aspiration à la retraite, santé et satisfaction au travail : une comparaison européenne

G2007/04 M. BARLET - L. CRUSSON
Quel impact des variations du prix du pétrole sur la croissance française ?

G2007/05 C. PICART
Flux d'emploi et de main-d'œuvre en France : un réexamen

G2007/06 V. ALBOUY - C. TAVAN
Massification et démocratisation de l'enseignement supérieur en France

G2007/07 T. LE BARBANCHON
The Changing response to oil price shocks in France: a DSGE type approach

G2007/08 T. CHANEY - D. SRAER - D. THESMAR
Collateral Value and Corporate Investment
Evidence from the French Real Estate Market

G2007/09 J. BOISSINOT
Consumption over the Life Cycle: Facts for France

G2007/10 C. AFSA
Interpréter les variables de satisfaction : l'exemple de la durée du travail

G2007/11 R. RATHELOT - P. SILLARD
Zones Franches Urbaines : quels effets sur l'emploi salarié et les créations d'établissements ?

G2007/12 V. ALBOUY - B. CRÉPON
Aléa moral en santé : une évaluation dans le cadre du modèle causal de Rubin

G2008/01 C. PICART
Les PME françaises : rentables mais peu dynamiques

G2008/02 P. BISCOURP - X. BOUTIN - T. VERGÉ
The Effects of Retail Regulations on Prices
Evidence form the Loi Galland

G2008/03 Y. BARBESOL - A. BRIANT
Économies d'agglomération et productivité des entreprises : estimation sur données individuelles françaises

G2008/04 D. BLANCHET - F. LE GALLO
Les projections démographiques : principaux mécanismes et retour sur l'expérience française

G2008/05 D. BLANCHET - F. TOUTLEMONDE
Évolutions démographiques et déformation du cycle de vie active : quelles relations ?

G2008/06 M. BARLET - D. BLANCHET - L. CRUSSON
Internationalisation et flux d'emplois : que dit une approche comptable ?

G2008/07 C. LELARGE - D. SRAER - D. THESMAR
Entrepreneurship and Credit Constraints - Evidence from a French Loan Guarantee Program

G2008/08 X. BOUTIN - L. JANIN
Are Prices Really Affected by Mergers?

G2008/09 M. BARLET - A. BRIANT - L. CRUSSON
Concentration géographique dans l'industrie manufacturière et dans les services en France : une approche par un indicateur en continu

G2008/10 M. BEFFY - É. COUDIN - R. RATHELOT
Who is confronted to insecure labor market histories? Some evidence based on the French labor market transition

G2008/11 M. ROGER - E. WALRAET
Social Security and Well-Being of the Elderly: the Case of France

G2008/12 C. AFSA
Analyser les composantes du bien-être et de son évolution
Une approche empirique sur données individuelles

G2008/13 M. BARLET - D. BLANCHET -
T. LE BARBANCHON
Microsimuler le marché du travail : un prototype

G2009/01 P.-A. PIONNIER
Le partage de la valeur ajoutée en France, 1949-2007

G2009/02 Laurent CLAVEL - Christelle MINODIER
A Monthly Indicator of the French Business Climate

G2009/03 H. ERKEL-ROUSSE - C. MINODIER
Do Business Tendency Surveys in Industry and Services Help in Forecasting GDP Growth?
A Real-Time Analysis on French Data

G2009/04 P. GIVORD - L. WILNER
Les contrats temporaires : trappe ou marchepied vers l'emploi stable ?

G2009/05 G. LALANNE - P.-A. PIONNIER - O. SIMON
Le partage des fruits de la croissance de 1950 à 2008 : une approche par les comptes de surplus

G2009/06 L. DAVEZIES - X. D'HAULTFOEUILLE
Faut-il pondérer ?… Ou l'éternelle question de l'économètre confronté à des données d'enquête

G2009/07 S. QUANTIN - S. RASPILLER - S. SERRAVALLE
Commerce intragroupe, fiscalité et prix de transferts : une analyse sur données françaises

G2009/08 M. CLERC - V. MARCUS
Élasticités-prix des consommations énergétiques des ménages

G2009/09 G. LALANNE - E. POULIQUEN - O. SIMON
Prix du pétrole et croissance potentielle à long terme

G2009/10 D. BLANCHET - J. LE CACHEUX - V. MARCUS
Adjusted net savings and other approaches to sustainability: some theoretical background

G2009/11 V. BELLAMY - G. CONSALES - M. FESSEAU - S. LE LAIDIER - É. RAYNAUD
Une décomposition du compte des ménages de la comptabilité nationale par catégorie de ménage en 2003

G2009/12 J. BARDAJI - F. TALLET
Detecting Economic Regimes in France: a Qualitative Markov-Switching Indicator Using Mixed Frequency Data

G2009/13 R. AEBERHARDT - D. FOUGÈRE -
R. RATHELOT
Discrimination à l'embauche : comment exploiter les procédures de *testing* ?

G2009/14 Y. BARBESOL - P. GIVORD - S. QUANTIN
Partage de la valeur ajoutée, approche par données microéconomiques

G2009/15 I. BUONO - G. LALANNE
The Effect of the Uruguay round on the Intensive and Extensive Margins of Trade

G2010/01 C. MINODIER
Avantages comparés des séries des premières valeurs publiées et des séries des valeurs révisées - Un exercice de prévision en temps réel
de la croissance trimestrielle du PIB en France

G2010/02 V. ALBOUY - L. DAVEZIES - T. DEBRAND
Health Expenditure Models: a Comparison of Five Specifications using Panel Data

G2010/03 C. KLEIN - O. SIMON
Le modèle MÉSANGE réestimé en base 2000
Tome 1 – Version avec volumes à prix constants

G2010/04 M.-É. CLERC - É. COUDIN
L'IPC, miroir de l'évolution du coût de la vie en France ? Ce qu'apporte l'analyse des courbes d'Engel

G2010/05 N. CECI-RENAUD - P.-A. CHEVALIER
Les seuils de 10, 20 et 50 salariés : impact sur la taille des entreprises françaises

G2010/06 R. AEBERHARDT - J. POUGET
National Origin Differences in Wages and Hierarchical Positions - Evidence on French Full-Time Male Workers from a matched Employer-Employee Dataset

G2010/07 S. BLASCO - P. GIVORD
Les trajectoires professionnelles en début de vie active : quel impact des contrats temporaires ?

G2010/08 P. GIVORD
Méthodes économétriques pour l'évaluation de politiques publiques

G2010/09 P.-Y. CABANNES - V. LAPÈGUE - E. POULIQUEN - M. BEFFY - M. GAINI
Quelle croissance de moyen terme après la crise ?

G2010/10 I. BUONO - G. LALANNE
La réaction des entreprises françaises à la baisse des tarifs douaniers étrangers

G2010/11 R. RATHELOT - P. SILLARD
L'apport des méthodes à noyaux pour mesurer la concentration géographique - Application à la concentration des immigrés en France de 1968 à 1999

G2010/12 M. BARATON - M. BEFFY - D. FOUGÈRE
Une évaluation de l'effet de la réforme de 2003 sur les départs en retraite - Le cas des enseignants du second degré public

G2010/13 D. BLANCHET - S. BUFFETEAU - E. CRENNER S. LE MINEZ
Le modèle de microsimulation Destinie 2 : principales caractéristiques et premiers résultats

G2010/14 D. BLANCHET - E. CRENNER
Le bloc retraites du modèle Destinie 2 : guide de l'utilisateur

G2010/15 M. BARLET - L. CRUSSON - S. DUPUCH - F. PUECH
Des services échangés aux services échangeables : une application sur données françaises

G2010/16 M. BEFFY - T. KAMIONKA
Public-private wage gaps: is civil-servant human capital sector-specific?

G2010/17 P.-Y. CABANNES - H. ERKEL-ROUSSE - G. LALANNE - O. MONSO - E. POULIQUEN
Le modèle Mésange réestimé en base 2000
Tome 2 - Version avec volumes à prix chaînés

G2010/18 R. AEBERHARDT - L. DAVEZIES
Conditional Logit with one Binary Covariate: Link between the Static and Dynamic Cases

G2011/01 T. LE BARBANCHON - B. OURLIAC - O. SIMON
Les marchés du travail français et américain face aux chocs conjoncturels des années 1986 à 2007 : une modélisation DSGE

G2011/02 C. MARBOT
Une évaluation de la réduction d'impôt pour l'emploi de salariés à domicile

G2011/03 L. DAVEZIES
Modèles à effets fixes, à effets aléatoires, modèles mixtes ou multi-niveaux : propriétés et mises en œuvre des modélisations de l'hétérogénéité dans le cas de données groupées

G2011/04 M. ROGER - M. WASMER
Heterogeneity matters: labour productivity differentiated by age and skills

G2011/05 J.-C. BRICONGNE - J.-M. FOURNIER V. LAPÈGUE - O. MONSO
De la crise financière à la crise économique L'impact des perturbations financières de 2007 et 2008 sur la croissance de sept pays industrialisés

G2011/06 P. CHARNOZ - É. COUDIN - M. GAINI
Wage inequalities in France 1976-2004: a quantile regression analysis

G2011/07 M. CLERC - M. GAINI - D. BLANCHET
Recommendations of the Stiglitz-Sen-Fitoussi Report: A few illustrations

G2011/08 M. BACHELET - M. BEFFY - D. BLANCHET
Projeter l'impact des réformes des retraites sur l'activité des 55 ans et plus : une comparaison de trois modèles

G2011/09 C. LOUVOT-RUNAVOT
L'évaluation de l'activité dissimulée des entreprises sur la base des contrôles fiscaux et son insertion dans les comptes nationaux

G2011/10 A. SCHREIBER - A. VICARD
La tertiarisation de l'économie française et le ralentissement de la productivité entre 1978 et 2008

G2011/11 M.-É. CLERC - O. MONSO - E. POULIQUEN
Les inégalités entre générations depuis le baby-boom

G2011/12 C. MARBOT - D. ROY
Évaluation de la transformation de la réduction d'impôt en crédit d'impôt pour l'emploi de salariés à domicile en 2007

G2011/13 P. GIVORD - R. RATHELOT - P. SILLARD
Place-based tax exemptions and displacement effects: An evaluation of the Zones Franches Urbaines program

G2011/14 X. D'HAULTFOEUILLE - P. GIVORD - X. BOUTIN
The Environmental Effect of Green Taxation: the Case of the French "Bonus/Malus"

G2011/15 M. BARLET - M. CLERC - M. GARNEO - V. LAPÈGUE - V. MARCUS
La nouvelle version du modèle MZE, modèle macroéconométrique pour la zone euro

G2011/16 R. AEBERHARDT - I. BUONO - H. FADINGER
Learning, Incomplete Contracts and Export Dynamics: Theory and Evidence form French Firms

G2011/17 C. KERDRAIN - V. LAPÈGUE
Restrictive Fiscal Policies in Europe: What are the Likely Effects?

G2012/01 P. GIVORD - S. QUANTIN - C. TREVIEN
A Long-Term Evaluation of the First Generation of the French Urban Enterprise Zones

G2012/02 N. CECI-RENAUD - V. COTTET
Politique salariale et performance des entreprises

G2012/03 P. FÉVRIER - L. WILNER
Do Consumers Correctly Expect Price Reductions? Testing Dynamic Behavior

G2012/04 M. GAINI - A. LEDUC - A. VICARD
School as a shelter? School leaving-age and the business cycle in France

G2012/05 M. GAINI - A. LEDUC - A. VICARD
A scarred generation? French evidence on young people entering into a tough labour market

G2012/06 P. AUBERT - M. BACHELET
Disparités de montant de pension et redistribution dans le système de retraite français

G2012/07 R. AEBERHARDT - P GIVORD - C. MARBOT
Spillover Effect of the Minimum Wage in France: An Unconditional Quantile Regression Approach

G2012/08 A. EIDELMAN - F. LANGUMIER - A. VICARD
Prélèvements obligatoires reposant sur les ménages : des canaux redistributifs différents en 1990 et 2010

G2012/09 O. BARGAIN - A. VICARD
Le RMI et son successeur le RSA découragent-ils certains jeunes de travailler ? Une analyse sur les jeunes autour de 25 ans

G2012/10 C. MARBOT - D. ROY
Projections du coût de l'APA et des caractéristiques de ses bénéficiaires à l'horizon 2040 à l'aide du modèle Destinie

G2012/11 A. MAUROUX
Le crédit d'impôt dédié au développement durable : une évaluation économétrique

G2012/12 V. COTTET - S. QUANTIN - V. RÉGNIER
Coût du travail et allègements de charges : une estimation au niveau établissement de 1996 à 2008

G2012/13 X. D'HAULTFOEUILLE - P. FÉVRIER - L. WILNER
Demand Estimation in the Presence of Revenue Management

G2012/14 D. BLANCHET - S. LE MINEZ
Joint macro/micro evaluations of accrued-to-date pension liabilities: an application to French reforms

G2013/01-F1301 T. DEROYON - A. MONTAUT - P-A PIONNIER
Utilisation rétrospective de l'enquête Emploi à une fréquence mensuelle : apport d'une modélisation espace-état

G2013/02-F1302 C. TREVIEN
Habiter en HLM : quel avantage monétaire et quel impact sur les conditions de logement ?

G2013/03 A. POISSONNIER
Temporal disaggregation of stock variables - The Chow-Lin method extended to dynamic models

G2013/04 P. GIVORD - C. MARBOT
Does the cost of child care affect female labor market participation? An evaluation of a French reform of childcare subsidies

G2013/05 G. LAME - M. LEQUIEN - P.-A. PIONNIER
Interpretation and limits of sustainability tests in public finance

G2013/06 C. BELLEGO - V. DORTET-BERNADET
La participation aux pôles de compétitivité : quelle incidence sur les dépenses de R&D et l'activité des PME et ETI ?

G2013/07 P.-Y. CABANNES - A. MONTAUT - P.-A. PIONNIER
Évaluer la productivité globale des facteurs en France : l'apport d'une mesure de la qualité du capital et du travail

G2013/08 R. AEBERHARDT - C. MARBOT
Evolution of Instability on the French Labour Market During the Last Thirty Years

G2013/09 J-B. BERNARD - G. CLÉAUD
Oil price: the nature of the shocks and the impact on the French economy

G2013/10 G. LAME
Was there a « Greenspan Conundrum » in the Euro area?

G2013/11 P. CHONÉ - F. EVAIN - L. WILNER - E. YILMAZ
Introducing activity-based payment in the hospital industry : Evidence from French data

G2013/12 C. GRISLAIN-LETRÉMY
Natural Disasters: Exposure and Underinsurance

G2013/13 P.-Y. CABANNES - V. COTTET - Y. DUBOIS - C. LELARGE - M. SICSIC
French Firms in the Face of the 2008/2009 Crisis

G2013/14 A. POISSONNIER - D. ROY
Households Satellite Account for France in 2010. Methodological issues on the assessment of domestic production

G2013/15 G. CLÉAUD - M. LEMOINE - P.-A. PIONNIER
Which size and evolution of the government expenditure multiplier in France (1980-2010)?

G2014/01 M. BACHELET - A. LEDUC - A. MARINO
Les biographies du modèle Destinie II : rebasage et projection

G2014/02 B. GARBINTI
L'achat de la résidence principale et la création d'entreprises sont-ils favorisés par les donations et héritages ?

G2014/03 N. CECI-RENAUD - P. CHARNOZ - M. GAINI
Évolution de la volatilité des revenus salariaux du secteur privé en France depuis 1968

G2014/04 P. AUBERT
Modalités d'application des réformes des retraites et prévisibilité du montant de pension

G2014/05 C. GRISLAIN-LETRÉMY - A. KATOSSKY
The Impact of Hazardous Industrial Facilities on Housing Prices: A Comparison of Parametric and Semiparametric Hedonic Price Models

G2014//06 J.-M. DAUSSIN-BENICHOU - A. MAUROUX
Turning the heat up. How sensitive are households to fiscal incentives on energy efficiency investments?

G2014/07 C. LABONNE - G. LAMÉ
Credit Growth and Capital Requirements: Binding or Not?

G2014/08 C. GRISLAIN-LETRÉMY et C. TREVIEN
The Impact of Housing Subsidies on the Rental Sector: the French Example

G2014/09 M. LEQUIEN et A. MONTAUT
Croissance potentielle en France et en zone euro : un tour d'horizon des méthodes d'estimation

G2014/10 B. GARBINTI - P. LAMARCHE
Les hauts revenus épargnent-ils davantage ?

G2014/11 D. AUDENAERT - J. BARDAJI - R. LARDEUX - M. ORAND - M. SICSIC
Wage Resilience in France since the Great Recession

G2014/12 F. ARNAUD - J. BOUSSARD - A. POISSONNIER - H. SOUAL
Computing additive contributions to growth and other issues for chain-linked quarterly aggregates

G2014/13 H. FRAISSE - F. KRAMARZ - C. PROST
Labor Disputes and Job Flows

G2014/14 P. GIVORD - C. GRISLAIN-LETRÉMY - H. NAEGELE
How does fuel taxation impact new car purchases? An evaluation using French consumer-level dataset

G2014/15 P. AUBERT - S. RABATÉ
Durée passée en carrière et durée de vie en retraite : quel partage des gains d'espérance de vie ?

G2015/01 A. POISSONNIER
The walking dead Euler equation
Addressing a challenge to monetary policy models

G2015/02 Y. DUBOIS - A. MARINO
Indicateurs de rendement du système de retraite français

G2015/03 T. MAYER - C. TREVIEN
The impacts of Urban Public Transportation: Evidence from the Paris Region

G2015/04 S.T. LY - A. RIEGERT
Measuring Social Environment Mobility

G2015/05 M. A. BEN HALIMA - V. HYAFIL-SOLELHAC
M. KOUBI - C. REGAERT
Quel est l'impact du système d'indemnisation maladie sur la durée des arrêts de travail pour maladie ?

G2015/06 Y. DUBOIS - A. MARINO
Disparités de rendement du système de retraite dans le secteur privé : approches intergénérationnelle et intragénérationnelle

G2015/07 B. CAMPAGNE - V. ALHENC-GELAS - J.-B. BERNARD
No evidence of financial accelerator in France

G2015/08 Q. LAFFÉTER - M. PAK
Élasticités des recettes fiscales au cycle économique : étude de trois impôts sur la période 1979-2013 en France

G2015/09 J.-M. DAUSSIN-BENICHOU, S. IDMACHICHE, A. LEDUC et E. POULIQUEN
Les déterminants de l'attractivité de la fonction publique de l'État

G2015/10 P. AUBERT
La modulation du montant de pension selon la durée de carrière et l'âge de la retraite : quelles disparités entre assurés ?

G2015/11 V. DORTET-BERNADET - M. SICSIC
Effet des aides publiques sur l'emploi en R&D dans les petites entreprises

G2015/12 S. GEORGES-KOT
Annual and lifetime incidence of the value-added tax in France

G2015/13 M. POULHÈS
Are Enterprise Zones Benefits Capitalized into Commercial Property Values? The French Case

G2015/14 J.-B. BERNARD - Q. LAFFÉTER
Effet de l'activité et des prix sur le revenu salarial des différentes catégories socioprofessionnelles

G2015/15 C. GEAY - M. KOUBI - G de LAGASNERIE
Projections des dépenses de soins de ville, construction d'un module pour Destinie

G2015/16 J. BARDAJI - J.-C. BRICONGNE - B. CAMPAGNE - G. GAULIER
Compared performances of French companies on the domestic and foreign markets

G2015/17 C. BELLÉGO - R. DE NIJS
The redistributive effect of online piracy on the box office performance of American movies in foreign markets

G2015/18 J.-B. BERNARD - L. BERTHET
French households financial wealth: which changes in 20 years?

G2015/19 M. POULHÈS
Fenêtre sur Cour ou *Chambre avec Vue* ?
Les prix hédoniques de l'immobilier parisien

G2016/01 B. GARBINTI - S. GEORGES-KOT
Time to smell the roses? Risk aversion, the timing of inheritance receipt, and retirement

G2016/02 P. CHARNOZ - C. LELARGE - C. TREVIEN
Communication Costs and the Internal Organization of Multi-Plant Businesses: Evidence from the Impact of the French High-Speed Rail

G2016/03 C. BONNET - B. GARBINTI - A. SOLAZ
Gender Inequality after Divorce: The Flip Side of Marital Specialization - Evidence from a French Administrative Database

G2016/04 D. BLANCHET - E. CAROLI - C. PROST - M. ROGER
Health capacity to work at older ages in France

G2016/05 B. CAMPAGNE - A. POISSONNIER
MELEZE: A DSGE model for France within the Euro Area

G2016/06 B. CAMPAGNE - A. POISSONNIER
Laffer curves and fiscal multipliers: lessons from Mélèze model

G2016/07 B. CAMPAGNE - A. POISSONNIER
Structural reforms in DSGE models: a case for sensitivity analyses

G2016/08 Y. DUBOIS et M. KOUBI
Relèvement de l'âge de départ à la retraite : quel impact sur l'activité des séniors de la réforme des retraites de 2010 ?

G2016/09 A. NAOUAS - M. ORAND - I. SLIMANI HOUTI
Les entreprises employant des salariés au Smic : quelles caractéristiques et quelle rentabilité ?

G2016/10 T. BLANCHET - Y. DUBOIS - A. MARINO - M. ROGER
Patrimoine privé et retraite en France

G2016/11 M. PAK - A. POISSONNIER
Accounting for technology, trade and final consumption in employment: an Input-Output decomposition

G2017/01 D. FOUGÈRE - E. GAUTIER - S. ROUX
Understanding Wage Floor Setting in Industry-Level Agreements: Evidence from France

G2017/02 Y. DUBOIS - M. KOUBI
Règles d'indexation des pensions et sensibilité des dépenses de retraites à la croissance économique et aux chocs démographiques

G2017/03 A. CAZENAVE-LACROUTZ - F. GODET
L'espérance de vie en retraite sans incapacité sévère des générations nées entre 1960 et 1990 : une projection à partir du modèle Destinie

G2017/04 J. BARDAJI - B. CAMPAGNE - M.-B. KHDER - Q. LAFFÉTER - O. SIMON (Insee)
A.-S. DUFERNEZ - C. ELEZAAR - P. LEBLANC - E. MASSON - H. PARTOUCHE (DG-Trésor)
Le modèle macroéconométrique Mésange : réestimation et nouveautés

G2017/05 J. BOUSSARD - B. CAMPAGNE
Fiscal Policy Coordination in a Monetary Union at the Zero-Lower-Bound

G2017/06 A. CAZENAVE-LACROUTZ - A. GODZINSKI
Effects of the one-day waiting period for sick leave on health-related absences in the French central civil service

G2017/07 P. CHARNOZ - M. ORAND
Qualification, progrès technique et marchés du travail locaux en France, 1990-2011

G2017/08 K. MILIN
Modélisation de l'inflation en France par une approche macrosectorielle

G2017/09 C.-M. CHEVALIER - R. LARDEUX
Homeownership and labor market outcomes: disentangling externality and composition effects

G2017/10 P. BEAUMONT
Time is Money: Cash-Flow Risk and Export Market Behavior

G2018/01 S. ROUX - F. SAVIGNAC
SMEs' financing: Divergence across Euro area countries?

G2018/02 C.-M. CHEVALIER - A. LUCIANI
Computerization, labor productivity and employment: impacts across industries vary with technological level

G2018/03 R.MONIN - M. SUAREZ CASTILLO
L'effet du CICE sur les prix : une double analyse sur données sectorielles et individuelles

G2018/04 R. LARDEUX
Who Understands The French Income Tax? Bunching Where Tax Liabilities Start

G2018/05 C.-M. CHEVALIER
Financial constraints of innovative firms and sectoral growth

G2018/06 R. S.-H. LEE - M. PAK
Pro-competitive effects of globalisation on prices, productivity and markups: Evidence in the Euro Area

G2018/07 C.-M. CHEVALIER
Consumption inequality in France between 1995 and 2011

G2018/08 A. BAUER - B. GARBINTI - S. GEORGES-KOT
Financial Constraints and Self-Employment in France, 1945-2014

G2018/09 P. BEAUMONT – A. LUCIANI
Prime à l'embauche dans les PME : évaluation à partir des déclarations d'embauche

G2018/10 C BELLÉGO – V. DORTET-BERNADET - M. TÉPAUT
Comparaison de deux dispositifs d'aide à la R&D collaborative public-privé

G2018/11 R. MONIN – M. SUAREZ CASTILLO
Réplication et rapprochement des travaux d'évaluation de l'effet du CICE sur l'emploi en 2013 et 2014

G2018/12 A. CAZENAVE-LACROUTZ - F. GODET – V. LIN
L'introduction d'un gradient social dans la mortalité au sein du modèle Destinie 2

G2019/01 F1901 M. ANDRÉ – A.-L. BIOTTEAU
Effets de moyen terme d'une hausse de TVA sur le niveau de vie et les inégalités : une approche par microsimulation

G2019/02 A. BOURGEOIS – A. BRIAND
Le modèle Avionic : la modélisation Input/Output des comptes nationaux

G2019/03 A. GODZINSKI – M. SUAREZ CASTILLO
Short-term health effects of public transport disruptions: air pollution and viral spread channels

G2019/04 F1903 L. AEBERHARDT - F. HATIER - M. LECLAIR - B. PENTINAT - J.-D. ZAFAR
L'économie numérique fausse-t-elle le partage volume-prix du PIB ?

G2019/05 A. CAZENAVE-LACROUTZ – E. YILMAZ
Dans quelle mesure les incitations tarifaires et la procédure de mise sous accord préalable ont-elles contribué au développement de la chirurgie ambulatoire ?

G2019/06 J.-P. CLING – S. EGHBAL-TEHERANI – M. ORZONI – C. PLATEAU
The Differences between EU Countries for Sustainable Development Indicators:
It is (mainly) the Economy!

G2019/07 P. CHONÉ – L. WILNER
Competition on Unobserved Attributes: The Case of the Hospital Industry

G2019/08 P. PORA – L. WILNER
Child Penalties and Financial Incentives: Exploiting Variation along the Wage Distribution

G2019/09 E. GAUTIER – S. ROUX – M. SUAREZ CASTILLO
Do Minimum Wages make Wages more Rigid ? Evidence from French Micro Data

G2019/10 M. ANDRÉ – A. SIREYJOL
Imposition des couples et des familles : effets budgétaires et redistributifs de l’impôt sur le revenu

G2019/11 K. MOHKAM – O. SIMON
L’empreinte matière de l’économie française : une analyse par matière et catégorie de produits

G2019/12 S. BUNEL – B. HADJIBEYLI
Évaluation du crédit d’impôt innovation

G2019/13 C. BONNET – F. GODET – A. SOLAZ
Gendered economic determinants of couple formation over 50 in France

G2019/14 P. GIVORD – M. SUAREZ CASTILLO
Excellence for all? Heterogeneity in high schools’ value-added

G2019/15 G. CETTE – L. KOEHL – T. PHILIPPON

Labor Share in Some Advanced Countries

G2020/01 J. SILHOL – B. VENTELOU

Les zones d’intervention prioritaire reflètent-elles des écarts de pratiques des médecins généralistes ?

G2020/02 B. BOUCHTENIK – R. LARDEUX

The Take-Up of Unemployment Benefit Extensions

G2020/03 J-M. GERMAIN

A Welfare Based Estimate of “Real Feel GDP” for Europe and the USA

G2020/04 J. BOUSSARD – R. LEE

Competition, Profit Share and Concentration

G2020/05 P. PORA

Keep Working and Spend Less?
Collective Childcare and Parental Earnings in France

G2020/06 R. MONIN – M. SUAREZ CASTILLO

Product Switching, market power and distance to core competency

G2020/07 L. WILNER

How do citizens perceive centralization reforms? Evidence from the merger of French regions

G2020/08 L. WILNER
The persistence of subjective well-being: permanent happiness, transitory misery?

G2020/09 J-M. GERMAIN – T. LELLOUCH

Comptabilité économique de la soutenabilité climatique

G2020/10 A. BAUER – M. ROTEMBERG

Tax avoidance in French Firms:
Evidence from the Introduction of a Tax Notch

G2020/11 P. AGHION - A. BERGEAUD

M. LEQUIEN - M. J. MELITZ

The Heterogeneous Impact of Market Size on Innovation: Evidence from French Firm-Level Exports

G2020/12 L. GALIANA – B. SAKAROVITCH

F. SÉMÉCURBE – Z. SMOREDA

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