Inequality of opportunity in France:

The role of spatial segregation in the ethnic gap

Master's Thesis

July, 2020

Louis Sirugue*
Paris School of Economics
Master 2 - APE

Supervisor:
Thomas Piketty
Paris School of Economics

Abstract

I take advantage of the recent improvements in the Permanent Demographic Sample to provide new estimations of intergenerational mobility in France. I show that despite very comparable educational investments, second-generation immigrants from Maghreb experience lower absolute upward mobility in earnings than children of French natives. Results suggest that this discrepancy stems from differences in terms of access to employment. In addition to the well-documented hiring discrimination on the French labor market, I hypothesize spatial segregation to be linked with this phenomenon. I provide the first estimations of segregation in France that simultaneously (1) are not restricted to the first generation and (2) cover the whole French territory (3) using variations at a level as granular as that of the municipality. While the French literature had only relied on nationality and place of birth so far, I develop for that purpose an algorithm that can infer an individual's origin based on her last name. Segregation indices are significantly associated with lower incomes for second-generation immigrants from Maghreb, and with a higher intergenerational persistence irrespective of individuals' origin, but to a greater extent for children of immigrants from Maghreb.

***K*eywords** Intergenerational mobility · Segregation · Ethnicity · France

JEL codes J61 · J62 · R23

*This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissement d'Avenir" program (reference: ANR-10-EQPX-17 - Centre d'accès sécurisé aux données - CASD). I thank all participants in the Annual Meeting on the Permanent Demographic Sample of the INSEE for sharing their knowledge of, and experience with, the database, as well as Paul Brandily-Snyers, Pascale Champalaune, Baptiste Coulmont, Youssef El Jai, and Yajna Govind, for their time and valuable comments along the way. I am particularly grateful to Patrick Simon for refereeing this Master's Thesis and to Thomas Piketty for accepting to supervise this project, for his precious advice, and continuous support.

Contents

1 Introduction

By international standards, the significant involvement of the French state in the education and the tax system is supposed to ensure a satisfactory level of equality of opportunity among its citizens. Under the assumption that ability is partly heritable, and that wealthier parents invest more in their children's education, one could argue that the distortionary cost of redistribution society should bear to dissolve any intergenerational persistence is too high for perfect mobility to be socially optimal. But irrespective of the level of intergenerational mobility in a country, differentiated expected outcomes on the sole basis of ethnic origins conditional on parental socioeconomic background would perpetuate earnings inequality between individuals from French and foreign origin over generations, and could hardly be economically and ethically legitimized. Quantifying such an inequality of opportunity between the offspring of natives and ethnic minorities, and understanding its potential determinants, be they related to discrimination or other factors such as spatial segregation, is of particular interest as inceptive evidence to relevant policy recommendations.

1.1 Main contributions

The contribution of this Master's Thesis to the academic literature is threefold. First, I take advantage of the recent improvements in the Permanent Demographic Sample to provide new estimations of intergenerational mobility in France. Using intergenerational elasticities and rank-rank estimators, and after a careful evaluation of the different biases they may be subject to, I show that levels of intergenerational mobility in France are comparable to the estimations put forward for the United States, i.e., among the lowest in previously studied OECD countries. I then contribute to the French literature on intergenerational mobility by exploring heterogeneity according to ethnic origins. I show that despite very comparable educational investments, second-generation immigrants from Maghreb experience lower absolute upward mobility in earnings than children of French natives. Results suggest that this discrepancy stems from differences in terms of access to employment: conditional on parental earnings, while hourly wages are not significantly different, the number of hours worked is systematically lower and the probability to have perceived unemployment benefits over the period studied is systematically larger for second-generation immigrants from Maghreb.

To investigate how spatial segregation may play a role in this relationship, the second main contribution consists in providing the first estimations of segregation in France that simultaneously

(1) are not restricted to the first generation and (2) cover the whole French territory (3) using variations at a level as granular as that of the municipality. The comparison between segregation estimations based on countries of birth and estimations based on last names indicates that traditional estimations understate the level of segregation as individuals of Arabic origin born in France tend to locate in municipalities where foreign-born individuals of Arabic origin are already over-represented. The estimated segregation indices are matched to the department of birth of individuals in the Permanent Demographic Sample to estimate how segregation in childhood's environment relates to intergenerational mobility, and whether the intensity of this relationship differs according to one's origin. Segregation indices are significantly associated with lower incomes for second-generation immigrants from Maghreb, and with a higher intergenerational persistence irrespective of individuals' origin, but to a greater extent for children of immigrants from Maghreb.

The provision of such estimations of segregation is made possible by what constitutes the third contribution of this Master's Thesis to the literature: I develop an algorithm that can infer an individual's origin based on her last name. Simply put, the algorithm compares in a flexible way the sub-sequences of letters forming the name whose origin has to be determined to the sub-sequences of letters appearing in well-defined corpora of names of French and Arabic origin. The parametrization of the algorithm is based on an axiomatic approach. Several robustness checks are performed and support that the algorithm correctly targets the population of interest. Given the restrictions of French data on the origins of individuals, the development of this tool paves the way for research advances not only on spatial segregation, but on any socio-demographic indicator for which the use of place of birth or nationality instead of the actual origin constitutes a restriction to the conduction of comprehensive analyses on well-defined groups of individuals.

1.2 Contextual matters

The traditional estimator used to evaluate intergenerational mobility is the elasticity between incomes of children and their parents'. The economic literature on the topic initially focused on providing such estimates at the country level. Some stylized facts were then put forward, the most cited one being the negative relationship between inequality and intergenerational mobility, often referred to as the "Great Gatsby Curve" (Corak, 2013a). Still, it is difficult to draw clear-cut conclusions based on cross-country comparisons as national contexts vary substantially and in a wide variety of ways. Thus, a more recent field of this literature investigated local variations of mobility within countries, notably with a series of articles by Raj Chetty and co-authors that gathers a collection of new evidence on the issue. Using federal income tax records, they

documented large variations in the correlation between the rank of children and their parents' in the income distribution across US counties and commuting zones (Chetty et al., 2014). Their latest contribution to this literature investigates racial disparities in economic opportunity (Chetty et al., 2020). Results notably reveal that Black Americans experience much lower rates of upward mobility than Whites, sustaining the income gap from one generation to the next.

In France however, little is known about intergenerational mobility, surely because no large-scale data linking individuals' incomes to their parents' have been collected yet. As a result, most studies on the French intergenerational mobility focused on social class rather than income (Bourdieu et al., 2009; Dherbécourt, 2015; Poncelet et al., 2016). But while socio-professional categories may keep the same appellation over two generations, what they encapsulate surely evolves dramatically over time. Pecuniary measures are arguably more time-consistent. But to estimate intergenerational mobility in earnings on French data, parental income can only be predicted based on socio-demographic characteristics, either using data on older individuals from previous waves of a same sample, or by using two different data sources with common covariates. This was done by Lefranc and Trannoy (2005) using several waves of INSEE's (French National Institute for Statistics and Economic Studies) FQP (*Formation, Qualification, Profession)* database to compute national estimates of intergenerational mobility, that were recently revised by Lefranc (2018) as part of a long-run evolution analysis. But apart from that, little evidence were yet gathered. The big picture remains to be drawn, and the links between intergenerational mobility and several key factors, such as ethnic origins and spatial segregation, need to be evaluated as inceptive evidence to relevant policy recommendations.

In the United States, Chetty and Hendren (2018a) investigated how spatial variations in intergenerational mobility are driven by the causal effect of commuting zones, and Chetty and Hendren (2018b) identified the features of areas producing good outcomes. They notably showed that a 1 standard deviation increase in segregation in a commuting zone is associated with a 5.2% reduction in children's incomes for families at the 25^{th} income percentile. Along with the fraction of single mothers and income inequality in the county, segregation seems particularly influential. These demographic indicators were then shown to be tightly linked with the perpetuation of the Black-White gap in the US (Chetty et al., 2020). Just like for intergenerational mobility, the French literature on spatial segregation according to ethnic origins is bounded by data constraints. Due to the restrictions on the collection of information on ethnicity in France, researchers had to rely either on nationality or place of birth to compute segregation indices so far. The main data

sources that were used for this purpose are the Labor Force Survey and the Population Censuses (Gobillon and Selod, 2007; Préteceille, 2011; Quillian and Lagrange, 2016; Safi, 2009; Pan Ké Shon and Verdugo, 2015; Verdugo, 2011). The first limit of these data is that their scope only allows to measure segregation for the biggest metropolitan areas. But most importantly, the reliance on country of birth and nationality restricts the analysis to first-generation immigrants, which lessens the relevance of these estimations, especially for most recent periods. The other source the literature relied on to study segregational patterns is survey data (McAvay and Safi, 2018). The most influential one certainly is *Trajectoires et Origines* (TeO), a French cross-sectional survey that notably allows to observe the generation of immigration up to the third one. This is particularly suitable to define accurately the different subgroups of the population over which to compute segregation, but the sample size does not allow to produce estimates on more than a few thousand observations. All these current limitations not only constrain the accuracy of the measurement of segregation in France, but also the thorough analysis of its implications. It is thus of particular interest to provide new methods that would allow to estimate segregation over the whole French territory, accounting for actual origin rather than place of birth, and using variations at a level as granular as that of the municipality.

A way to infer origin when ethnic variables are not directly available consists in relying on individuals' last names. The cultural anchor of the tight relation between one's last name and origin has raised the attention of researchers from various disciplines, starting by Genetics (Jobling, 2001; King et al., 2006; Lasker, 1985), Biology and Medical research (Choi et al., 1993; Lasker, 1980; Polednak, 1993; Shah et al., 2010), as well as Genealogy (King and Jobling, 2009). The appropriation of such methods by social scientists is relatively recent, and has notably been used in research on intergenerational mobility either from a historical perspective (Clark et al., 2015), or by looking at the joint distribution of surnames and economic outcomes (Güell et al., 2013), and to study the under-representation of some origins in specific occupations and education levels (Mazieres and Roth, 2018). The method the literature relied on so far is based on drawing country-specific lists of names (generally by linking the name and institution of authors of scholarly publications), the origin of a name then being inferred from its occurrence in those different lists[1]. These standards of inference are quite rigid and have some obvious shortcomings. They are notably sensitive to misspellings and their validity is threatened by migration patterns (Mateos et al., 2014). These limitations were partially tackled by Mazieres and Roth (2018): rather than relying on the occurrence of the name in different corpora, they considered the

[1]See Mateos (2007) for a methodological review.

occurrence of its sub-sequences of letters. But this literature is still in its infancy. In order to provide estimations of origin based on last names that are reliable enough to be applied to the large-scale measurement of segregation and other socio-demographic indicators, the standards that consist in relying solely on relative frequencies in contaminated corpora must be improved by the construction of flexible algorithms, either based on parametric assumptions or on Machine Learning techniques.

Section 2 describes the data, how parental income is predicted, and the variables used throughout the analysis. Section 3 tests how sensitive the intergenerational persistence estimates are to the main sources of bias identified in prior literature, in order put forward national estimations on an appropriate sample. It then investigates heterogeneity according to ethnic origins and its potential underlying channels. Section 4 concentrates on spatial segregation. It first describes the algorithm constructed to infer one's origin from her last name, performs robustness checks to assess the validity of the predictions, and applies it on more than 12,000,000 individuals to provide segregation indices at the department level. Prior to concluding and discussing avenues for future research in Section 6, Section 5 analyzes the links between intergenerational mobility, ethnic origins, and the level of segregation experienced during childhood.

2 Data

2.1 Sample definitions

The Permanent Demographic Sample was established in 1968 by the INSEE as the first large-scale socio-demographic panel in France. It tracks all individuals born during the first 4 days of October[2] through birth, marriage, and death records, periodic censuses, electoral registers, annual declarations of social data, and since 2011, fiscal data. Thus, detailed information about individuals' earnings is available from 2010 to 2016, but earnings of their parents can only be inferred based on the socio-demographic variables collected in the population census waves that took place in 1975, 1982, 1990, and 1999.

For the purpose of the analysis, the sample is restricted to all individuals whose fiscal data are observed at least once from 2010 to 2016 while they were aged between 30 and 55, and whose parental information is observed at least once for at least one parent aged between 30 and 55 at the time of the observation as well, while individuals are declared as dependent children of

[2]The EDP selection criterion progressively widened to include people born during the first days of January, April, and July.

the household they belong to. As parental information is missing for virtually all foreign-born individuals, first-generation immigrants are practically absent from the analysis. Also, as parental income can only be predicted for the employed population due to restrictions of the sample the prediction model will be based on, only wages of individuals whose main source of income is stated as being their wage are considered. Depending on the income variable used, between 70,000 and 75,000 individuals meet these selection conditions.

2.2 Parental income predictions

Even if parental income cannot be observed directly, the Employee Panel (*Panel d'Actifs*) of the EDP contains the annual wage and socio-demographic characteristics of employed people in the sample on a yearly basis. A prediction model can thus be estimated with the Employee Panel based on gender, age, socio-professional category, year, and place of work, to predict parental income from census data based on these very characteristics.

The geographical units used to define the place of work are the French departments, which divide France into about 100 territories whose borders did not change over the whole period. Parents' place of work can be integrated to the prediction model estimated on the Employee Panel, but is not observed in census data. Thus, one can only use place of residence as proxy instead. As individuals who commute to a different place from the one they live in may earn more on average because of systematic differences in motivation or other unobservable characteristics, relying on place of residence could bias the estimation. But another effect, more specific to France, could theoretically select poorer individuals. Indeed, the city of Paris is a department in itself, where rents are particularly high. But the Paris commuting zone spreads over 7 other departments, and already gathered 17.26% of the whole French population in 1975[3]. Thus, a substantial share of the population that is observed commuting to a different department could actually be poorer, living in the departments surrounding Paris to commute to this city-department. With that in mind, it is not clear in which direction the selection goes. But the fact that place of work and place of residence are both observed in censuses for individuals in working age allows to check that 83% to 85% of them live and work in the same geographical unit at each of the points in time considered for the predictions, i.e., 1975, 1982, 1990, and 1999. With more than 80% of accuracy, it seems reasonable to use the place of residence as a proxy for the place of work, especially given that some of the individuals who commute to a different department are probably living close to a border and should thus not be the source of any particular bias.

[3]This share slightly increased over the period, up to 19.79% in 1999.

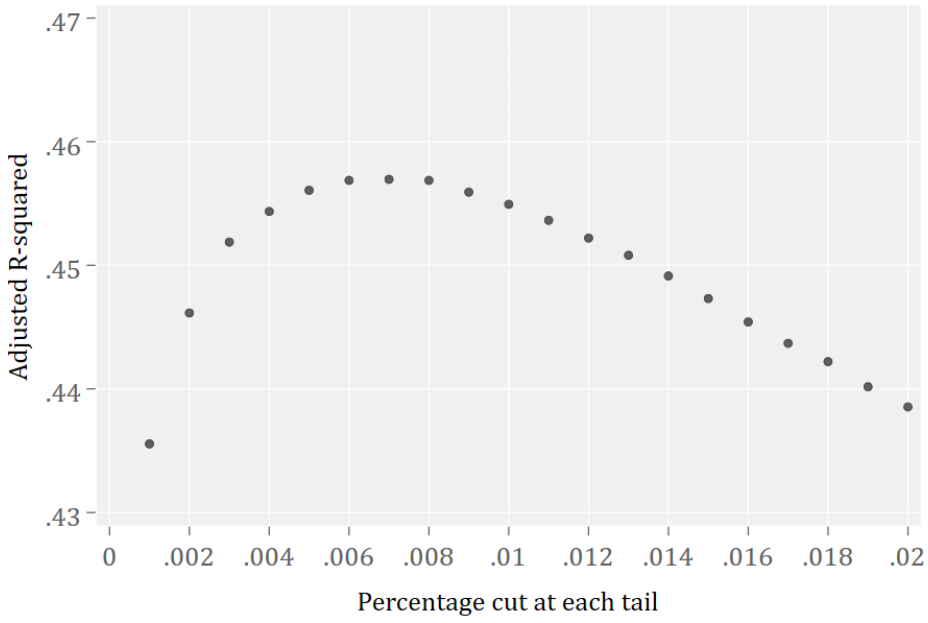
Due to nomenclature changes over time, and to differences between the two samples, parental social classes have to be reduced to 5 different socio-professional categories[4], and neither farmers nor unemployed individuals can be kept in the sample. As farmers (Lefranc et al., 2009) and the unemployed (O'Neill and Sweetman, 1998) tend to have relatively low earnings and to face a strong occupational inheritance, this exclusion may bias final estimates downwards. Finally, as 1990 is one of the few years for which annual declarations of social data have never been treated by the INSEE, data from 1991 are used instead. The prediction model writes as follows.

$$income_i = \alpha + \beta \times gender_i + \gamma \times age_i + \delta \times age_i^2 + \sum_{\phi=1}^{\Phi-1} \eta_\phi \times \mathbb{1}\{j = \phi\} + \sum_{\kappa=1}^{K-1} \theta_\kappa \times \mathbb{1}\{k = \kappa\} + \sum_{\lambda=1}^{\Lambda-1} \mu_\lambda \times \mathbb{1}\{l = \lambda\} + \varepsilon_i, \quad (1)$$

where j denotes the social class and k the place of work of individual i observed during year l.

To prevent extreme values from spuriously inflating the coefficients they are associated with, it is conventional to remove the top and bottom 1% income earners from the distribution. Yet, to avoid arbitrary decisions, Figure 1 shows how the fit of the model evolves with the share of the population cut at each tail of the income distribution.

Figure 1: Model's fit response to tails cut



[4]The resulting nomenclature is consistent with the current PCS-ESE classification of the INSEE. See Table 11 in Appendix for a description of each category.

Estimating the model on the sample as is, without considering any trimming at all, would yield an adjusted R^2 of .365. By simply removing the top and bottom .1%, this indicator jumps to .436, and by trimming more and more the tails of the distribution, the adjusted R^2 concavely increases up to .457 for a .7% cut at each tail, before declining almost linearly. The concavity of the relationship can be explained by the tradeoff between removing enough outliers and starting to get rid of relevant observations. Indeed, the decreasing returns to tails cut up to the maximum value of the adjusted R^2 obtained at .7% probably reflects the fact that theoretically, removing the highest income from the distribution should impact the fit of the model to a higher extent than removing the second-highest income, and so on. Then, the linear decrease of the adjusted R^2 must indicate that the cut at each tail becomes too large, and entails a loss in relevant information. This clearly shows that (1) the major part of the issue related to outlying values is circumvented by getting rid of the top and bottom .1%, and (2) removing the whole top and bottom 1% may not always be optimal. Before estimating the model, top and bottom .7% of the distribution are thus removed.

The results of the OLS estimation of the prediction model are presented in Table 1. The model accurately captures the gender wage gap, the concavity of the age-earnings profile, and social class differentials display a relevant hierarchy from blue collar jobs to executive positions[5]. Yet, relying on parental income predictions can hamper intergenerational elasticity estimations in two ways[6]. First, due to their consequential noise, inaccurate predictions are likely to understate the coefficients. Second, as predictions may prevent from estimating correctly outlying values, by presumably underestimating them, if top-income parents have top-income children, then the IGE would be upward biased, as for this population a fallaciously small increase in parental earnings would be associated with a substantial increase in children's earnings. The potential biases final estimates presented in Section 3.2 may be subject to are discussed in more detail in Section 3.3.

[5]The reference category includes artisans, tradesmen, and entrepreneurs.

[6]Theoretically, naively estimated standard errors are subject to a downward bias as well. But due to the large sample size and the high explanatory power of the estimated prediction model compared to the second stage regressions, the fact that parental income is predicted has a minor impact on standard errors. Indeed, by following the correction procedure presented by Inoue and Solon (2010), the standard errors associated with, for instance, the typical father-son relationship based on the wage variable, must be multiplied by a corrective term whose order of magnitude is $1 + (1.79 \times 10^{-2})$.

Table 1: Parental income prediction model

	Income
Gender:	
Female	−4,846.073***
	(35.922)
Age	1,621.426***
	(8.650)
Age^2	−17.193***
	(.110)
Social class:	
Executive position	4,708.892***
	(197.859)
Intermediary profession	−7,967.18***
	(194.056)
Employee	−14,562.47***
	(193.824)
Blue-collar job	−16,674.42***
	(192.842)
Work department	*Included*
Year	*Included*
Constant	3,610.872
	(6,859.244)
Nb. observations	362,573
Adjusted R^2	.457

Standard errors in parentheses

2.3 Variable definitions

Income can either be that of the individual, or that of the household she belongs to. While Lefranc (2018) uses the former definition, Chetty et al. (2014) favor the latter. In this study estimates are computed for both of these variables, but individual wage is the only variable available for parents. Both parental and children's wages are the annual net salaries and are expressed in 2015 euros.

For both child income and parental income, individuals are ranked by ascending order for each year of observation, and are attributed the number of the percentile of the distribution they belong to[7]. Ranks as defined by Chetty et al. (2014) shall be computed by child cohort, which could be done here for children but not for parents[8]. While ranks within birth cohorts theoretically follow a uniform distribution, using the rank in the whole income distribution of the corresponding year shifts the rank distribution to the right, as the bottom ranks are mainly occupied by people outside the 30 to 55 year-old age range. When dealing with male income, this phenomenon is accentuated as females are also over-represented at the bottom of the distribution.

Yet, as defining ranks the same way for children and parents is certainly crucial, both are computed out of the whole distribution of the corresponding year. But given that age-earnings profiles are most likely monotonically increasing from age 30 to 55 for both parents and children, and that their permanent income are inferred following comparable age restrictions, the fact that parental annual ranks are not computed separately by child cohort should not hamper the estimations of rank-rank correlations.

3 Intergenerational mobility

3.1 Sample restrictions

There are two essential priors to consider before estimating intergenerational correlations: the attenuation bias and the lifecycle bias. The latter is a direct consequence of the concavity of the age-earnings profile. As younger individuals have steeper earnings profiles, early measures of intergenerational persistence could suffer from a downward bias (Haider and Solon, 2006). The attenuation bias, pointed out by Solon (1992), and further documented by Mazumder (2005), arises when the number of income observations is too low to guarantee that their average value

[7]Table 12 in Appendix shows that using the percentile individuals belong to rather than their exact position in the income distribution has no impact on the coefficients estimated in Section 3 up to the third digit, nor on the standard errors up to the fourth digit.

[8]This would by construction reduce the size of the distributions ranks are computed from, and even more as the year of birth of the child is only observed for parents whose revenue is predicted, not for parents from the Employee Panel that constitute the major part of the distribution from which parental rank is computed.

represents a permanent level of income rather than a transitory one. The severity of these biases can be examined graphically, and for the sake of parsimony, only two relations are documented in this subsection: the correlation between the log wage of children and that of their father, i.e., the intergenerational elasticity, used by Lefranc (2018) on French data, and the rank-rank correlation based on children household income, as computed by Chetty et al. (2014) on US data.

A preliminary concern to deal with is the order in which both biases should be evaluated. Starting off with the lifecycle bias is presumably the best option, as it avoids overestimating the severity of the attenuation bias. Indeed, the latter could be accentuated by the fact that the individuals close to the bounds of the selected age range, i.e., close to 30 or 50 years old, have by construction less observations than individuals whose revenues are observed at intermediary ages and who are consequently less impacted by the bias related to the age at the time of the observation. Thus, Section 3.1.1 starts by evaluating the lifecycle bias, Section 3.1.2 analyzes the attenuation bias, and Sections 3.1.3 and 3.1.4 investigate how parental income variables can be subject to these two biases, respectively.

3.1.1 Lifecycle bias

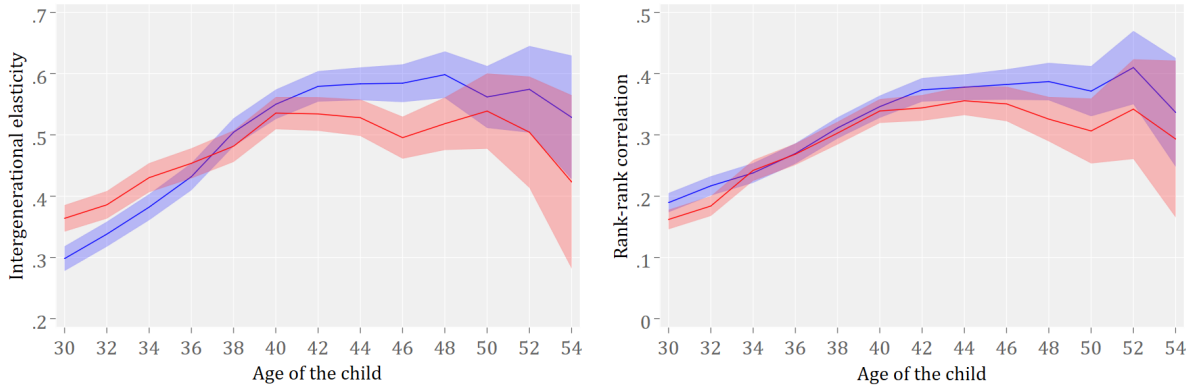
To determine when in the lifetime of an individual it is optimal to observe her income to estimate intergenerational mobility, Figures 2a and 2b show how the main estimators evolve with the age of the child. To reduce noise and increase the number of observations per estimation, for each group of 2 years between 30 and 55 years old a coefficient is computed based on the income of individuals whose ages are comprised in that range. In other words, the coefficient corresponding to age a is actually estimated using all individuals whose earnings are observed from age a to age $a + 1$ in the sample.

As the period considered is rather long, to prevent the relationship from reflecting a time trend or differences between cohorts[9] rather than the lifecycle bias, year of birth is controlled for[10]. The blue line shows the estimations for fathers and sons, the red line for fathers and daughters, and the corresponding shaded areas represent their confidence intervals. Be it for the intergenerational wage elasticity or the household income rank-rank correlation, the relationship seems to stabilize between ages 40 and 50. This is in line with the results of Haider and Solon (2006) for the United States, and of Lefranc (2018) for France, suggesting that the lifecycle bias is minimized around age 40.

[9] See Böhlmark and Lindquist (2006) for evidence on the issue.

[10] Still, as individuals are born between 1962 and 1987 as a direct consequence of the initial selection criteria, and as the availability of fiscal data from 2010 to 2016 allows to control for an eventual trend 7 years by 7 years only, it is not possible to fully rule out the potential implication of a time trend in the relationship.

Figure 2: Lifecycle bias of the child



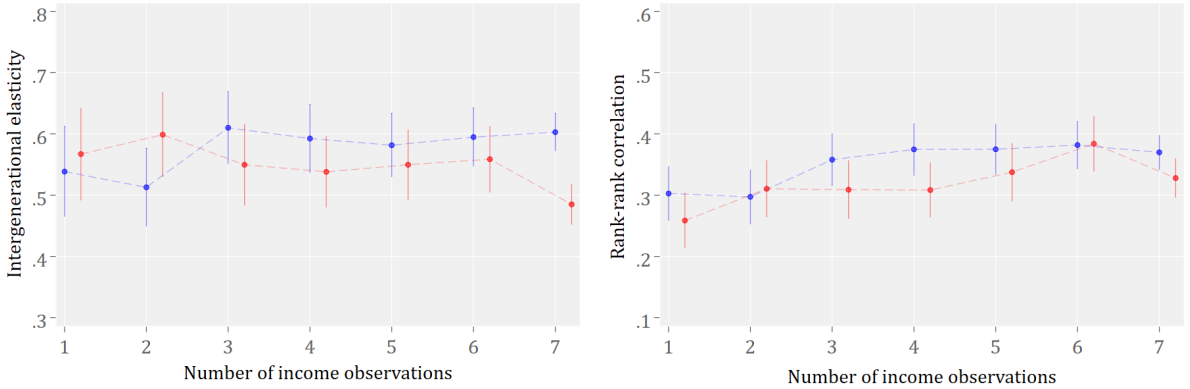
(a) Intergenerational wage elasticity (b) Household income rank-rank correlation

Note: The blue line shows the estimations for fathers and sons, the red line for fathers and daughters, and the corresponding shaded areas represent their confidence intervals. The IGE amounts to .55 when son's wage is observed at age 40-41, and the household income rank-rank correlation amounts to .35 when daughter's income is observed at age 44-45.

3.1.2 Attenuation bias

To produce Figures 3a and 3b, estimates were computed separately for individuals whose fiscal information is observed once, twice, and so on, up to 7 times in the data. To account for the lifecycle bias, only incomes observed while individuals were in their forties are kept.

Figure 3: Attenuation bias of the child



(a) Intergenerational wage elasticity (b) Household income rank-rank correlation

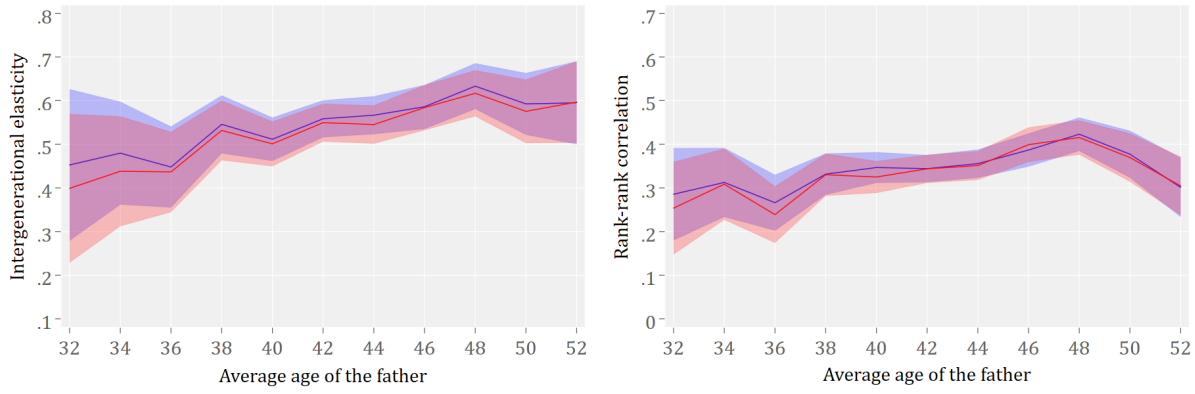
Note: Intergenerational elasticity and rank-rank correlation estimates, and the corresponding confidence intervals, are depicted in blue for father-son specifications, and in red for father-daughter specifications. The IGE amounts to .6 when computed with sons whose wage is observed 3 times in the data, and the rank-rank correlation amounts to .39 when computed with daughters whose household income rank is observed 6 times in the data.

The attenuation bias is salient for sons, but a bit less for daughters. This is probably due to the fact that estimates are less stable with respect to age for daughters than for sons, so that enlarging the number of observations does not impact estimates only directly, but also through age variations. Indeed, Figure 2a shows that the IGE decreases for daughters in their mid forties. By construction, the oldest and youngest daughters of the sample have fewer income observations, that are located at both extremities of the selected age range. Thus, by extending the number of observations as a selection criterion to compute the IGE, daughters whose income is observed in their mid forties are progressively included, and put a downward pressure on the slope that should theoretically have reflected an attenuation bias if the IGE was stable for daughters from age 40 to age 50. But overall, both sons and daughters seem to share a common stability threshold of at least 3 income observations.

3.1.3 Parental lifecycle bias

The age of the individual at income measurement is shown to have an influence on the magnitude of the estimates, and so could the age of her parents when their income is measured. For individuals whose income is observed at least 3 times in their forties, Figures 4a and 4b show how the estimates evolve with respect to the mean of the ages at which father income was predicted to compute parental earnings.

Figure 4: Lifecycle bias of the father



(a) Intergenerational wage elasticity

(b) Household income rank-rank correlation

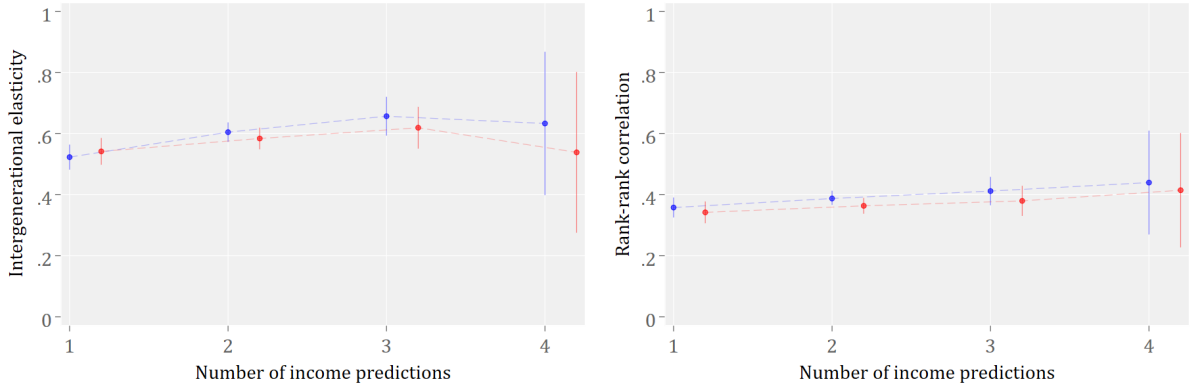
Note: The blue line shows the estimations for fathers and sons, the red line for fathers and daughters, and the corresponding shaded areas represent their confidence intervals. The wage-IGE amounts to .5 between fathers and daughters when fathers' income is observed on average at age 40-41, and the household income rank-rank correlation between fathers and sons amounts to .35 when fathers' income is observed on average at age 44-45.

Once again, to reduce noise and to increase the number of observations per estimation, the coefficient associated with age a was actually estimated using all individuals whose parental mean age at income observation is either a or $a+1$. Despite a weakly positive slope during the forties, this age range seems to be the most appropriate to use for parents as well.

3.1.4 Parental attenuation bias

As parental income can only be predicted at most 4 times from census waves, and as more than 80% of the sample has at most 2 parental income predictions, there is not much room to evaluate the severity of the potential parental attenuation bias. Still, Figures 5a and 5b, which plot the evolution of the estimates computed for the 4 groups defined by the number of parental income predictions, reveal seemingly upward-sloping relationships.

Figure 5: Parental attenuation bias



(a) Intergenerational wage elasticity (b) Household income rank-rank correlation

Note: Intergenerational elasticity and rank-rank correlation estimates, and the corresponding confidence intervals, are depicted in blue for father-son specifications, and in red for father-daughter specifications. The father-son IGE amounts to .6 when computed with fathers whose income is predicted twice, and the father-daughter rank-rank correlation amounts to .4 when estimated with fathers whose income rank is computed 4 times.

As dropping more than 80% of the sample would not allow to estimate conveniently intergenerational mobility coefficients on different subgroups, keeping individuals whose parental earnings are observed at least twice is the only affordable option given the sample size. To check whether or not this sample reduction would significantly improve the accuracy of the estimations, Table 2 shows by how many standard deviations the intergenerational mobility estimates shift when considering only individuals whose parental income is observed at least twice[11].

[11]Table 13 in Appendix displays the estimates blind to the parental attenuation bias.

Table 2: Standard-deviation differences between naive and corrected estimates

	Wage	Hh. income	Wage	Hh. income
	Intergenerational elasticities		Rank-rank correlations	
Father-son	3.59	3.37	3.12	2.77
Father-daughter	1.34	2.41	1.84	2.61
Mother-son	3.49	3.60	4.17	3.81
Mother-daughter	7.26	6.03	9.12	5.27

Note: The father-boy intergenerational elasticity amounts to .602 (s.e. = .013) when estimated blind to the parental attenuation bias (see Table 13 in Appendix), and increases by 3.59 standard deviations, i.e., up to .647, when the sample is restricted to individuals whose parental income is predicted at least twice.

The fact that each of the 16 estimated coefficients increases once the parental attenuation bias is taken into account provides support to the idea that absent this sample restriction, individuals whose parental earnings are observed only once put a downward pressure on the estimates by failing to capture a permanent level of income. Nonetheless, the bias does not impact uniformly all coefficients. Correlations based on variables related to mothers are the most affected, especially when considering daughters' outcomes, with differences in IGE that range from 5.27 to 9.12 standard deviations. When using father income, these differences are smaller in general, especially when considering daughters' outcomes. Even if the severity of the bias was not striking graphically, these differences in relative magnitude between naive and corrected estimates clearly indicate that parental earnings are not immune to the attenuation bias, which is thus accounted for in the computation of national estimates.

3.2 National estimates

Table 3 shows the national intergenerational elasticities and rank-rank correlations for the two previously defined variables of interest, once all the aforementioned biases are taken into account. One could compute each estimate with different sample selection rules according to the specific bias sensitivities and stability bounds associated with each variable, but for these coefficients to be comparable with each other, common selection criteria are applied to the sample according to the general sensitivity of the estimates to the different biases.

The sample is thus restricted to all individuals whose income is observed at least 3 times during their forties and whose parental income could be predicted at least twice between ages 40 and 50 on average. The resulting sample size amounts to about 25,000 observations.

Table 3: National estimates

	Wage	Hh. income	Wage	Hh. income
	Intergenerational elasticities		Rank-rank correlations	
Father-son	.647	.534	.254	.415
	(.015)	(.014)	(.007)	(.012)
Father-daughter	.567	.490	.289	.380
	(.017)	(.016)	(.008)	(.014)
Mother-son	.542	.501	.165	.310
	(.024)	(.021)	(.008)	(.015)
Mother-daughter	.723	.550	.291	.323
	(.024)	(.023)	(.009)	(.016)

Standard errors in parentheses

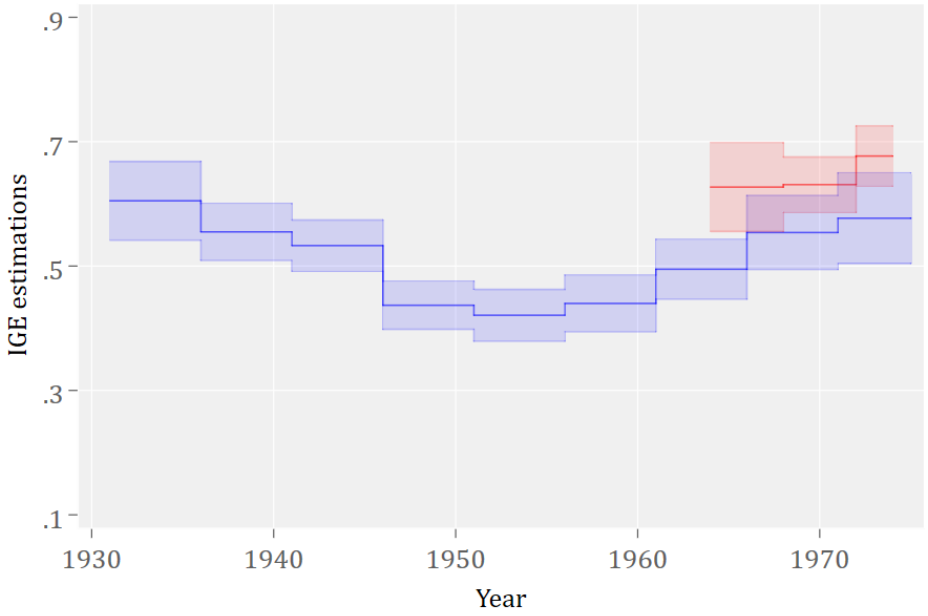
Results suggest that intergenerational elasticities are particularly high between children and parents of the same gender. While cross-gender elasticities are lower than .6, the father-son wage elasticity amounts to .647, and the estimate reaches .723 for mothers and daughters. A 40 year-old woman (resp. man) in France would on expectation earn 7.23% (resp. 6.47%) more if her mother (resp. his father) was earning 10% more at the same age. Thus, intergenerational persistence in France is among the highest in previously studied OECD countries (Acciari et al., 2019; Chetty et al., 2014; Corak, 2013b; Mazumder, 2016)[12].

The father-son estimation is larger than, even though not significantly different from, what Lefranc (2018) estimated using the French FQP dataset[13]. These differences in magnitude probably stem from the fact that, despite a rigorous control for the lifecycle bias that justifies the use of income at age 40, the income variables used by Lefranc (2018) may still not reflect a permanent level of income, as they were not averaged over several years. As he used education (which is in general stable by age 40) rather than socio-professional category to predict parental income, his estimations should be less sensitive to the parental attenuation bias than what is depicted by Figure 5. Yet, as shown in Section 3.1.2, not averaging child income over several years may be a source of downward bias. By splitting the post-selection sample according to birth year, cohort-specific estimations reveal an upward-sloping trend from the mid 60's to the mid 70's. As depicted in Figure 6, this is consistent with the long-run evolution of the IGE documented by

[12]See Figure 1 in Corak (2013b).
[13]See Figure 5 in Lefranc (2018).

Lefranc (2018), and the further results he put forward suggest that this may illustrate a response of the IGE to changes in cross-sectional inequality.

Figure 6: Comparison with Lefranc (2018)



Note: Coefficients and their confidence interval lay over the time-period corresponding to the birth cohort they were computed for. Estimations from Lefranc (2018), for 5-year birth cohorts from 1931-1935 to 1971-1975, are depicted in blue, while estimations conducted on the post-selection sample for cohorts 1964-1967, 1968-1971, and 1972-1974, are depicted in red.

Contrarily to what the results displayed in Table 3 indicate, as fathers earn on average more than mothers, one could have hypothesized that father income would have yielded higher estimates for both sons and daughters than mother income. Indeed, as a larger share of the household income is expected to be earned by the father, and as there is little if any reason for the daughter to benefit less from it than the son, it would seem plausible to guess that father income should be a stronger determinant of children income whatever their gender. Yet, what is observed in the data is more suggestive of a "like father like son", and "like mother like daughter" phenomenon. One reason for that could be that while boys identify more with their father, girls rather identify with their mother (Starrels, 1994), which could later translate into within-gender occupational reproduction. In addition, both children and parental occupation choices are shaped by a certain degree of occupational gender segregation[14]. Thus, as the occupational decision of the individual is likely to be partially framed by gender segregation, and influenced to a greater extent by the occupation of her parent of the same gender, whose occupational choice was probably also subject

[14]The fact that men and women tend to end up in different types of occupations is remarkably robust across settings (Watt, 2010).

to gender segregation, the fact that within-gender IGEs are higher than cross-gender ones is not particularly at odds with the expectations one could have based on the literature.

Rank-rank correlations are also noticeably high. The father-child rank-rank correlation in household income without gender distinction amounts to .399, and the estimate reaches .415 when looking at fathers and sons only. The French social ladder is thus particularly hard to climb compared to what was estimated by Chetty et al. (2014) for the United States. Indeed, they estimated rank-rank correlations in household income that lie around .34[15], which is itself higher than in Italy (Acciari et al., 2019), Canada, or Denmark[16], for instance. Yet, more recent results obtained by Mazumder (2016) suggest that Chetty et al. (2014)'s coefficients are underestimated because of the age structure and the limited panel dimension of their data, and that corrected rank-rank correlations would rather be close to .4. Thus, it seems that the French level of intergenerational persistence is relatively close to that observed in the United States.

Finally, while intergenerational elasticities are higher between fathers and sons than between fathers and daughters, rank-rank correlations show opposite patterns. This peculiarity presumably comes from the sensitivity of intergenerational elasticities to extreme values. Indeed, removing the top and bottom 1% of the income distribution of the sample is sufficient to decrease the father-boy estimate down to a point at which it is no longer significantly different from the father-girl one, as shown by Table 14 in Appendix. This is probably due to a strong intergenerational persistence in the top 1% of the post-selection sample, in which 84% of individuals are men.

3.3 Robustness

Despite all the attention paid to the enfeeblement of the biases inherent to the study of intergenerational mobility, some data-specific biases may still remain. For instance, as parental income predictions presumably underestimate large outlying values, if top-income parents have top-income children, the IGE would probably be upward-biased as for this population a fallaciously small increase in parental earnings would be associated with a substantial increase in children's earnings. The natural way to test this hypothesis is to remove the top and bottom 1% of the parental income distribution from the sample. Table 15 in Appendix shows how the results displayed in Table 3 are impacted by this change. Even though most estimates decrease, the difference never even reaches 1 standard deviation. Thus, if this potential upward bias does have an impact, it is arguably negligible.

[15]The standard error associated with the French coefficient amounts to .0092, and that of Chetty et al. (2014) on US data amounts to .0003. Yet, local estimations at the commuting zone level in the United States can exceed the national estimate for France, especially in the South-East of the country.

[16]See Figures 2 (a) and 2 (b) in Chetty et al. (2014).

The other potential bias induced by the fact that parental income is predicted goes in the opposite direction. Indeed, as any noisy measurement, the inevitable imprecision of the predictions would theoretically bias the estimates towards zero. As no data linking an individual's earnings to their parents' were yet collected in France, evaluating the severity of this potential bias is out of reach for the time being, but the high correlation between parental and child's income suggests that parental income predictions are likely to be reasonably accurate, and the downward bias relatively small.

A more problematic potential source of downward bias is the exclusion of both farmers and the unemployed, two categories of individuals that tend to be associated with relatively low earnings and to face a strong occupational inheritance (Lefranc et al., 2009; O'Neill and Sweetman, 1998). This is probably the main limitation of these estimations, and this highlights the fact that the collection of data linking earnings of French individuals to their parents' earnings would be highly valuable for research advances on the issue.

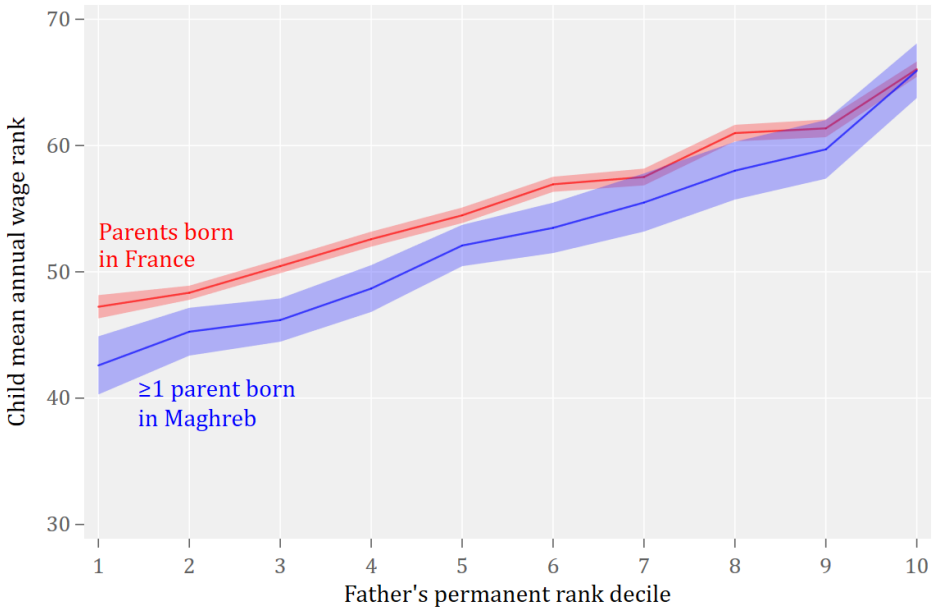
The last source of downward bias comes from the fact that the sample size and scope do not allow to fully tackle the parental attenuation bias. According to Figures 5a and 5b, a minimum of 2 observations is not enough to compute a permanent, as opposed to transitory, level of income, and more comprehensive data sources would be required to fully account for that. Thus, these two remaining potential sources of bias that could hamper the presented estimates are all theoretically threatening to understate their magnitude. All the intergenerational persistence coefficients estimated in this analysis should consequentially only be seen as lower bounds.

3.4 Ethnic inequalities

The strong intergenerational persistence documented in the previous section could be a reason why the offspring of natives earns on average more than individuals with foreign-born origins: by starting at a lower level in the socioeconomic ladder at birth, second-generation immigrants would on average end up with lower outcomes once on the labor market. This section aims at evaluating whether observed differences in labor market outcomes between the offspring of natives and second-generation immigrants from Maghreb can entirely be explained by intergenerational persistence itself, and if not, to identify the plausible underlying channels of this phenomenon[17]. Figure 7 shows the average income rank of individuals for each parental income decile depending on whether both parents were born in France or at least one was born in Maghreb.

[17]To reach a reasonably large sample size of second-generation immigrants, the age restriction defined in the previous subsection is loosened to 30 to 55 year-old individuals. Given the closeness of the age distributions of both groups of individuals documented in Table 17 in Appendix, this should not have a differentiated impact on the resulting intergenerational persistence estimations, but the parental sample restrictions and the minimum of 3 income observations for children are still maintained to limit other potential sources of bias.

Figure 7: Average income rank conditional on parental income decile



Except for the last parental income decile, second-generation immigrants from Maghreb have systematically lower expected ranks in the income distribution even by conditioning on parental income. In other words, by starting at comparable socioeconomic levels during childhood, second-generation immigrants from Maghreb still end up lower in the social ladder than children of native parents. Table 4 quantifies the different aspects of the ethnic gap elicited in Figure 7.

Table 4: Rank-rank correlation & parental origin

	Child income rank			
Father income rank	.323***		.322***	.316***
	(.006)		(.006)	(.006)
≥ 1 parent born in Maghreb		−3.08***	−2.753***	−6.93***
		(.343)	(.333)	(1.324)
Interaction				.062***
				(.019)
Constant	33.675***	55.895***	34.006***	34.441***
	(.407)	(.108)	(.408)	(.43)

Nb obs: 53,350 - *Standard errors in parentheses*

In the first column, the coefficient associated with the parental income rank variable, i.e., the rank-rank correlation, has the same interpretation as in Table 3 but is estimated on individuals of both genders whose parental place of birth is observed. The coefficient in the second column is the unconditional ethnic gap. Individuals for whom at least one parent is born in Maghreb are on average located 3 percentile ranks below children of French natives. In the last column, the coefficient associated with the interaction between parental rank and parental place of birth encapsulates the difference in the extent to which parental income matters between individuals for whom at least one parent is born in Maghreb and children of French natives, i.e., the difference in relative mobility. It corresponds to the difference in the steepness of the fitted regression lines of the two curves depicted in Figure 7.

These results corroborate the graphical evidence depicted in Figure 7, namely that second-generation immigrants from Maghreb are not only located at lower ranks of the income distribution unconditionally, but also for a given level of parental income rank, and that intergenerational persistence is also stronger for them than for children of French natives. Table 16 in Appendix elicits the same patterns based on intergenerational elasticity estimations rather than rank-rank correlations.

Figures 17a to 17f in Appendix investigate the heterogeneity of this relationship with respect to gender, employment sector, and self-employment status. No clear heterogeneity pattern is observed between males and females nor between employees of the public and the private sector. But while the conditional ethnic gap is also observed for the subsample of employed individuals, it does not seem to hold for self-employed individuals. Yet, this absence of significant discrepancy could possibly be attributed to the low number of observations, as self-employed individuals account for barely 4% of the sample.

To investigate the potential mechanisms at stake, the following four figures show the average of various outcomes separately for children with French parents and children with at least one parent born in Maghreb, conditional on parental income decile.

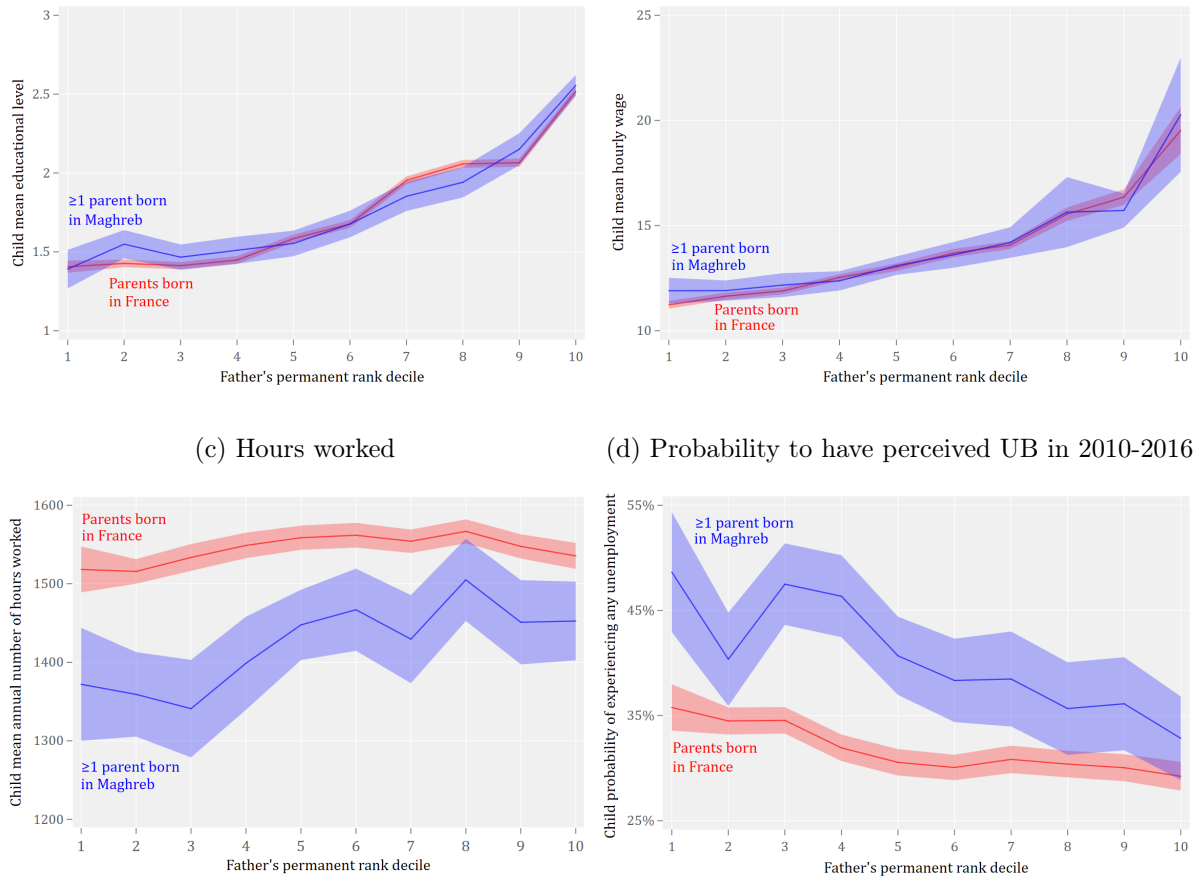
Figure 8a indicates that conditional on parental earnings, children of parents born in France and in Maghreb have very comparable educational outcomes[18], and according to Figure 8b, very similar hourly wage levels as well. These findings discredit the hypotheses of differences in human capital investments and in ability/productivity to explain the robustness of the ethnic gap to the conditioning on parental earnings. The phenomena depicted on Figures 8c and 8d rather

[18]Educational level is coded on a scale from 1 to 4, corresponding respectively to lower than high-school level, high-school level or equivalent, undergraduate level, and postgraduate level.

Figure 8: Average economic outcomes conditional on parental income decile

(a) Educational level

(b) Hourly wage



(c) Hours worked

(d) Probability to have perceived UB in 2010-2016

advocate differences in terms of access to employment. Indeed, the number of hours worked is systematically lower and the probability to have perceived unemployment benefits over the period studied is systematically larger for second-generation immigrants from Maghreb. As similar educational investments are observed for both groups, these gaps are arguably not likely to result from differentiated preferences in terms of labor provision. These findings echo the well-documented hiring discrimination towards applicants of North-African origin observed on the French labor market (Adida et al., 2016; Cahuc et al., 2019; Edo et al., 2019; Silberman et al., 2007). To my knowledge, no dataset that would allow to quantify the implications of hiring discrimination in the difference between economic opportunities of second-generation immigrants and individuals whose parents are native yet exists. But hiring discrimination may not be the only factor that plays a role in the relationship between ethnic origins and inequality of opportunity. Indeed, residential segregation was shown by Chetty and Hendren (2018b) to be associated with lower rates of upward mobility, and further results suggest that it would play a role in the perpetuation of the Black-White gap in the US (Chetty et al., 2020). Because of data restrictions

on ethnic variables, the current estimations of segregation in France are too sparse for their spatial variations to be exploited so as to investigate the relationship between intergenerational mobility, origins, and segregation. As motivated earlier, these limitations can be circumvented by relying on an alternative source of information on an individual's origin, her last name.

4 Spatial segregation

4.1 Predicting origins from last names

The methodology I consider consists in comparing the sub-sequences of letters forming the name whose origin has to be determined to the sub-sequences of letters appearing in a corpus of Arabic names and in a corpus of French names. How a name matches a given corpus should translate into a probability ideally satisfying the following 5 axioms.

(1) The probability that a name belongs to a given origin should lie between 0 and 1;

(2) The probability that a name belongs to a given origin should be equal to 1 if and only if it is included in the corpus of the corresponding origin;

(3) A longer sub-sequence of letters should have a larger impact on the probability;

(4) The more frequent a sub-sequence is in a given corpus, the larger should be the probability that a name containing this sub-sequence belongs to the corresponding origin;

(5) The length of a name should not have a systematic impact on value of the probability.

For exposition purposes, the description of the algorithm will be illustrated using first names. The second axiom can be applied strictly by considering that Côme does not belong to the corpus {Pierre, Pacôme}, and in a looser way that would attribute a probability 1 for the name Côme to belong to that corpus. I consider the strict version of this axiom as it allows to give a larger weight to the first and the last letter of a name, which may be more informative than letters that are located at intermediate positions in the name. For instance, the letter A at the end of a name may indicate with more certainty a potential Arabic origin if the letter A is more frequently the last letter of Arabic names than of French names. In other words, a name that ends with the letter A should be assigned a higher probability to belong to an origin whose corresponding corpus contains names that systematically end in A than if the letter A is not over-represented in the last position. To account for that, I add to each name the prefix ¿ and the suffix ?, which converts the problem Côme $\overset{?}{\in}$ {Pierre, Pacôme} into ¿Côme? $\overset{?}{\in}$ {¿Pierre?, ¿Pacôme?}.

Suppose a name $\mathcal{M}$ made of n_c characters. Because of the characters ¿ and ? indicating the beginning and the end of the sequence of letters, the name contains sub-sequences of letters of length $j \in \{1, n_c + 2\}$. Let n_j be the number of sub-sequences of length j, $s_{i,j}$ denotes the sub-sequence $i \in \{1, n_j\}$ in the group of sub-sequences of length j. Note that $n_j = n_c - j + 3$ $\forall\ j \neq 1$, and that $n_{j=1} = n_c$ as the characters ¿ and ? are not considered as sub-sequences of length 1. The most straight-forward way to satisfy the first three axioms consists in summing on the whole set of sub-sequences in the name a binary operator that takes the value 1 if the sub-sequence belongs to the corresponding corpus $\mathcal{C}$, weighted by the length of the sub-sequence relative to that of the whole name:

$$\Gamma_1 = \sum_{j=1}^{n_c+2} \left[\frac{j}{\sum j} \times \left(\sum_{i=1}^{n_j} \frac{1}{n_j} \times \mathbb{1}\{s_{i,j} \in \mathcal{C}\} \right) \right] \tag{2}$$

Table 5 lists the values taken by the different components of the index Γ_1 for each sub-sequence of the name Louis. The value of $\omega(s_{i,j})$ corresponds to the overall impact of the presence of the sub-sequence $s_{i,j}$ on the value of the index. For instance, if one of the names in the corpus contains the sub-sequence L, this would increase the probability that the name Louis belongs to the corresponding origin by .7 percentage points, while the occurrence of the sub-sequence $OUIS$ in the corpus would increase the probability by 3.6 percentage points directly as a sub-sequence of 4 characters, and by 10.7 additional percentage points *via* the occurrence of the shorter sub-sequences it contains in the corpus, OUI, UIS, OU, UI, IS, O, U, I, S, for a total impact of 14.3 percentage points on the index.

Table 5: Decomposition of the index Γ_1 according to the sub-sequences of the name Louis

j	$s_{1,j}$	$s_{2,j}$	$s_{3,j}$	$s_{4,j}$	$s_{5,j}$	$s_{6,j}$	n_j	$j/\sum j$	$1/n_j$	$\omega(s_{i,j})$
1	L	O	U	I	S		5	1/28	1/5	.007
2	¿L	LO	OU	UI	IS	S?	6	2/28	1/6	.012
3	¿LO	LOU	OUI	UIS	IS?		5	3/28	1/5	.021
4	¿LOU	$LOUI$	$OUIS$	UIS?			4	4/28	1/4	.036
5	¿$LOUI$	$LOUIS$	$OUIS$?				3	5/28	1/3	.060
6	¿$LOUIS$	$LOUIS$?					2	6/28	1/2	.107
7	¿$LOUIS$?						1	7/28	1	.25

This table notably stresses that the weight associated with the full sub-sequence, i.e., the name itself, is so large that it generates a discrete jump in the probability associated to a name that belongs to the corpus relatively to a name that is only different by a single character. Thus, it seems preferable to omit the complete sub-sequence of letters by considering the more reasonable following index:

$$\Gamma = \sum_{j=1}^{n_c+1} \left[\frac{j}{\sum j} \times \left(\sum_{i=1}^{n_j} \frac{1}{n_j} \times \mathbb{1}\{s_{i,j} \in \mathcal{C}\} \right) \right] \tag{3}$$

The equivalent of Table 5 for this second index is given in Table 18 in Appendix. Even though Γ seems more appropriate than Γ_1 in terms of weight distribution, it requires to depart slightly from the second axiom. Indeed, according to Γ the probability would be equal to 1 if and only if the name is included in the corpus, or if it both starts and ends a name included in the corpus. For instance, if a corpus contains the name Yahya, as the sequence "Ya" both starts and ends the name Yahya, the Γ value of "Ya" would be equal to 1 even if it is not in the corpus *per se*. This modification of the second axiom could yield erroneous conclusions for some very short names, but is likely not to cause significant issues for the vast majority of names.

A limit of this first type of indices is that the mere inversion of two letters or the removal of a single letter can drastically reduce the value of the probability, especially if the change takes place at the middle of the name. One way to palliate this problem consists in considering the potential permutations of characters that form the name. For instance, if a name of 7 characters has 5 common letters with a name of 8 characters in the corpus, the Γ index could be raised to the power $\left[1 - (\frac{5}{7} \times \frac{5}{8})\right]$ so as to increase the probability that the name belongs to the origin of that corpus even if the location of the characters do not match perfectly one of the traditional names it includes. Indeed, as $\Gamma \in [0, 1]$, the index $\Gamma^{(1-\Lambda)}$ with $\Lambda \in [0, 1]$ increases in Λ, and always lies in the interval $[0, 1]$. Yet, defining Λ as suggested above would only account for the share of common letters, not for how they are ordered in the name. It is obviously not desirable that any random permutation of the letters constituting a name in the corpus is attributed a probability 1 to belong to the corresponding origin. Thus, in order to quantify the similarities in terms of placement for the letters that are common to a name A and a name B, I consider the ratio of the number of letters $s^A_{i,1}$ of the first name whose preceding letter $s^A_{i-1,1}$ is also the preceding letter $s^B_{j-1,1}$ of the initial letter in name B ($s^A_{i,1} = s^B_{j,1}$), or whose subsequent letter $s^A_{i+1,1}$ is also the subsequent letter $s^B_{j+1,1}$ of the initial letter in name B, over the total number of letters in the

name A. For instance, in the name Louis, the letter L is followed by an O as in Clovis, the letter O is preceded by an L, the letter I is followed by an S, and the letter S is in the last position. Thus, 4 letters out of the 5 that make the name Louis are located at similar places to where they are located in the name Clovis. The index Λ that quantifies similarities between Louis and Clovis accounting for the number of letters they share relative to their respective length and how close their placement is in these two names can then write $(\frac{4}{5} \times \frac{4}{6} \times \frac{4}{5})$. According to this definition, the random permutation Iluso of the name would not share any similarity in terms of location of letters with the name Louis, such that the exponent $(1 - \Lambda)$ would take the value $(1 - \frac{5}{5} \times \frac{5}{5} \times \frac{0}{5}) = 1$, and its initial probability would not be raised by the presence of Louis in the corpus. In practice, that value of Λ should be computed for each name in the corpus so that the highest one is considered for the calculation of the index. Formally,

$$\Lambda = \max_{\mathcal{M}^{\mathcal{C}} \in \mathcal{C}} \left\{ \frac{\sum_{i=1}^{n_c} \mathbb{1}\{s_{i,1} \in \mathcal{M}^{\mathcal{C}}\}}{n_c} \times \frac{\sum_{i=1}^{n_c^{\mathcal{M}^{\mathcal{C}}}} \mathbb{1}\{s_{i,1}^{\mathcal{M}^{\mathcal{C}}} \in \mathcal{M}\}}{n_c^{\mathcal{M}^{\mathcal{C}}}} \times \frac{\sum_{i=1}^{n_c} \mathbb{1}\left\{\exists (s_{i-1,1} \cup s_{i+1,1}).\left[s_{i-1,1} = s_{k-1,1}^{\mathcal{M}^{\mathcal{C}}} \cup s_{i+1,1} = s_{k+1,1}^{\mathcal{M}^{\mathcal{C}}}\right] \ni (k : s_{i,1} = s_{k-1,1}^{\mathcal{M}^{\mathcal{C}}})\right\}}{n_c} \right\} \quad (4)$$

The fourth axiom to account for states that the more frequent a sub-sequence is in a given corpus, the larger should be the probability that a name containing this sub-sequence belongs to the corresponding origin. It is notably desirable, *ceteris paribus*, (1) that the probability increases as the sub-sequences the name contains appear more frequently in the corresponding corpus $\mathcal{C}^o$, (2) that the probability decreases as they appear more frequently in the other corpus $\mathcal{C}^{-o}$, (3) that the probability is not affected by a sub-sequence whose frequency is the same in both corpora, and (4) that the impact of the frequency of a sub-sequence in a corpus is proportional to its number of characters. This can be achieved by simply raising the probability $\Gamma^{(1-\Lambda)}$ to the power Δ, which writes:

$$\Delta = \frac{\sum_{j=1}^{n_c+1} \left[\frac{j}{\sum j} \times \left(\sum_{i=1}^{n_j} \frac{1}{n_j} \times \mathbb{1}\{s_{i,j} \in \mathcal{C}^{-o}\} \times \frac{\#s_{i,j}^{\mathcal{C}^{-o}}}{n_j} \Big/ \max_l \left\{ \frac{\#s_{l,j}^{\mathcal{C}^{-o}}}{n_j} \right\} \right) \right]}{\sum_{j=1}^{n_c+1} \left[\frac{j}{\sum j} \times \left(\sum_{i=1}^{n_j} \frac{1}{n_j} \times \mathbb{1}\{s_{i,j} \in \mathcal{C}^{o}\} \times \frac{\#s_{i,j}^{\mathcal{C}^{o}}}{n_j} \Big/ \max_m \left\{ \frac{\#s_{m,j}^{\mathcal{C}^{o}}}{n_j} \right\} \right) \right]} \quad (5)$$

In that way, if every sub-sequence $s_{i,j}$ of a name $\mathcal{M}$ has the same frequency in the groups of

sub-sequences of length j of each corpus, the probability $\Gamma^{(1-\Lambda)}$ would not be affected by raising it to the power Δ. If the sub-sequences occur more often in the corresponding corpus $\mathcal{C}^o$ than in the other corpus $\mathcal{C}^{-o}$, Δ would be lower than 1, and as $\Gamma^{(1-\Lambda)} \in [0,1]$, the probability would increase. Naturally, when the sub-sequences are more frequent in the other corpus, Δ would be larger than 1, what would decrease the probability that the name belongs to the same origin as the names in the corpus $\mathcal{C}^o$.

The last axiom to consider states that the length of a name should not systematically affect the value of the probability *ceteris paribus*. This axiom is particularly relevant for names that are a juxtaposition of two single names. Indeed, while the approach followed so far consists in testing whether it is plausible that a name belongs to a given corpus, the relevant question for this very type of names is whether it is plausible that the names in a given corpus belong to the name whose origin has to be predicted. To design such an indicator, I first consider the $1+\sum_{i=1}^{n_c-1} 2^{(i-1)}$ juxtapositions of sub-sequences of characters that form the name, as illustrated in Table 6.

Table 6: List of the juxtapositions of sub-sequences of letters that form the name Louis

$\#s_{i,j}^d$	Juxtapositions of $s_{i,j}$ that form Louis					
1	$LOUIS$					
2	$L \cdot OUIS$	$LO \cdot UIS$	$LOU \cdot IS$	$LOUI \cdot S$		
3	$L \cdot O \cdot UIS$	$L \cdot OU \cdot IS$	$L \cdot OUI \cdot S$	$LO \cdot U \cdot IS$	$LO \cdot UI \cdot S$	$LOU \cdot I \cdot S$
4	$L \cdot O \cdot U \cdot IS$	$L \cdot O \cdot UI \cdot S$	$L \cdot OU \cdot I \cdot S$	$LO \cdot U \cdot I \cdot S$		
5	$L \cdot O \cdot U \cdot I \cdot S$					

Let $\mathcal{D}$ be the set of disaggregations d whose juxtaposition of sub-sequences $s_{i,j}$ forms the name $\mathcal{M}$. The index Λ can be transposed from the name level to the sub-sequence level so as to compute for each sub-sequence $s_{i,j}$ of the name $\mathcal{M}$ the index value given by its closest match with a sub-sequence $s_{i,j}^{\mathcal{C}}$ from the corpus.

$$\Lambda(s_{i,j},\mathcal{C}) = \max_{s_{i,j}^{\mathcal{C}} \in \mathcal{C}} \left\{ \frac{\sum_{i=1}^{n_c^{s_{i,j}}} \mathbb{1}\{s_{i,1} \in s_{i,j}^{\mathcal{C}}\}}{n_c^{s_{i,j}}} \times \frac{\sum_{i=1}^{n_c^{s_{i,j}^{\mathcal{C}}}} \mathbb{1}\{s_{i,1}^{s_{i,j}^{\mathcal{C}}} \in s_{i,j}\}}{n_c^{s_{i,j}^{\mathcal{C}}}} \times \right.$$
$$\left. \frac{\sum_{i=1}^{n_c^{s_{i,j}}} \mathbb{1}\left\{\exists(s_{i-1,1} \cup s_{i+1,1}).\left[s_{i-1,1} = s_{k-1,1}^{s_{i,j}^{\mathcal{C}}} \cup s_{i+1,1} = s_{k+1,1}^{s_{i,j}^{\mathcal{C}}}\right] \ni (k : s_{i,1} = s_{k-1,1}^{s_{i,j}^{\mathcal{C}}})\right\}}{n_c^{s_{i,j}}} \right\} \quad (6)$$

In that way, every sub-sequence $s_{i,j}$ of a disaggregations d can be associated to a score $\Lambda(s_{i,j}, \mathcal{C})$, so that the score associated to d can be computed as the sum of the scores of each of its sub-sequences weigthed by the ratio of the length of the sub-sequence over that of the name. The highest score associated with a disaggregations $d \in \mathcal{D}$ whose juxtaposition of sub-sequences $s_{i,j}$ forms the name $\mathcal{M}$ writes:

$$\Omega_1 = \max_{d \in \mathcal{D}} \left\{ \sum_{i=1}^{\#s_{i,j}^d} \frac{j}{n_c} \Lambda(s_{i,j}, \mathcal{C}) \right\}, \tag{7}$$

where $\#s_{i,j}^d$ is the number of sub-sequences in the disaggregations d. Note that without imposing any restriction on Ω_1, the algorithm could maximize the score $\Lambda(s_{i,j}, \mathcal{C})$ by dividing the name into a succession of very short sub-sequences to be matched with very short names in the corpus. As this would yield fallaciously high probabilities for long names, it seems appropriate to restrain Ω_1 to use no more than two names from the corpus to explain the name whose origin has to be determined, and to explain only sub-sequences of more the two letters. The resulting index Ω should then be raised to the Δ to account for the fourth axiom.

The final estimator writes:

$$\mathcal{P}(\mathcal{M} \in \mathcal{C}^o) = \max\{\Gamma^{(1-\Lambda)}, \Omega\}^{\Delta}, \tag{8}$$

with

$$\Gamma = \sum_{j=1}^{n_c+1} \left[\frac{j}{\sum j} \times \left(\sum_{i=1}^{n_j} \frac{1}{n_j} \times \mathbb{1}\{s_{i,j} \in \mathcal{C}^o\} \right) \right],$$

$$\Lambda = \max_{\mathcal{M}^{\mathcal{C}^o} \in \mathcal{C}^o} \left\{ \frac{\sum_{i=1}^{n_c} \mathbb{1}\{s_{i,1} \in \mathcal{M}^{\mathcal{C}^o}\}}{n_c} \times \frac{\sum_{i=1}^{n_c^{\mathcal{M}^{\mathcal{C}^o}}} \mathbb{1}\{s_{i,1}^{\mathcal{M}^{\mathcal{C}^o}} \in \mathcal{M}\}}{n_c^{\mathcal{M}^{\mathcal{C}^o}}} \times \right.$$
$$\left. \frac{\sum_{i=1}^{n_c} \mathbb{1}\left\{ \exists (s_{i-1,1} \cup s_{i+1,1}). \left[s_{i-1,1} = s_{k-1,1}^{\mathcal{M}^{\mathcal{C}^o}} \cup s_{i+1,1} = s_{k+1,1}^{\mathcal{M}^{\mathcal{C}^o}} \right] \ni (k : s_{i,1} = s_{k-1,1}^{\mathcal{M}^{\mathcal{C}^o}}) \right\}}{n_c} \right\},$$

$$\Omega = \max_{d \in \mathcal{D}} \left\{ \frac{j_q}{n_c} \Lambda(s_{q,j}, \mathcal{C}^o) + \frac{j_r}{n_c} \Lambda(s_{r,j}, \mathcal{C}^o) \ \forall \ q, r \in i = \{1, \#s_{i,j}^d\} \ \middle| \ j_i > 2 \right\},$$

$$\Lambda(s_{i,j}, \mathcal{C}^o) = \max_{s_{i,j}^{\mathcal{C}^o} \in \mathcal{C}^o} \left\{ \frac{\sum_{i=1}^{n_c^{s_{i,j}}} \mathbb{1}\{s_{i,1} \in s_{i,j}^{\mathcal{C}^o}\}}{n_c^{s_{i,j}}} \times \frac{\sum_{i=1}^{n_c^{s_{i,j}^{\mathcal{C}^o}}} \mathbb{1}\{s_{i,1}^{s_{i,j}^{\mathcal{C}^o}} \in s_{i,j}\}}{n_c^{s_{i,j}^{\mathcal{C}^o}}} \times \right.$$
$$\left. \frac{\sum_{i=1}^{n_c^{s_{i,j}}} \mathbb{1}\left\{\exists(s_{i-1,1} \cup s_{i+1,1}).\left[s_{i-1,1} = s_{k-1,1}^{s_{i,j}^{\mathcal{C}^o}} \cup s_{i+1,1} = s_{k+1,1}^{s_{i,j}^{\mathcal{C}^o}}\right] \ni (k : s_{i,1} = s_{k-1,1}^{s_{i,j}^{\mathcal{C}^o}})\right\}}{n_c^{s_{i,j}}} \right\},$$

$$\Delta = \frac{\sum_{j=1}^{n_c+1} \left[\frac{j}{\sum j} \times \left(\sum_{i=1}^{n_j} \frac{1}{n_j} \times \mathbb{1}\{s_{i,j} \in \mathcal{C}^{-o}\} \times \frac{\#s_{i,j}^{\mathcal{C}^{-o}}}{n_j} \Big/ \max_l \left\{ \frac{\#s_{l,j}^{\mathcal{C}^{-o}}}{n_j} \right\} \right) \right]}{\sum_{j=1}^{n_c+1} \left[\frac{j}{\sum j} \times \left(\sum_{i=1}^{n_j} \frac{1}{n_j} \times \mathbb{1}\{s_{i,j} \in \mathcal{C}^{o}\} \times \frac{\#s_{i,j}^{\mathcal{C}^{o}}}{n_j} \Big/ \max_m \left\{ \frac{\#s_{m,j}^{\mathcal{C}^{o}}}{n_j} \right\} \right) \right]}.$$

4.2 Validity of the predictions

Until very recently, the link between one's name and ethnic origins was given relatively little attention by the French literature. There are still two studies whose results can be used as a benchmark to test the validity of the predictions.

First, Fourquet and Manternach (2019) quantify the share of Arabic first names given to newborns in France using the publicly available first names file (*Fichier des prénoms*) of the INSEE. Their methodology relies on a manual classification of names between an Arabic and non-Arabic origin. As the algorithm can be applied either to first names or to last names as long as the comparison corpora are suitably adapted, a first test can consist in applying the algorithm to the first names of INSEE's first names file, and to compare the evolution of the share of Arabic first names among newborns in France whose origin is attributed manually by Fourquet and Manternach (2019), to that resulting from the classification by the algorithm.

To my knowledge, no such statistics were yet computed directly for last names in France, as this variable is generally absent from diffusion datasets. Yet, as the electoral registers contain both the last name and the first name of voters, a second test can consist in comparing the origin the algorithm attributes to the first name and to the last name of voters. French electoral registers are neither yet centralized nor always digitized, but those of the Paris city-department still contain data on more than one million voters.

Finally, Coulmont and Simon (2019) use the survey *Trajectoires et Origines* (TeO) to study the most frequent first names of immigrants in France, conditionally on country of origin and on generation of immigration. As the electoral registers also contain country birth, a last test of whether the algorithm indeed identifies the population of interest can consist in comparing the most frequent first names among second-generation immigrants whose origin is obtained by survey data (TeO) to the most frequent first names among people born in France whose origin is inferred from their last name by the algorithm.

4.2.1 Comparison to Fourquet and Manternach (2019)

To estimate the proportion of Arabic first names among newborns in France and compare it to the results of Fourquet and Manternach (2019)[19], two corpora of first names should be built. These first names should be "root" names, anchored in the culture of the corresponding origin, and most importantly more subject to serve as a basis for derivative names rather than to be derivative names themselves. For the definition of the corpora to be the most objective, I constitute the corpus of French names based on the French Christian calendar, and I gather names that are transliterations of common words in Arabic for the corpus of Arabic names. As female first names tend to be less distinctive in their cultural affiliations, just as Fourquet and Manternach (2019) I will only consider male first names. I exclude the first name Habib from the 245 male first names of the Christian calendar as it is also a word in Arabic meaning *beloved*. I also exclude first names that are not historically French such as Ulrich, Igor, or Donald. To build the corpus of Arabic names, I gather the first names from the excerpt of the list used by Fourquet and Manternach (2019) that the authors assented to share, and the names listed on the Wikipedia page of Arabic first names[20]. Out of these more than 600 names, only those whose presumed origin is not ambiguous and that are actual words commonly used in Arabic are kept[21]. The final corpora contain 210 Arabic first names and 214 French first names.

In combination with these two corpora, the algorithm is applied to every male first name of INSEE's first names file. Virtually every male first name given in France since 1900[22] is thus associated with a probability of French origin and a probability of Arabic origin. I distinguish 5 zones of equal area on the domain of the joint distribution of these two probability variables. Figure 9 plots on this domain each of the 16,269 male first names observed in the data.

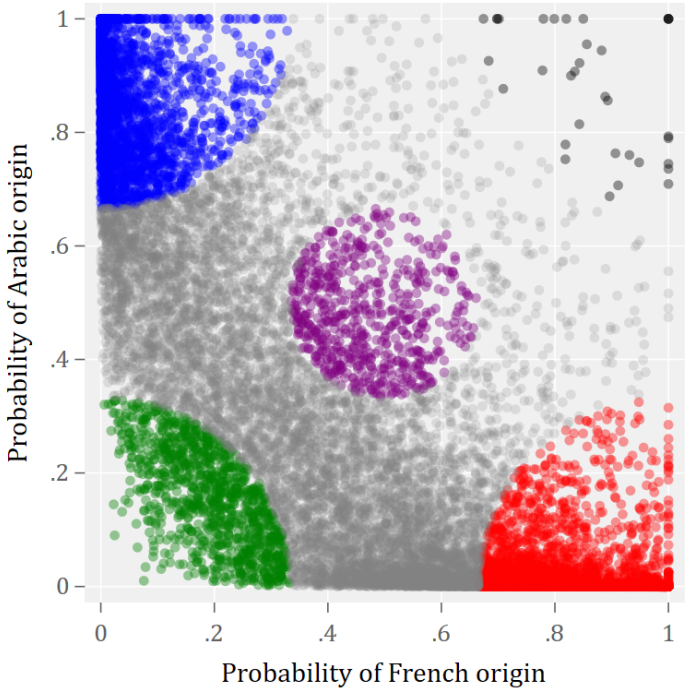
[19] A reappraisal of Fourquet and Manternach (2019)'s study based on the algorithm developed in this section is proposed in Appendix.

[20] *https://fr.wikipedia.org/wiki/Liste_de_prénoms_arabes*

[21] To ensure that names in this final list meet this definition, it was proofread by a native Arabic speaker. Each of the corpora used in this Master's Thesis is available upon request.

[22] The file is exhaustive from 1946.

Figure 9: Classification of the 16,269 male first names



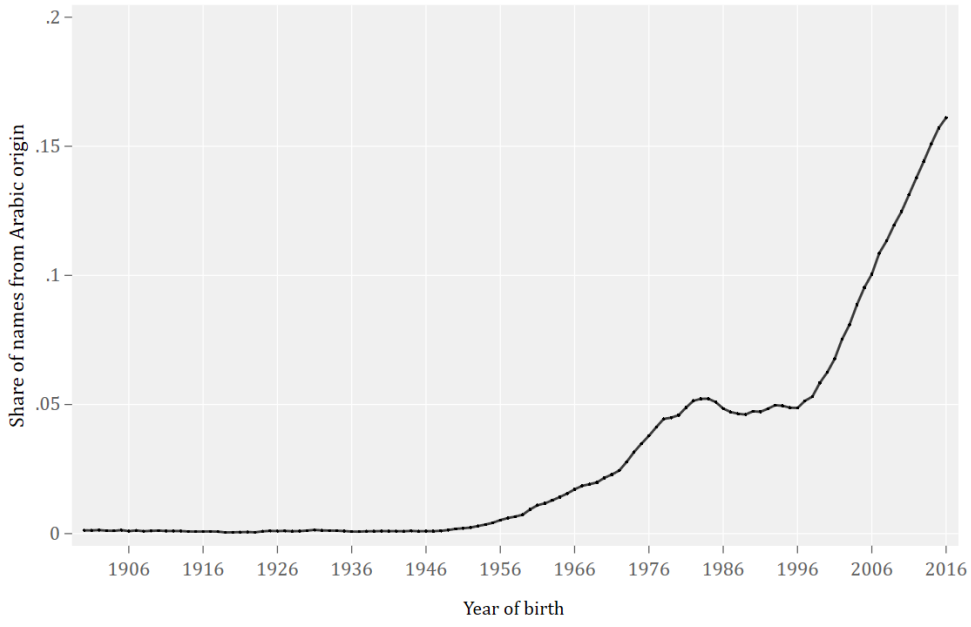
I attribute an Arabic origin to the names located in the top-left corner that are associated both with a low probability of French origin and with a high probability of Arabic origin. Conversely, the names located in the bottom-right corner are attributed a French origin. These two types of names are either included in the corpora, or relatively close to the listed names in their sub-sequences of letters. The names located in the bottom-left corner of the domain have both a low probability of French origin and a low probability of Arabic origin. These are mostly names from diverse third origins like Dimitri, Giovanni, Jimmy, or Donovan. In the top-right corner, very few names are attributed both a high probability of French origin and a high probability of Arabic origin, which is coherent with the fact that French and Arabic names are quite distinct etymologically. Some names still have a probability 1 of being of each origin: Al, An, L, and N. The attribution of a probability 1 to these very short names is a direct consequence of the loosening of the first axiom: in each of the corpora, at least one first name begins and/or ends with "al", and likewise for "an", "l", and "n". As expected in the previous subsection, since these first names are highly under-represented among newborns in France, this does not constitute a major issue. Finally, the names located within the circle at the center of domain are etymologically half-way between a French and an Arabic origin. This category notably includes first names like Soan and Nael that Fourquet and Manternach (2019) consider as Arabic, as well as many first names whose spelling seems to borrow as much from the French as from the Arabic registers like Nathis (Nathan/Anis), or Adryan (Abd Rayan/Adrien).

Figures 10a and 10b show the evolution of the share of Arabic names among male newborns in France from 1900 to 2016. In Figure 10a the Arabic origin results from the automatic attribution of the algorithm, while in Figure 10b the origin was attributed manually by Fourquet and Manternach (2019). The strong resemblance between the two curves suggests that the algorithm can target quite accurately the groups of origin that one would obtain *via* a manual classification. The curve produced by the algorithm is still less steep than that of Fourquet and Manternach (2019), especially at the end of the period.

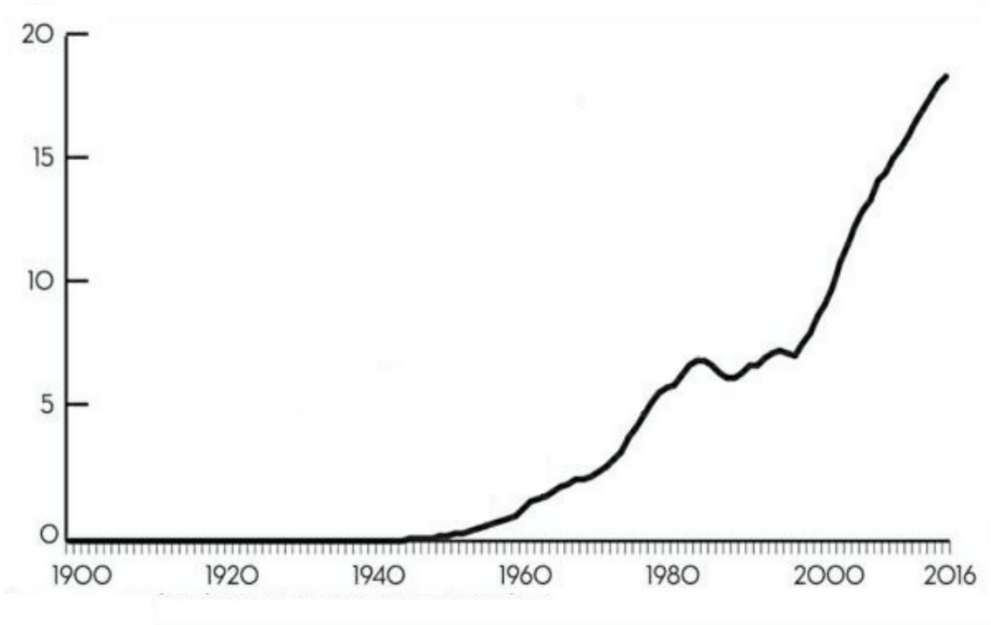
The reason for this slight dissonance is twofold. First, some names like Soan or Nael were manually classified as Arabic by Fourquet and Manternach (2019), but are attributed intermediate probabilities of each origin by the algorithm. The increasing popularity of such relatively short and etymologically ambiguous names contributes to the flattening of the curve produced by the algorithm. Second, the bounds of the area of the joint probability distribution that encloses names with a high probability of Arabic origin is set so as to form a reasonable set of areas of expected origins, not to maximize the fit between the two curves. Even though it may be argued that the latter approach could minimize arbitrariness, the lack of clear discontinuities in the joint probability distribution depicted in Figure 9 advocates rather conservative restrictions so as to minimize the risk of misclassification.

Figure 10: Share of Arabic first names according to different inference methods of origin

(a) Share of male first names whose origin is predicted to be Arabic by the algorithm



(b) Share of male first names manually classified as Arabic by Fourquet and Manternach (2019)



4.2.2 Comparison between predictions of first names and last names

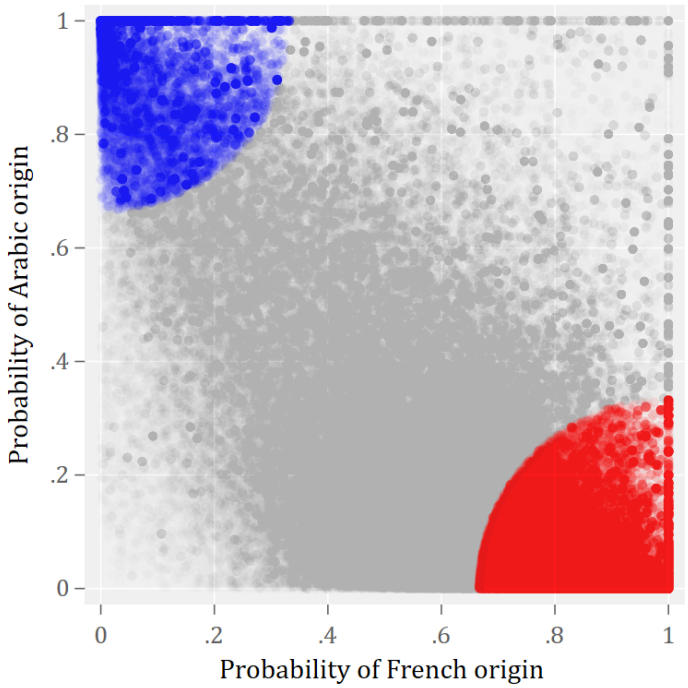
A second test of the validity of the predictions can be performed based on the Parisian electoral registers by comparing the origin attributed by the algorithm to the first name and to the last name of voters. As the results presented in the previous subsection support the reliability of the predictions of the origin of first names on expectation, the adequacy between the predicted origin of the first name and last name of Parisian voters would be additional evidence of the accuracy of the algorithm, and corroborate its transposability to the prediction of last names. Conceptually the algorithm can be applied directly to the study of last names, but the corpora must be adapted to account for the etymological differences between first names and last names.

I draw the list of last names of each origin by exploiting the Parisian electoral registers. This data source notably includes the name, the place of birth, and the date of birth of each Parisian voter. Constructing the corpora solely based on a birth-place criterion could yield misclassifications resulting from the French occupation in Maghreb during the most distant periods for the Arabic corpus, and from the immigration from Maghreb for the French corpus. To build the corpus of French last names, I thus focus on individuals born before 1940, when newborns in France are the most likely to be of French origin. Even if the main immigration waves from Maghreb took place after the Second World War, it seems preferable not to include the last name of every individual born in France before 1940 in the corpus. I keep only the names

whose frequency is higher among individuals born in France than among foreign-born individuals, and among those, the 500 most common names among people born in France are kept. Apart from the period chosen, the same methodology is followed to construct the corpus of Arabic names. Indeed, birth years prior to 1970 are ignored to limit the probability that an individual born in Maghreb is of French origin. The 500 most common last names of each origin are added to the previously defined lists of first names to form the final corpora.

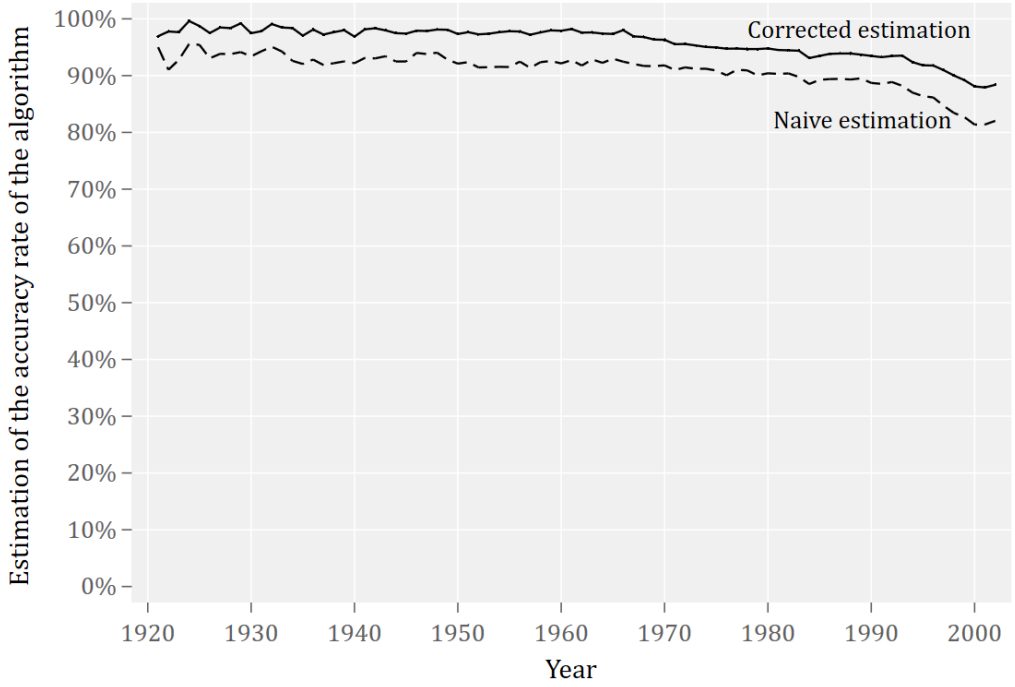
Based on these, the algorithm can attribute a probability of Arabic origin and a probability of French origin to each last name of the electoral register. The origin of first names is inferred in the exact same way as in the previous subsection. Thus, only males are considered. I split the domain of the joint distribution of the two probability variables in the same way as for first names. Indeed, in order to minimize the risk of misclassification and the contamination of last names coming from other origins than French and Arabic, I only keep the names located in the quadrants of radius $\frac{1}{3}$ centered at $(0,1)$ and $(1,0)$. These 17.5% of the total area of the domain of the joint probability distribution highlighted in Figure 11 gather 56.4% of the sample.

Figure 11: Classification of the 622,288 last names



The dashed line in Figure 12 shows the share of individuals for whom the predicted origin of the name coincides with the most likely origin associated to the corresponding first name by the algorithm.

Figure 12: Estimation of the percentage of accuracy of the algorithm



Among individuals with conflicting origin attributions, in many instances the algorithm rightly detects a last name of Arabic origin and a first name of French origin. To account for the validity of these predictions, the accuracy ratio can be adjusted as follows. Denote p_1 the share of attuned predictions and p_2 the share of individuals whose first name is classified as French and whose last name is classified as Arabic. A reasonable estimation of the accuracy rate p^* of the algorithm can be obtain as: $p^* = p_1 + p^* \times p_2$. The solid line in Figure 12 shows the solution p^* to this equation.

The estimation of the validity ratio of the algorithm exceeds 90% over virtually all the period. Despite cautious sample restrictions, the decrease in the accuracy starting in the late 60's is mainly due to misclassifications resulting from the internationalization of the Parisian population, and is amplified at the very end of the period by the popularization of neological and neographical first names.

4.2.3 Comparison to Coulmont and Simon (2019)

Coulmont and Simon (2019) study the most popular first names within ethnic groups in France based on the survey *Trajectoires et Origines*. Through the over-representation of immigrants and their offspring, this survey allows to distinguish the generation of immigration from the first to third for a large scope of origins. The third test consists in comparing the most popular first names by generation of immigration for individuals whose origin is identified by survey data

(TeO), and for individuals whose origin is determined by the algorithm. Among individuals whose origin is classified as Arabic based on their last name, I consider those born in Maghreb as first-generation immigrants and those born in France as second-generation immigrants. Table 7 compares the most popular first names in these groups to the first names Coulmont and Simon (2019) identified as the most popular among individuals of North-African origin for the first and the second generation of immigration.

Table 7: Most popular first names among individuals with North-African origins

Rank	Origin from survey data		Origin from last name	
	1st generation	2nd generation	Born in Maghreb	Born in France
1	Mohamed	Mohamed	Mohamed	Mohamed
2	Ahmed	Karim	Mohammed	Karim
3	Rachid	Medhi	Ahmed	Medhi

The match between the results for groups of first-generation immigrants is not particularly informative about the performances of the algorithm given that individuals born in Maghreb were selected to reproduce the results. But the perfect match between the most popular first names among second-generation North-African immigrants whose origin is identified by survey data and those born in France whose origin is inferred based on their last name by the algorithm seems to confirm that the latter correctly identifies the population that should be considered in order to produce reliable estimations of spatial segregation.

4.3 Comparison with the literature

There are two essentials priors to the computation of segregation indices at the department level. First, it is important to check whether the restriction to last names whose origin is well-identified modifies the composition of the sample in a way that significantly impacts the measures of segregation. Second, the extent to which using the last name rather than the nationality affects these indices should be quantified. For the sake of comparability with the existing literature on segregation in France, I consider the most commonly used segregation index, that of dissimilarity (Duncan and Duncan, 1955). This index quantifies the homogeneity of the distribution of two sub-populations into spatial units. Let A_m and F_m be the numbers of individuals of Arabic and French origin in municipality m and A_d and F_d the numbers of individuals of Arabic and French

origin in department d. The dissimilarity index D_d for department d across its M municipalities writes:

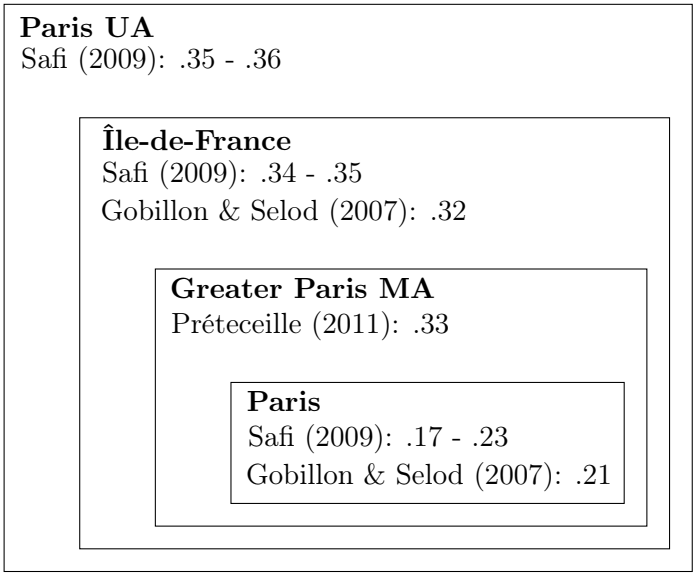
$$D_d = \frac{1}{2} \sum_{m=1}^{M} \left| \frac{A_m}{A_d} - \frac{F_m}{F_d} \right| \tag{9}$$

This index lies between 0 and 1 and can be interpreted as the share of the minority population that should move without being replaced to yield an equal repartition across sub-units. Relying on the actual origin of an individual based on her last name rather than on her nationality or her place of birth can affect the dissimilarity index of segregation in two ways. If individuals of Arabic origin born in France locate in municipalities where foreign-born individuals of Arabic origin are under-represented, using place of birth would overestimate segregation based on actual origins. But if individuals of Arabic origin born in France locate in municipalities where foreign-born individuals of Arabic origin are already over-represented, segregation according to place of birth would underestimate segregation according to actual origins. Finally, if individuals of Arabic origin born in France are homogeneously distributed relative to the rest of the French majority, both estimations should coincide.

To have an idea of the direction of this potential bias, both methods can be applied to the Parisian electoral registers. This data source is indeed particularly suited for this exercise as it includes the country of birth as well as the full name of each voter. Yet, the resulting estimates would only reflect segregation in the city-department of Paris at the time of the extraction of the data, i.e., in early 2020. This renders the comparison with the literature rather difficult, as to my knowledge (1) most recent studies on the segregation in the Paris area relied on the 1999 population census, and (2) only Gobillon and Selod (2007) and Safi (2009) provided estimates specifically for Paris (while the former focuses on the dissimilarity between French natives and North-African immigrants, the latter provides separate estimates for each country of birth).

The results put forward so far by the literature on segregation in the Paris region can be summarized as depicted in Figure 13. All the coefficients it mentions are dissimilarity indices computed between French and North-African individuals whose origin is either based on nationality or place of birth, using variations between municipalities. Note that each of the administrative areas represented as a nested rectangle includes the smaller rectangles: the Paris Urban Area includes Île-de-France, which includes the Greater Paris Metropolitan Area, which includes Paris.

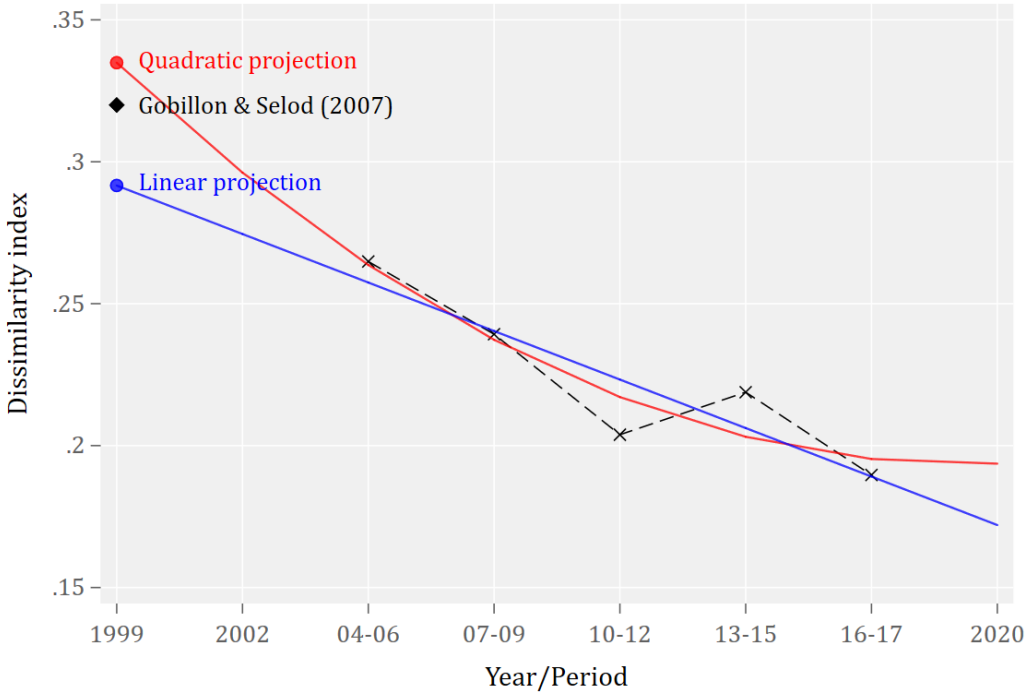
Figure 13: Dissimilarity indices of Paris & surroundings in 1999



A pattern seems to emerge from the results of these three studies: the wider the zone considered around Paris, the larger the segregation index. As mentioned above, the dissimilarity index of .21 computed for Paris by Gobillon and Selod (2007) is based on the use of the population census of 1999, and thus reflects the level of segregation 20 years before what the electoral registers would reflect. I will use this estimate as a benchmark as it was computed between French natives and a comprehensive group of North-African immigrants, and as it lies between the lowest and highest estimates provided by Safi (2009) separately for North-African countries of birth. Indeed, it is still possible to relate estimations in 1999 and in 2020 using the annual censuses carried out since 2004 and included in the Permanent Demographic Sample. Using this data source, a trend can be estimated for the dissimilarity index between 2004 and 2017, so as to compare its projections in 1999 and in 2020 to the index provided by Gobillon and Selod (2007) and the index estimated on the electoral register.

Due to sample size restrictions, the trend cannot be accurately estimated specifically for Paris, but that of Île-de-France can be computed based on the 162 municipalities in which more than 50 individuals either born in France or in Maghreb are observed in working age in each 3-year period from 2004 to 2015 and in the period 2016-2017. The evolution of the dissimilarity index and its linear and quadratic trends are depicted in Figure 14. To compute the trends, the index was assigned to the middle year of each 3-year period, and to the year 2017 for the period 2016-2017.

Figure 14: Evolution of the dissimilarity index in Île-de-France



Results indicate that segregation decreased from 2004 to 2017 in Île-de-France. The linear and quadratic projections in 1999 are of comparable magnitude to the estimation provided by Gobillon and Selod (2007). The dissimilarity index fell by 40% according to the quadratic trend, and by 46% according to the linear trend. For Paris specifically, Gobillon and Selod (2007) put forward an index of .21. Assuming that segregation in Paris evolved the same way as in Île-de-France, the growth rates between 1999 and 2020 estimated based on linear and quadratic trends are applied to the index for Paris to produce expected magnitudes in 2020. The dissimilarity index between people in working age born in France and in Maghreb is then computed for Paris at the municipality level in early 2020 using the electoral registers. To evaluate whether restricting the sample to the names whose predictions of origin are the most reliable has a noticeable impact on the dissimilarity index, the latter is computed both on the full and on the restricted sample. Finally, the dissimilarity index is computed using origins inferred from last names rather than from country of birth.

The resulting estimates are gathered in Table 8. The magnitudes of the indices computed on the electoral registers are very similar to those of the projections of the estimate of Gobillon and Selod (2007) based on the trend of Île-de-France. Also, the sample restriction does not seem to have a sizable impact on the index. But while all estimations and projections based on country of birth lie around .12, the index that relies on the prediction of origin from last names reaches .15.

Table 8: Comparison between segregation indices in Paris and benchmark estimations

2020 projections of Gobillon and Selod (2007)	Using country of birth	
	Based on linear trend	Based on quadratic trend
	[.113	- .126]
2020 estimations on electoral registers	Using country of birth	
	Full sample	Restricted sample
	[.112	- .122]
	Using last names	
	.151	

This suggests that individuals of Arabic origin born in France tend to locate in municipalities where foreign-born individuals of Arabic origin are already over-represented, such that the use of place of birth as a proxy for the actual origin underestimates segregation levels.

4.4 Department level estimates

4.4.1 Choice of indices

Even though the dissimilarity index is historically the main reference in the literature, perhaps because of its particularly transparent interpretation, contemporaneous standards advocate the use of multiple indices to draw accurate conclusions on segregation levels[23]. Moreover, Duncan and Duncan (1955)'s dissimilarity index was subject to various criticisms with respect to its properties. First, it tends to attribute a larger weight to the most populated sub-units (James and Taeuber, 1985). Second, it is not composition-invariant in the sense that for a given relative distribution of the minority group, the size of the minority group matters. Even if these properties are not of primary concern when the objective is to describe segregation in a specific area, it seems particularly important to avoid such limitations when it comes to comparing segregation levels between heterogeneous spatial units. To account for that, Taylor et al. (2000) put forward the following alternative segregation index:

$$S_d = \frac{1}{2} \sum_{m=1}^{M} \left| \frac{A_m}{A_d} - \frac{A_m + F_m}{A_d + F_d} \right| \tag{10}$$

[23]See Alivon (2016) for a comprehensive review of the various types of segregation indices.

The strong composition invariance of this estimator makes it better-suited for the comparison of segregation levels between unequal units in their share of individuals from the minority group. Just as the dissimilarity index, this segregation index captures a level of unevenness. But another important facet of segregation is often argued to be how it is experienced by individuals, a notion encapsulated in the so-called *contact indices* (Safi, 2009). The isolation index proposed by Cutler et al. (1999) falls in this category by quantifying the extent to which individuals from the minority group are only exposed to each other:

$$I_d = \frac{\sum_m \left(\frac{A_m}{A_d} \times \frac{A_m}{A_m + F_m} \right) - \frac{A_d}{A_d + F_d}}{\min \left\{ \frac{A_d}{\min_m (A_m + F_m)},\ 1 \right\} - \frac{A_d}{A_d + F_d}} \tag{11}$$

Thus, I consider both the segregation index and the isolation index to provide comparable estimations of (1) distribution unevenness and (2) likelihood to live close to co-ethnics, between French departments.

4.4.2 INSEE's last names file

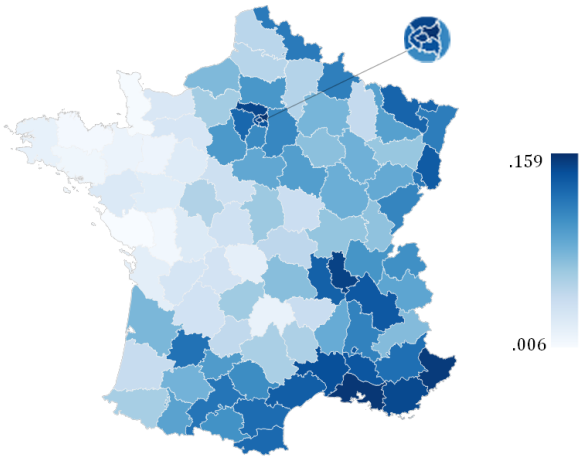
The last names file[24] is a database that gathers the last name as well as the municipality of birth of every individual born in France whose birth act was collected between 1891 and 1990. Birth counts per name and municipality are gathered in 4 periods: 1881-1915, 1916-1940, 1941-1965, and 1966-1990. The algorithm described in Section 4.1 is applied to each last name listed for the 1966-1990 period, and using the sample selection rules described above, only the predictions that are reliable enough are kept. To avoid irrelevant classifications of last names from various origins, only mainland France is considered. I also ignore municipalities with less than 50 births in the period, as they could yield fallaciously high or low shares of individuals of Arabic origin. The final sample gathers the name and municipality of birth of 12,315,972 individuals.

Figure 15a displays the share of individuals of Arabic origin according to their last name in the population. The highest shares are notably observed where large metropolitan areas are located: in Seine-Saint-Denis/Paris, Bouches-du-Rhône, and Rhône. Using the population censuses included in the Permanent Demographic Sample, these estimations can be compared to what is obtained by considering the country of birth rather than the name, as shown in Figure 15b. As the period used in the last name file is 1966-1990, I rely on the censuses from 1968 to 1990.

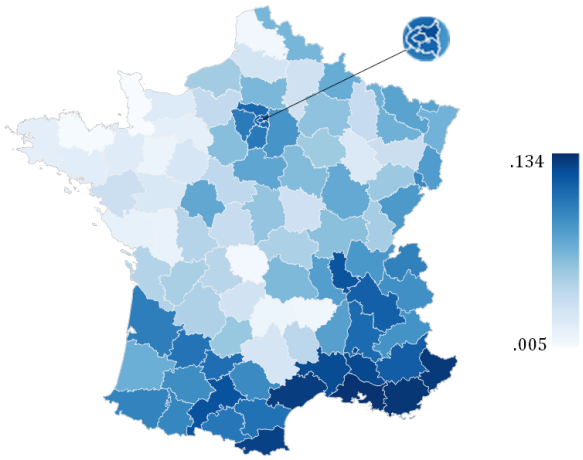
[24] "*Fichiers des noms patronymiques de 1891 à 1990 - Édition 1999*", INSEE (production), ADISP (diffusion).

Figure 15: Share of individuals of Arabic origin (1966-1990)

(a) Estimations based on last name



(b) Estimations based on country of birth



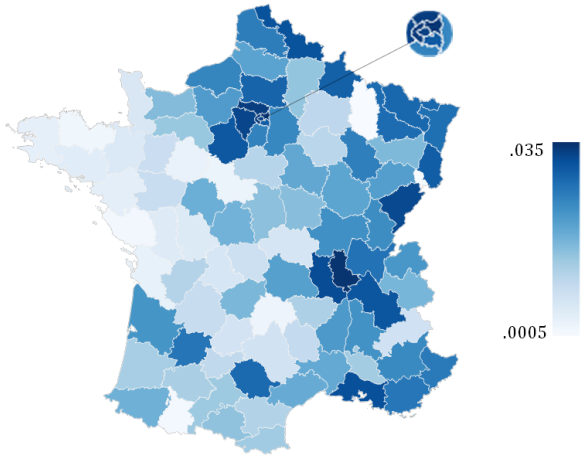
Spatial distributions are very similar when using the origin inferred from the last name and from the place of birth, but the latter approach tends to yield lower immigrant shares. This is both due to the fact that Figure 15b only accounts for individuals from Maghreb because of data restrictions on the precise country of origin for the first censuses, and to the omission by construction of second- and higher-generation immigrants when relying on place of birth.

4.4.3 Estimations

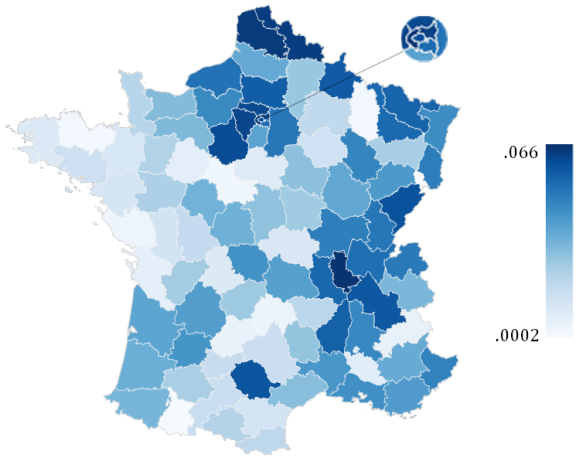
Figure 16 lays side to side the estimations of segregation (unevenness of the distribution) and isolation (likelihood to live close to co-ethnics) for the period 1966-1990 in France. These two indices exhibit similar variations over the French territory, with the highest values being generally associated with the departments where the share of individuals of Arabic origin is the highest: in Seine-Saint-Denis, Paris, and Rhône. There are still discrepancies for some departments like Bouches-du-Rhône where the segregation index is noticeably higher than the isolation index relatively to the other French departments. This would suggest than in Bouches-du-Rhône, individuals of Arabic origin tend to be less isolated from individuals of French origin than they are unevenly distributed across municipalities relatively to the other French departments.

Figure 16: Segregation indices at the department level (1966-1990)

(a) Segregation index



(b) Isolation index



5 Interactions between mobility and segregation

The variations of the segregation and the isolation indices across French departments can be used to study its potential links with intergenerational mobility. I associate to each child from the Permanent Demographic Sample the indices of segregation and isolation that correspond to her department of birth. As these indices apply to the period 1966-1990, and as most individuals in the sample are born around the 1970's, these variables reflect the level of segregation individuals experienced during childhood.

Table 9 investigates how intergenerational mobility, segregation levels, and the origin of an individual interplay. The *Origin* variable takes the value 1 if the individual has at least one parent born in Maghreb, and the father income rank variable is denoted *Parental rank*. Department fixed effects are included in each regression. Coefficients of the first column are interpreted in the same way as those in Table 4. They consistently indicate that children of individuals born in Maghreb tend to be located at lower ranks in the income distribution and tend to experience a stronger intergenerational persistence. The coefficient associated with the segregation indices does not show any significant relationship between segregation and children's income rank, nor any difference depending on individuals' origin. Yet, the last two columns present significantly positive coefficients associated with the interaction between segregation and parental income rank, implying that more segregated areas are associated with lower rates of intergenerational mobility. The last interaction does not elicit any significant heterogeneity in this effect depending on individuals' origin.

Table 9: Rank-rank correlation, parental origin, and segregation

	Child income rank				
Parental rank	.298***	.298***	.298***	.282***	.28***
	(.006)	(.006)	(.006)	(.01)	(.01)
Origin	−7.066***	−7.066***	−6.972***	−6.741***	−7.765***
	(1.336)	(1.336)	(1.372)	(1.376)	(2.287)
Parental rank × Origin	.06***	.06***	.061***	.056***	.072**
	(.019)	(.019)	(.02)	(.02)	(.034)
Segregation		−.24	−.222	−1.385	−1.478
		(.959)	(.96)	(1.088)	(1.1)
Segregation × Origin			−.105	−.037	.743
			(.347)	(.348)	(1.434)
Segregation × Parental rank				.015**	.017**
				(.007)	(.007)
Segregation × Parental rank × Origin					−.011
					(.02)
Constant	37.583***	37.844***	37.836***	39.063***	39.161**
	(1.379)	(2.317)	(2.317)	(2.379)	(2.385)

Nb obs: 53,253 - *Standard errors in parentheses - Department fixed effects included in each regression*

By replicating the analysis using intergenerational elasticities rather than rank-rank correlations, Table 10 provides additional insights on the relationships between intergenerational mobility, segregation levels, and individuals' origin. First, the last three columns elicit a significantly negative link between segregation and income for second-generation immigrants from Maghreb that does not hold for children of natives. In addition, while the negative effect of segregation on intergenerational mobility is still significant, its magnitude is now shown to be even higher for individuals with at least one parent born in Maghreb than for individuals whose parents are both born in France. Thus, it appears that as hypothesized earlier, higher segregation levels are associated not only with lower economic outcomes, but also with a higher intergenerational persistence, particularly for second-generation immigrants from Maghreb. As shown by Tables 19 and 20 in Appendix, the exact same patterns are observed for the isolation index. Hence, just as hiring discrimination, it is indeed likely that segregation levels actually shape individuals' socioeconomic perspectives in a way that accentuates the gap between the mobility opportunities of second-generation immigrants and these of children of French natives.

Table 10: Intergenerational mobility, parental origin, and segregation

	Child log income				
Parental log income	.451***	.451***	.449***	.407***	.417***
	(.011)	(.011)	(.011)	(.017)	(.018)
Origin	−3,727***	−3,727***	−3,158***	−2,958***	−314
	(859)	(860)	(911)	(913)	(1,539)
Parental log income × Origin	.057*	.057*	.062*	.053	−.054
	(.033)	(.033)	(.033)	(.033)	(.06)
Segregation		−285	−221	−1,185	−971
		(727)	(727)	(786)	(792)
Segregation × Origin			−496*	−433*	−2,223**
			(263)	(264)	(879)
Segregation × Parental log income				.035***	.027**
				(.011)	(.012)
Segregation × Parental log income × Origin					.071**
					(.033)
Constant	10,560***	10,876***	10,822***	11,976***	11,712***
	(1,027)	(1,745)	(1,745)	(1,781)	(1,785)

Nb obs: 54,474 - *Standard errors in parentheses - Department fixed effects included in each regression*

6 Conclusion

Throughout this study, the richness of the data gathered in INSEE's Permanent Demographic Sample allowed to draw a general depiction of the extent and features of intergenerational mobility in France. Intergenerational wage elasticities are shown to be higher within genders than across genders. This "like father like son" and "like mother like daughter" phenomenon could be due to the fact that children identify more with their parent of the same gender (Starrels, 1994), which could later translate into within-gender occupational reproduction behaviors, strengthened by a certain degree of occupational gender segregation (Watt, 2010). Intergenerational wage elasticity estimations suggest that a 40 year-old woman (resp. man) in France would on expectation earn 7.23% (resp. 6.47%) more if her mother (resp. his father) was earning 10% more at the same age. Thus, intergenerational persistence in France is among the highest in previously studied OECD countries (Acciari et al., 2019; Chetty et al., 2014; Corak, 2013b; Mazumder, 2016). In line with the results of Lefranc (2018), splitting the sample by birth cohort shows that the IGE followed

an increasing trend from the mid 1960's to the mid 1970's. Rank-rank correlations based on household income are also noticeably high. Between fathers and sons, for instance, the estimate reaches .415, which confirms that the French social ladder is particularly hard to climb, and to fall from.

This is particularly true for children with at least one parent born in Maghreb. They have systematically lower expected ranks in the income distribution even by conditioning on parental income. In other words, by starting at comparable socioeconomic levels during childhood, second-generation immigrants from Maghreb still end up lower in the social ladder than children of native parents. However, results indicate that they have very comparable educational outcomes and hourly wage levels, which discredits the hypotheses of differences in human capital investments and in ability/productivity to explain the robustness of the ethnic gap to the conditioning on parental earnings. The explanation rather seems to lie in differences in terms of access to employment: the number of hours worked is systematically lower and the probability to have perceived unemployment benefits over the period studied is systematically larger for second-generation immigrants from Maghreb.

These findings echo the well-documented hiring discrimination towards applicants of North-African origin observed on the French labor market. Another potential underlying channel to these differentiated intergenerational mobility prospects is residential segregation. As previous estimations of segregation are not comprehensive enough to exploit their spatial variations at the department level, I develop a new approach that allows to estimate segregation indices for all departments of mainland France using variations at the municipality level. More specifically, I develop an algorithm that can infer an individual's origin based on her name, that I apply to more than 12,000,000 individuals using the INSEE's *fichier patronymique*. Segregation indices tend to be higher in large metropolitan areas, particularly in Paris, in Seine-Saint-Denis, in Bouches-du-Rhône, and in Rhône.

The interactions between segregation levels experienced during childhood, ethnic origins, and intergenerational mobility, reveal several patterns. First, segregation is shown to be negatively correlated with income for second-generation immigrants from Maghreb but not for children of French natives. Second, segregation is also associated with a higher intergenerational persistence irrespective of individuals' origin, but the effect is shown to be stronger for children with at least one parent born in Maghreb. Thus, just as hiring discrimination, segregation levels in the environment children grow up in appear as a plausible driver of the ethnic gap conditional on parental earnings. This calls for additional empirical research to understand the underlying

mechanisms behind these relationships, be they related to an increased difficulty to build a diversified network, or to an impact on motivation and ambition through a feeling of exclusion.

One limitation of the INSEE's *fichier patronymique* for the study of segregation is that municipality of birth is an imperfect proxy for municipality of residence, especially as the medicalization of childbirths tends to over-allocate individuals to larger cities. Yet, the approach developed in this Master's Thesis can be applied to more ambitious databases to obtain extremely precise estimates of segregation at very granular spatial levels. For instance, the soon available *Répertoire Électoral Unique* will gather the name and address of residence of every French voter. By inferring the origin from the last name and matching the address to a precise longitude and latitude, segregation indices could be computed at virtually any geographical scale, and their variations could be thoroughly analyzed in order to better understand its likely causes and consequences. More generally, this approach can catalyze research advances on any socio-demographic indicator for which the use of place of birth or nationality instead of the actual origin constitutes a restriction to the conduction of comprehensive analyses on well-defined groups of individuals. For instance, it would allow to better identify populations that are likely to be discriminated against on the labor market than what can be done by relying on place of birth or nationality.

References

Acciari, Paolo, Alberto Polo, and Gianluca Violante, "'And yet, it moves': Intergenerational mobility in Italy," *IZA Discussion Papers, No. 12273*, 2019.

Adida, Claire L, David D Laitin, and Marie-Anne Valfort, ""One Muslim is enough!" Evidence from a field experiment in France," *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, 2016, *121/122*, 121–160.

Alivon, Fanny, "La segrégation spatiale et économique: Une analyse en termes d'emploi et d'éducation dans les espaces urbains." PhD dissertation, Dijon 2016.

Böhlmark, Anders and Matthew J Lindquist, "Life-cycle variations in the association between current and lifetime income: Replication and extension for Sweden," *Journal of Labor Economics*, 2006, *24* (4), 879–896.

Bourdieu, Jérôme, Joseph P Ferrie, and Lionel Kesztenbaum, "Vive la différence? Intergenerational mobility in France and the United States during the nineteenth and twentieth centuries," *Journal of Interdisciplinary History*, 2009, *39* (4), 523–557.

Cahuc, Pierre, Stéphane Carcillo, Andreea Minea, and Marie-Anne Valfort, "When Correspondence Studies Fail to Detect Hiring Discrimination," *IZA Discussion Paper*, 2019.

Chetty, Raj and Nathaniel Hendren, "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects," *The Quarterly Journal of Economics*, 2018, *133* (3), 1107–1162.

_ and _ , "The impacts of neighborhoods on intergenerational mobility II: County-level estimates," *The Quarterly Journal of Economics*, 2018, *133* (3), 1163–1228.

_ , _ , Maggie R Jones, and Sonya R Porter, "Race and economic opportunity in the United States: An intergenerational perspective," *The Quarterly Journal of Economics*, 2020, *135* (2), 711–783.

_ , _ , Patrick Kline, and Emmanuel Saez, "Where is the land of opportunity? The geography of intergenerational mobility in the United States," *The Quarterly Journal of Economics*, 2014, *129* (4), 1553–1623.

Choi, Bernard CK, JG Hanley, Eric J Holowaty, and Darlene Dale, "Use of surnames to identify individuals of Chinese ancestry," *American journal of epidemiology*, 1993, *138* (9), 723–734.

Clark, Gregory, Neil Cummins, Yu Hao, and Daniel Diaz Vidal, *The son also rises: Surnames and the history of social mobility*, Princeton University Press, 2015.

Corak, Miles, "Income inequality, equality of opportunity, and intergenerational mobility," *Journal of Economic Perspectives*, 2013, *27* (3), 79–102.

_ , "Inequality from generation to generation: The United States in comparison," *The Economics of Inequality, Poverty, and Discrimination in the 21st Century*, 2013, *1*, 107–126.

Coulmont, Baptiste, *Sociologie des prénoms*, la Découverte, 2011.

_ **and Patrick Simon**, "How do immigrants name their children in France?," *Population & Sociétés*, 2019, *4*, 1–4.

Cutler, David M, Edward L Glaeser, and Jacob L Vigdor, "The rise and decline of the American ghetto," *Journal of political economy*, 1999, *107* (3), 455–506.

Dherbécourt, Clément, "La géographie de l'ascenseur social français," *France Stratégie, Document de travail*, 2015, *6*.

Duncan, Otis Dudley and Beverly Duncan, "A methodological analysis of segregation indexes," *American sociological review*, 1955, *20* (2), 210–217.

Edo, Anthony, Nicolas Jacquemet, and Constantine Yannelis, "Language skills and homophilous hiring discrimination: Evidence from gender and racially differentiated applications," *Review of Economics of the Household*, 2019, *17* (1), 349–376.

Fourquet, Jérôme, *L'Archipel français*, Le Seuil, 2019.

_ **and Sylvain Manternach**, "Cent ans d'immigration racontés par les prénoms," *Herodote*, 2019, *3*, 113–140.

Gobillon, Laurent and Harris Selod, "Les déterminants locaux du chômage en région parisienne," *Economie prevision*, 2007, *4*, 19–38.

Güell, Maia, José V Rodrıguez Mora, and Christopher I Telmer, "Intergenerational mobility and the informational content of surnames," *Working Paper*, 2013.

Haider, Steven and Gary Solon, "Life-cycle variation in the association between current and lifetime earnings," *American Economic Review*, 2006, *96* (4), 1308–1320.

Inoue, Atsushi and Gary Solon, "Two-sample instrumental variables estimators," *The Review of Economics and Statistics*, 2010, *92* (3), 557–561.

James, David R and Karl E Taeuber, "Measures of segregation," *Sociological methodology*, 1985, *15*, 1–32.

Jobling, Mark A, "In the name of the father: Surnames and genetics," *TRENDS in Genetics*, 2001, *17* (6), 353–357.

King, Turi E and Mark A Jobling, "What's in a name? Y chromosomes, surnames and the genetic genealogy revolution," *Trends in Genetics*, 2009, *25* (8), 351–360.

_ , Stéphane J Ballereau, Kevin E Schürer, and Mark A Jobling, "Genetic signatures of coancestry within surnames," *Current biology*, 2006, *16* (4), 384–388.

Lasker, Gabriel W, "Surnames in the study of human biology," *American Anthropologist*, 1980, *82* (3), 525–538.

Lasker, Gabriel Ward, *Surnames and genetic structure*, Vol. 1, Cambridge University Press, 1985.

Lefranc, Arnaud, "Intergenerational earnings persistence and economic inequality in the long run: Evidence from French cohorts, 1931–75," *Economica*, 2018, *85* (340), 808–845.

_ and Alain Trannoy, "Intergenerational earnings mobility in France: Is France more mobile than the US?," *Annales d'Économie et de Statistique*, 2005, pp. 57–77.

_ , Nicolas Pistolesi, and Alain Trannoy, "Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France," *Journal of Public Economics*, 2009, *93* (11-12), 1189–1207.

Mateos, Pablo, "A review of name-based ethnicity classification methods and their potential in population studies," *Population, Space and Place*, 2007, *13* (4), 243–263.

_ et al., "Names, Ethnicity and Populations," *Advances in Spatial Science*, 2014.

Mazieres, Antoine and Camille Roth, "Large-scale diversity estimation through surname origin inference," *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 2018, *139* (1), 59–73.

Mazumder, Bhashkar, "Fortunate sons: New estimates of intergenerational mobility in the United States using social security earnings data," *Review of Economics and Statistics*, 2005, *87* (2), 235–255.

_ , "Estimating the intergenerational elasticity and rank association in the United States: Overcoming the current limitations of tax data," in "Inequality: Causes and Consequences," Emerald Group Publishing Limited, 2016, pp. 83–129.

McAvay, Haley and Mirna Safi, "Is there really such thing as immigrant spatial assimilation in France? Desegregation trends and inequality along ethnoracial lines," *Social science research*, 2018, *73*, 45–62.

O'Neill, Donal and Olive Sweetman, "Intergenerational mobility in Britain: Evidence from unemployment patterns," *Oxford Bulletin of Economics and Statistics*, 1998, *60* (4), 431–447.

Polednak, Anthony P, "Estimating cervical cancer incidence in the Hispanic population of Connecticut by use of surnames," *Cancer*, 1993, *71* (11), 3560–3564.

Poncelet, Thomas, Lauren Trigano, and Mariette Sagot, "Gravir l'échelle sociale est plus aisé en Ile-de-France qu'en province," *Insee Analyses*, 2016, *50*.

Préteceille, Edmond, "Has ethno-racial segregation increased in the greater Paris metropolitan area?," *Revue française de sociologie*, 2011, *52* (5), 31–62.

Quillian, Lincoln and Hugues Lagrange, "Socioeconomic segregation in large cities in France and the United States," *Demography*, 2016, *53* (4), 1051–1084.

Safi, Mirna, "La dimension spatiale de l'intégration: évolution de la ségrégation des populations immigrées en France entre 1968 et 1999," *Revue française de sociologie*, 2009, *50* (3), 521–552.

Shah, Baiju R, Maria Chiu, Shubarna Amin, Meera Ramani, Sharon Sadry, and Jack V Tu, "Surname lists to identify South Asian and Chinese ethnicity from secondary data in Ontario, Canada: A validation study," *BMC medical research methodology*, 2010, *10* (1), 42.

Shon, Jean-Louis Pan Ké and Gregory Verdugo, "Forty years of immigrant segregation in France, 1968–2007. How different is the new immigration?," *Urban Studies*, 2015, *52* (5), 823–840.

Silberman, Roxane, Richard Alba, and Irène Fournier, "Segmented assimilation in France? Discrimination in the labour market against the second generation," *Ethnic and Racial studies*, 2007, *30* (1), 1–27.

Solon, Gary, "Intergenerational income mobility in the United States," *The American Economic Review*, 1992, pp. 393–408.

Starrels, Marjorie E, "Gender differences in parent-child relations," *Journal of Family Issues*, 1994, *15* (1), 148–165.

Taylor, Chris, Stephen Gorard, and John Fitz, "A re-examination of segregation indices in terms of compositional invariance," *Social Research Update*, 2000, *30*, 1–4.

Verdugo, Gregory, "Public housing and residential segregation of immigrants in France, 1968-1999," *Population*, 2011, *66* (1), 169–193.

Watt, Helen MG, "Gender and occupational choice," in "Handbook of gender research in psychology," Springer, 2010, pp. 379–400.

Appendix

Table 11: Socio-professional categories

Category	Description
Reference category	Artisans, tradesmen, and entrepreneurs.
Executive position	Engineers, professors, and other intellectual professions and executive positions.
Intermediary profession	Technicians, foremen, school teachers, intermediary public servant occupations, and other intermediary occupations.
Employee	Lower civil servant positions, policemen, military, intermediary administrative positions, personal services workers, and other employee occupations.
Blue-collar job	Industrial and artisanal qualified and unqualified workers, agricultural workers, drivers, and other blue-collar jobs.

Table 12: Impact of the percentile discretization

	Wage	Hh. income	Wage	Hh. income
	Discrete percentile		Continuous exact position	
Father-son	.2538	.4155	.2536	.4153
	(.0068)	(.0123)	(.0068)	(.0123)
Father-daughter	.2893	.3797	.2893	.3797
	(.0082)	(.0139)	(.0082)	(.0139)
Mother-son	.1646	.3096	.1646	.3096
	(.0082)	(.0147)	(.0082)	(.0147)
Mother-daughter	.2914	.3232	.2914	.3234
	(.0089)	(.0159)	(.0089)	(.0158)

Standard errors in parentheses

Table 13: Estimations blind to the parental attenuation bias

	Wage	Hh. income	Wage	Hh. income
	Intergenerational elasticities		Rank-rank correlations	
Father-son	.602	.495	.236	.386
	(.013)	(.012)	(.006)	(.011)
Father-daughter	.549	.458	.276	.348
	(.014)	(.013)	(.007)	(.012)
Mother-son	.485	.448	.141	.272
	(.016)	(.015)	(.006)	(.010)
Mother-daughter	.599	.450	.234	.265
	(.017)	(.017)	(.006)	(.011)

Standard errors in parentheses

Table 14: Coefficients without top and bottom 1% individuals

	Wage	Hh. income	Wage	Hh. income
	Intergenerational elasticities		Rank-rank correlations	
Father-son	.543	.486	.239	.399
	(.014)	(.013)	(.007)	(.012)
Father-daughter	.536	.433	.283	.357
	(.016)	(.015)	(.008)	(.014)
Mother-son	.447	.455	.156	.297
	(.022)	(.021)	(.008)	(.015)
Mother-daughter	.686	.479	.286	.305
	(.023)	(.022)	(.009)	(.016)

Standard errors in parentheses

Table 15: Coefficients without top and bottom 1% fathers

	Wage	Hh. income	Wage	Hh. income
	Intergenerational elasticities		Rank-rank correlations	
Father-son	.645	.534	.253	.412
	(.015)	(.014)	(.007)	(.012)
Father-daughter	.563	.492	.287	.377
	(.017)	(.016)	(.008)	(.014)
Mother-son	.537	.506	.163	.316
	(.025)	(.023)	(.009)	(.015)
Mother-daughter	.723	.538	.290	.316
	(.025)	(.024)	(.009)	(.016)

Standard errors in parentheses

Table 16: Intergenerational elasticity & parental origin

	Child log income			
Father log income	.49***		.489***	.482***
	(.01)		(.01)	(.011)
≥ 1 parent born in Maghreb		−2,329***	−2,144***	−3,587***
		(259.419)	(254.134)	(853.167)
Interaction				.058*
				(.033)
Constant	8,620***	21,098***	8,868***	9,023***
	(265.598)	(82.227)	(267.047)	(281.129)

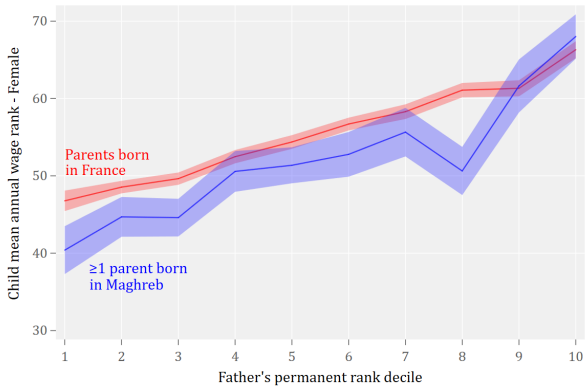
Nb obs: 54,575 - *Standard errors in parentheses*

Table 17: Age distribution by parental place of birth

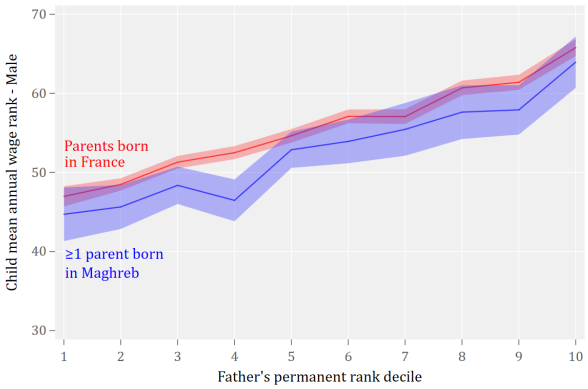
	1%	5%	10%	25%	50%	Mean	SD	75%	90%	95%	99%
France	30	31	32	36	41	40.48	5.74	45	48	49	50
Maghreb	30	31	32	35	39	39.67	5.7	44	48	49	50

Figure 17: Average economic outcomes conditional on parental income decile

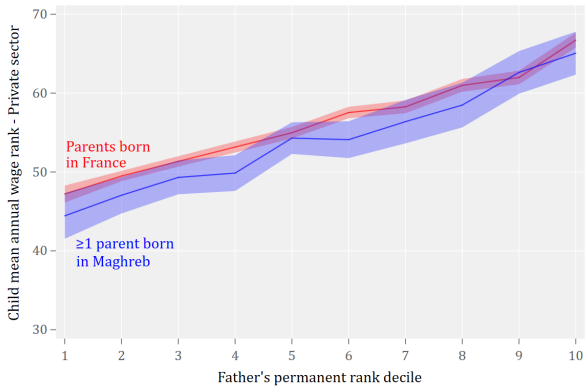
(a) Female



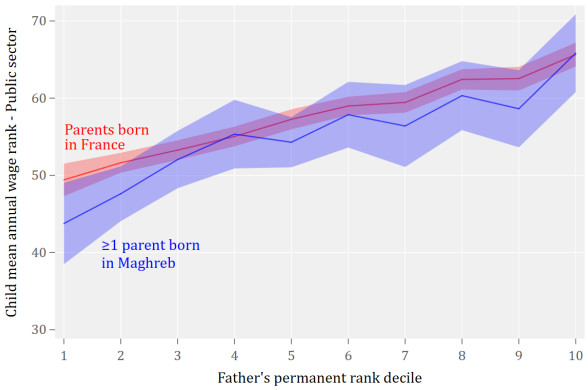
(b) Male



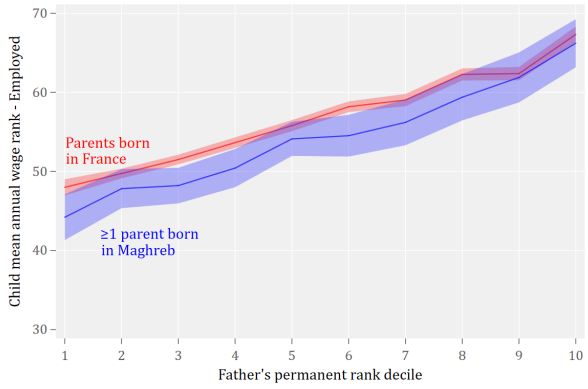
(c) Private sector



(d) Public sector



(e) Employed



(f) Self-employed

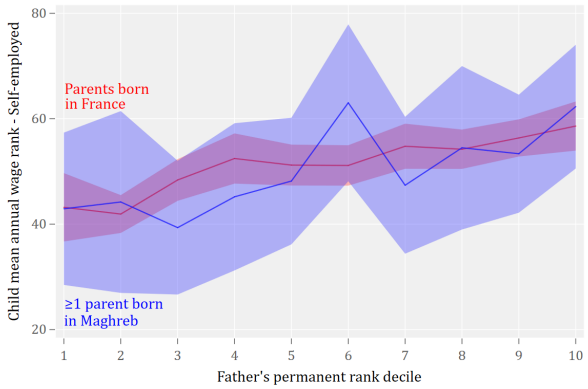


Table 18: Decomposition of the index Γ according to the sub-sequences of the name Louis

j	$s_{1,j}$	$s_{2,j}$	$s_{3,j}$	$s_{4,j}$	$s_{5,j}$	$s_{6,j}$	n_j	$j/\sum j$	$1/n_j$	$\omega(s_{i,j})$
1	L	O	U	I	S		5	1/21	1/5	.010
2	¿L	LO	OU	UI	IS	S?	6	2/21	1/6	.016
3	¿LO	LOU	OUI	UIS	IS?		5	3/21	1/5	.029
4	¿LOU	$LOUI$	$OUIS$	UIS?			4	4/21	1/4	.048
5	¿$LOUI$	$LOUIS$	$OUIS$?				3	5/21	1/3	.079
6	¿$LOUIS$	$LOUIS$?					2	6/21	1/2	.143

Table 19: Rank-rank correlation, parental origin, and isolation

	Child income rank				
Parental rank	.298***	.298***	.298***	.276***	.274***
	(.006)	(.006)	(.006)	(.009)	(.009)
Origin	−7.066***	−7.067***	−7.066***	−6.675***	−8.058***
	(1.336)	(1.336)	(1.336)	(1.363)	(1.899)
Parental rank × Origin	.06***	.06***	.06***	.06***	.081***
	(.019)	(.019)	(.019)	(.019)	(.028)
Isolation		−.284	−.251	−1.355	−1.443
		(1.156)	(1.156)	(1.192)	(1.195)
Isolation × Origin			−.247	−.175	.646
			(.199)	(.2)	(.81)
Isolation × Parental rank				.014***	.016***
				(.004)	(.004)
Isolation × Parental rank × Origin					−.012
					(.012)
Constant	37.583***	38.263***	38.056***	39.764***	39.902***
	(1.379)	(3.263)	(3.263)	(3.294)	(3.296)

Nb obs: 53,253 - *Standard errors in parentheses - Department fixed effects included in each regression*

Table 20: Intergenerational elasticity, parental origin, and isolation

	Child log income				
Parental log income	.451***	.451***	.449***	.415***	.419***
	(.011)	(.011)	(.011)	(.016)	(.016)
Origin	−3,727***	−3,727***	−3,071***	−3,037***	−1,992
	(859)	(860)	(897)	(897)	(1,288)
Parental log income × Origin	.057*	.057*	.061*	.057*	.014
	(.033)	(.033)	(.033)	(.033)	(.05)
Isolation		−344	−290	−850	−789
		(875)	(875)	(893)	(894)
Isolation × Origin			−390***	−346**	−894*
			(152)	(152)	(507)
Isolation × Parental log income				.02***	.018***
				(.006)	(.007)
Isolation × Parental log income × Origin					.022
					(.02)
Constant	10,560***	11,172***	11,106***	12,067***	11,960***
	(1,027)	(2,464)	(2,464)	(2,482)	(2,484)

Nb obs: 54,474 - *Standard errors in parentheses - Department fixed effects included in each regression*

Reappraisal of Fourquet and Manternach (2019):

An algorithmic approach to the onomatological inference of genealogical origins

In their follow-up study of Fourquet (2019), Fourquet and Manternach (2019) document a large increase in the share of Arabic first names among newborns in France (Figure 10b): this category of first names account for 20% of male births today, against 1% in 1960. The methodology this result relies on consists in manually attributing an Arabic or non-Arabic origin to each first name contained in the exhaustive list of first names given to individuals born in France, which is publicly shared and updated yearly by the INSEE. This approach potentially suffers from two main limits that this reappraisal aims to tackle.

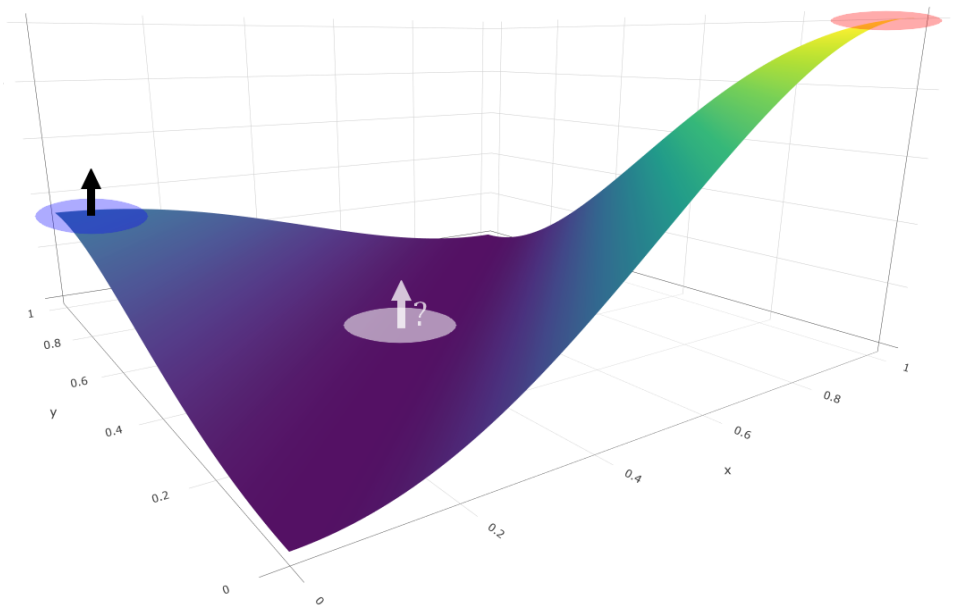
First, the binary attribution of an Arabic origin to first names may be too restrictive, especially for most recent period. Indeed, Coulmont (2011) shows that during the period over which Fourquet (2019) documents an increase in Arabic first names, the minimum number of names necessary to account for half of newborns increases by four times, and that the share of rare first names is now five times higher than what was observed in 1960. This diversification and enrichment of the list of names given to French newborns fade the delimitations between the different groups of origins, and sophisticate the classification exercise. Indeed, neological first names are increasingly common, and may in some cases borrow their etymological origins from different cultures. The restriction imposed by the dichotomous classification obliterates the potentiality of the formation of a continuity between the two groups of origin considered instead of a cleavage in the cultural borrowings of the first names attributed to newborns over the last decades.

The second limitation is the looseness of the decision process allowed by the manual classification of first names according to their supposed origin. For instance, Fourquet and Manternach (2019) assign Adam to the "modern Arabic names" category, while one may just as well not consider Adam to be an Arabic name in the first place due to its relatively cross-cultural nature. This illustrates the issue raised by the room for arbitrariness and subjective appreciation let by the non-systematic methodology employed by Fourquet and Manternach (2019), that the algorithmic aspect of the approach proposed in this reappraisal aims to tackle.

Indeed, these two limitations can be jointly addressed by relying on an algorithm that automatically assigns a continuous probability of etymological origin to first names based on the correspondence between their sub-sequences of letters and those appearing in given corpora of specific origins. This allows to test the hypothesis of the reinforcement of a dichotomy against that of a continuity in the cultural borrowings of the first names given to newborns in France.

Following the methodology described in Section 4.1, these hypotheses can be formulated in terms of evolution of the joint density of probabilities of etymological origin.

Figure 18: Illustration of the hypotheses on the joint density



Depicted function : $f(x,y) = (\mathbb{1}\{x > y\}\frac{1}{4} + \mathbb{1}\{x \leq y\}\frac{1}{12}) \times \left[1 + cos(\pi(x-y) - \pi)\right]$

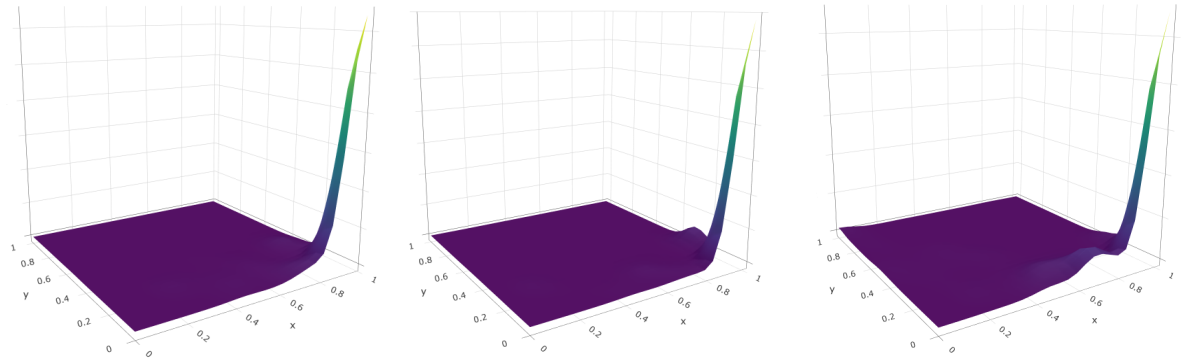
Figure 18 schematically represents the expected shape of the joint distribution of the probability of French origin (x) and the probability of Arabic origin (y). The area indicated by the red disk corresponds to the location of the first names with a high probability of French origin and a low probability of Arabic origin, and the blue disk indicates the area of low probability of French origin and high probability of Arabic origin. Given the evidence put forward by Fourquet (2019), the density under the blue disk is supposed to have increased over the last decades. The hypothesis of interest is whether the density under the white disk did increase as well or not, i.e., whether or not names that would etymologically be in between traditional French and Arabic names also gained in popularity in a way that formed a continuum between the two polar zones of the joint distribution.

Figures 19a to 19i describe the evolution of the joint density from 1900 to 2015. Over the first half of the XXth century, virtually all the density is concentrated on the area of traditional French names. The left corner of the domain of the joint distribution, where traditional Arabic names are located, then continuously rises until 2015, as documented by Fourquet (2019) in Figure 10b. The evolution of the joint density also clearly indicates the formation of a continuum between traditional French names and traditional Arabic names, which progressively accompanied the increase in the share of Arabic first names. Along this continuum are notably located names such

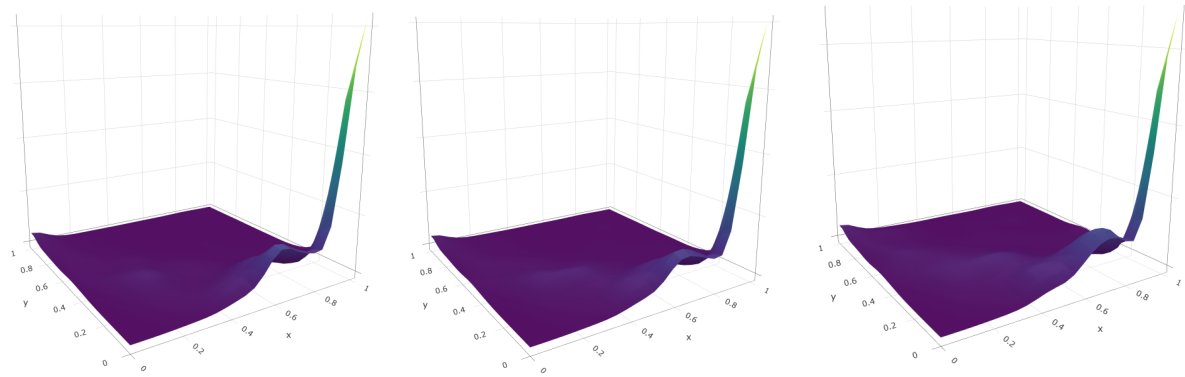
as Soan, Milan, Nael, Nils, Timaël, Ilian, Aédan, but also Nathis, which seem to be etymologically in between the names Nathan et Anis, or Adryan between Abd Rayan and Adrien. Thus, while the binary classification translates by construction into a dichotomous trend, the flexibility of the algorithmic approach allows to identify the development of a continuity in the joint distribution of origins due to an increasing share of names that are etymologically in between French and Arabic origins.

Figure 19: Evolution of the joint density of the probabilities of French (x) and Arabic (y) origin

(a) 1900 (b) 1950 (c) 1970



(d) 1990 (e) 1995 (f) 2000



(g) 2005 (h) 2010 (i) 2015