Ethnic differentials on the labor market in the presence of asymmetric spatial sorting: Set identification and estimation

Roland Rathelot[§]

July 1, 2014

Abstract

This paper aims to isolate the ethnic gap on the labor market that can be attributed to ethnicity and not to differences in individual characteristics or residential location. Controlling for residential location is important as ethnic minorities often live in distressed neighborhoods. It is also challenging because spatial sorting is likely to differ across ethnicities because of labor- or housing-market discrimination. This paper shows that controlling for neighborhoods and observed individual characteristics fails to provide a consistent estimate for the component of the gap accountable to ethnicity only. However, under some assumptions, the quantity of interest is set identified even when heterogeneous sorting patterns across ethnicities are allowed for and the set estimate can still be informative. A two-step estimation method is presented and applied to explain the ethnic employment differential in France, between French individuals of North African ancestry and those with non-immigrant parents. Most of the gap is not due to differences in residential location or individual characteristics, but rather to ethnicity itself.

Keywords: ethnic employment gaps, spatial sorting, set identification, discrimination, spatial mismatch

JEL: C23, J71, R23.

[§]CREST, roland.rathelot@ensae.fr, CREST - 15 Boulevard Gabriel Péri - 92245 Malakoff Cedex - France - Tel.: 33141176036 - Fax.: 33141176029

1 Introduction

In most countries, ethnic minorities are disproportionately more likely to live in distressed neighborhoods. To what extent can their lower labor market performances be attributed to ethnicity, and not to location? As Hellerstein and Neumark (2012) put it, is the issue of ethnic gaps on the labor market place-based or race-based? Should public policies target areas or ethnicities? Disentangling the effects of ethnicity from those of location may seem straightforward: introducing a measure of neighborhood quality in the employment or the wage equation would solve the problem. This paper shows that it is not the case: when ethnic minorities have spatial sorting behaviors that differ from the majority population, controlling for location does not yield consistent estimates. This article makes three contributions. First, a theoretical model shows that ethnically-asymmetric spatial sorting is likely to occur under reasonable assumptions. Second, we show that the ethnic gap attributable to ethnicity can only be partially identified and we provide a method to estimate the bounds. Finally, this method is applied to the employment gap of French workers of North African ancestry compared to those with French parents: the ethnic gap in the employment rate is mostly due to ethnicity and not to differentials in individual traits or residential location.

Residential location may affect employment status through several channels. First, the spatial mismatch hypothesis, first postulated by Kain (1968), states that living further away from jobs reduces workers' employment probability.[1] Second, human capital externalities may play a role: living in a place where everyone is unemployed makes it harder for a job-seeker to find work (Cutler and Glaeser, 1997; Bayer, Ross, and Topa, 2008; Ioannides, 2011). Because location is endogenous, individuals with different preferences or characteristics are going to sort across places. This sorting can result in statistical hiring discrimination based on residential location (redlining), decreasing the employment rate in some neighborhoods.

Spatial mismatch, local human-capital externalities or redlining may explain why, regardless of ethnicity, some areas exhibit lower employment rates than others. However, ethnic minorities have lower employment rates, regardless of where they live. Ethnic labor-market discrimination is an appealing explanation as it has been documented to hamper hiring for Blacks in the US as well as French of North African ancestry in France.[2] Recent empirical evidence for the US (Ritter and Taylor, 2011) and for France (Aeberhardt, Fougère, Pouget,

[1] See *e.g.* Ellwood (1986); Ihlanfeldt and Sjoquist (1998); Ihlanfeldt (2006); Gobillon and Selod (2007) for empirical elements about spatial mismatch in the US and in France and Gobillon, Selod, and Zenou (2007) for a comprehensive theoretical survey. See also Zenou (2009, Part 3).

[2] See Bertrand and Mullainathan (2003) and Duguet, Leandri, L'Horty, and Petit (2010) for correspondence studies on ethnic hiring discrimination, in the US and in France.

and Rathelot, 2010) suggest that, if some ethnic discrimination occurs on the labor market, it is more likely to occur at the hiring stage than at the wage-setting stage.[3] In addition to ethnic discrimination, other factors may be at work, such as cultural transmission of labor participation (Fernandez and Fogli, 2009) or differences in social-network quality.

This paper aims to separate the component of the ethnic employment gap due to ethnicity from differences in neighborhoods of residence and individual characteristics. Controlling for location seems to be important, because ethnic minorities tend to live in more distressed neighborhoods and location is believed to have a causal impact on labor-market outcomes.[4] However, we show that introducing a proxy for neighborhood quality (or, even better, fixed effects for location) in the employment equation is only valid when ethnic groups have symmetric sorting behaviors, *i.e.* when, conditional on their characteristics (whether observed or unobserved), individuals from the majority and the minority locate in places of similar quality. We present a simple theoretical framework showing that, if discrimination towards the minority exists on either the housing or the labor markets, individuals from the minority are less likely to be located in good places, conditional on their characteristics. This also implies that, conditional on neighborhood quality and observed characteristics, individuals from the minority will have better unobserved traits. Interestingly, we show that, conditional on neighborhood quality, minority individuals do not necessarily have better observed traits, so that computing ethnic differentials in observables within a location is not a valid indicator for spatial sorting. Empirically, we find that workers of North African origin tend to live in worse neighborhoods than those with French parents that have similar observed characteristics, which signals a potential asymmetric spatial sorting.

When sorting is asymmetric, controlling by location is not enough, because the expectation of unobservables is different across groups even conditional on location and observables. For instance, discriminated minorities that live in the best neighborhoods probably have extremely good unobservable traits. A second contribution of this paper is to show that, even when sorting is asymmetric, the ethnic employment gap attributable to ethnicity can still be partially identified under reasonable assumptions. When the minority locates in worse neighborhoods than the majority conditional on individual characteristics, controlling for characteristics only provides a lower bound while controlling for both characteristics and location provides an upper bound. We propose a semi-parametric two-step method to estimate these bounds.

[3]See also Abowd and Killingsworth (1984), Fairlie and Sundstrom (1999) for other evidence about the ethnic employment differentials and Neal and Johnson (1996) about the small size of the ethnic wage gap in the US.

[4]This is related to Black, Kolesnikova, Sanders, and Taylor (2013) who study the effect of controlling for location – MSAs or regions, in their case – on the black-white wage gap in the US.

Finally, we apply this method to compare individuals whose parents are North-African migrants to those with French non-migrant parents, using the French Labor Force Survey (Insee, Paris) from 2005 to 2011. While the raw ethnic employment gap is equal to 21 percentage points (55% vs. 76%), 13 to 17 percentage points are due to ethnicity only and not due to differences in observable characteristics or residential location. The main result of this study is in line with the one of Hellerstein, Neumark, and McInerney (2008) in the US case: hurdles associated with residential location are not key to explain ethnic minorities' unemployment. Using a different methodology, Gobillon, Rupert, and Wasmer (2014) also find that spatial factors are not the primary explanation of ethnic gaps in France.

The next section presents the data and some summary statistics. In section 3, a simple theoretical model explains how discrimination on the housing or on the labor market may generate asymmetric residential sorting. Moreover, we provide insights for the existence of bounds in a simple linear framework. Section 4 presents the main identification results in a more general setting and details the estimation strategy. The results are presented and discussed in section 5. Section 6 concludes.

2 Data and summary statistics

2.1 Data source and sample

In this paper, the empirical analysis is based on the French Labor Force Survey (LFS, Insee), from 2005-Q1 to 2011-Q4. The sampling frame of the LFS involves geographical cluster sampling and goes as follows. First, using information from the 1999 Census (until 2010) and the 2006 Housing Tax files (from 2010 on), *primary sample units* (of several thousands inhabitants) are selected using stratified random sampling. Then, within each of these primary units, at least one *cluster*, consisting of between 120 and 240 contiguous households, is defined. The cluster level is useful to control for very local neighborhood effects. Some local characteristics affecting one household in a given cluster will undoubtedly affect the other households in the same cluster. Note also that, by definition, clusters are stricty included in municipalities (the smallest administrative unit), so that the inhabitants of a given cluster are assumed to be supplied with comparable public goods. For privacy reasons, the data associate each household to a cluster ID, but the geographical location of clusters is unknown.

In this study, we compare a minority group to a majority group. In line with the literature, the minority group we focus on have at least one parent born with a North African citizenship.[5] As a comparison, some results are also given on individuals with Southern European parents. The majority group have both parents born French in France. Individuals from both groups are all born in France and have a French citizenship. Therefore, the analysis deals with individuals who are not migrants themselves. There are two reasons for this: first, education or labor experience acquired in France or abroad may be viewed differently by French employers; second, a certain command of French may also account for variability in employment rates.

Our outcome of interest is the employment status.[6] Gender, education and age are used as explanatory variables. The education variable reflects both the level and the field of the obtained degree, resulting in twenty dummies. Age and age squared are included in all specifications. The sample is restricted to individuals aged 20 to 59.

2.2 Disparities in individual characteristics

Table 1 presents some summary statistics on the three subpopulations: the reference group in the first column, French individuals of North African and Southern European ancestries in columns 2 and 3.

[Insert here Table 1]

The most striking fact is that individuals of North African origin have worse labor market outcomes than other groups; they are less likely to work (55% vs. 76%) and those who do earn around 16% less. They are less likely to be executive or professional (6% vs. 13%), to occupy technical or sales occupations (15% vs. 20%) or to work in agriculture (0% vs. 2%). They are slightly more likely to be office workers or blue collars and far more likely to have no reported occupation (31% vs. 18%). By constrast, the employment rate of the individuals of Southern European origin is very close to the one of the reference population. They earn 4% less and are less often employed in executive positions (10% vs. 13%).

The low employment rate in the group of North African ancestry is to some extent linked to their individual characteristics. First, they have less education: 4.6% of them hold a Master's degree, a diploma from a Elite university, or a PhD, while 7.8% of the reference population do so. They also frequently end up with no qualifications at all. 33% of them dropped out of the system with no diploma at all or the basic *Brevet des Collèges* (taken at the end of the 9th grade), while this is only the case for 24% of the French with French parents. Second,

[5] Since 2005, the LFS includes questions about one's parents' nationality at birth.

[6] The analysis has been replicated using the log-wages as the outcome. However, in line with the previous literature, the ethnic wage gap is entirely explained by differences in education and age. Detailed results are available from the author upon request.

this minority group is on average 8 years younger than the rest of the population. Individuals of Southern Europe origin do not differ much from those with French parents on all these characteristics. They are slightly less likely to hold a postgraduate degree and their average age is only 9 months lower than the majority group.

2.3 Spatial disparities

Clusters differ in several ways. Table 2 provides some summary statistics at the cluster level. The first two columns provide descriptive statistics for all clusters, unweighted (column 1) or weighted by cluster size (column 2). Weighting by size provides an insight about the typical cluster from the point of view of an individual. The last two columns provide descriptive statistics for clusters in which the majority group coexist with at least one individual from a minority: North African ancestry (column 3) or Southern European ancestry (column 4). We name these kinds of clusters "mixed clusters" and denote the set of mixed clusters with the minority of North African ancestry as $\mathcal{M}$.

[Insert here Table 2]

Depending on the sampling frame, the local response rate and local characteristics, some clusters may be substantially larger than others. Overall, the median cluster contains 36 individuals. From the point of view of individuals, the median size is 62. Mixed clusters are larger, but the direction of causality is not clear, as larger clusters are more likely to exhibit diversity. Figure 1 displays the distribution of the number of individuals per cluster in more detail.

[Insert here Figure 1]

The figure displays a large diversity of situations, with a minimum of 1 observation, a mode at 15 observations and a maximum at 196 observations.

As clusters can be considered, by construction, as independent draws over the French territory of groups of individuals living contiguously, the heterogeneity of clusters reflects social and ethnic residential disparities.[7] The first and third quartiles of the employment rate are 67% and 83%, which reflects the heterogeneity of economic conditions across clusters. Only half of the clusters mix individuals from both the majority group and the minority with North African ancestry. In the median cluster in which they are present, individuals from this minority group represent 5% of the inhabitants.

[7] More elements about the distribution of the employment rate and the share of the minority group across clusters are provided in appendix A.1.

In mixed clusters where people with North African ancestry are present, the employment rate is 3 percentage points lower. Figure 2 shows the results of non-parametric (kernel) regressions of the employment rate (left panel) and of the share of dropouts (right panel) on the minority share in the cluster.

[Insert here Figure 2]

The employment rate and the share of dropouts are displayed for both the majority population and for individuals with North African parents. For the employment rate, both curves exhibit a pronounced downward trend. For the share of dropouts, both curves are increasing (past a slight initial decrease) and the curve of the majority group is steeper than the one for the minority.

The key descriptive facts are the following. First, people with North African parents have significantly lower employment chances, which might be explained to some extent by lower levels of education and less experience. Second, they seem to be more concentrated in areas in which the employment rate is lower.

3 Conceptual framework

This subsection presents a simple theoretical framework of spatial sorting. Individuals may live in two locations, center c or suburbs s. Locations provide different levels of amenities depending on individuals' employment status and also affect the probability of employment. The population is assumed to be composed of two ethnic groups: the majority group 0 and the minority group 1. The minority group potentially faces housing- and labor-market discrimination. This simple model shows that such discrimination can generate asymmetric sorting: minority individuals with similar characteristics will live more often in the suburbs, which offer less amenities and provide lower probability to work.

Imagine first that there exist only group-0 individuals, in number N_0. The utility of living in location $a \in \{c, s\}$ depends on the individual's probability to work, the utility associated with the employment status and the rent. An individual i's probability to work depends on her observed characteristics X_i, her unobserved characteristics u_i and her location a: $p_i(a) = p(X_i, u_i, a)$. Living in the suburbs is assumed to reduce one's probability of working, so that $p_i(s) < p_i(c)$. This effect results from several phenomena. First, spatial mismatch implies that living further away from jobs harms one's probability to learn about the existence of vacancies (Kain, 1968). Second, redlining, as defined in Zenou and Boccard (2000), implies that workers living in the suburbs may suffer from some of kind of spatial discrimination, irrespective of ethnicity. Different reasons may lead to the existence of this discrimination, whether it is based on prejudice about inhabitants' human capital or propensity to commit crimes (Zenou and Boccard, 2000) or statistical discrimination due to a productivity loss associated with commuting (Zenou, 2002). On top of these two effects, equilibrium network effects may also play a role: peer effects, combined with sorting, are likely to exacerbate the existing differences between the center and the suburbs. Finally, we assume for simplicity that $p_i(s) = \rho_p p_i(c)$ with $\rho_p \in (0,1)$, and we denote $p_i = p_i(c)$ and $p(X_i, u_i) = p(X_i, u_i, c)$

Being employed in c brings utility $\bar{v}$, while being employed in s brings a lower utility $\underline{v} < \bar{v}$. The utility of being unemployed is equal to $\rho_v \in (0,1)$ times the utility of being employed. Note that, together, these two assumptions mean that the gap in utility between employed and unemployed individuals is larger in the center than in the suburbs. This is justified by the fact that living in the center provides amenities that may be valued more by employed people. In particular, given the geography of jobs and the structure of transportation systems, living in the center entails lower commuting times than living in the suburbs; see *e.g.* the benchmark model presented in Zenou (2009, chap.1).

Housing in s is infinitely supplied, so that rents in s are equal to zero. Conversely, housing in the center is assumed to be constrained: the C available units are allocated to the highest bidders. Rents in c are denoted as h. The utilities of group-0 individuals can thus be written as:

$$U_i^0(c) = p_i\bar{v} + (1-p_i)\rho_v\bar{v} - h$$
$$U_i^0(s) = \rho_p p_i\underline{v} + (1-\rho_p p_i)\rho_v\underline{v}$$

Note that $U_i^0(c)$, $U_i^0(s)$ and the difference $\Delta U_i^0 \doteq U_i^0(c) - U_i^0(s)$ are all increasing in p_i. Therefore, individuals with the best characteristics are those with the highest gains from being in the center rather than in the suburbs. These individuals secure housing in the center by bidding

$$h = \rho_v(\bar{v} - \underline{v}) + (1-\rho_v)(\bar{v} - \rho_p\underline{v})p^* \text{ , with } p^* = F_0^{-1}\left(\frac{N_0 - C}{N_0}\right)$$

where $F_0(.)$ is the cumulative distribution function of p_i in the majority group. Note that this distribution depends on the distributions of X_i and u_i.

Now, N_1 individuals from the minority group 1 are introduced in the market. The major difference across ethnic groups in this model is that individuals from the minority may suffer from housing and/or labor market discrimination. Housing discrimination is assumed to increase their rent from h to h/τ_h, with $\tau_h \in (0,1)$. Labor-market discrimination is assumed to decrease their probability to work by a factor $\tau_p \in (0,1)$, to $\tau_p p(X_i, u_i, a)$. Consistently with the assumptions made throughout the paper, the distribution of u conditional on X is assumed to be identical across groups. However, ethnic groups may differ in terms of their observed characteristics X. In the minority group, the cumulative distribution function of p_i (the pre-discrimination probability to work) is denoted as $F_1(.)$. The individual utilities for an individual i from the minority can be written as:

$$U_i^1(c) = \tau_p p_i \bar{v} + (1 - \tau_p p_i)\rho_v \bar{v} - h/\tau_h$$
$$U_i^1(s) = \rho_p \tau_p p_i \underline{v} + (1 - \rho_p \tau_p p_i)\rho_v \underline{v}$$

The utility gains of living in the center rather than in the suburbs $\Delta U_i^g \doteq U_i^g(c) - U_i^g(s)$, with $g \in \{0,1\}$, are different across ethnic groups:

$$\Delta U_i^0 = \rho_v(\bar{v} - \underline{v}) + (1 - \rho_v)(\bar{v} - \rho_p \underline{v})p_i - h$$
$$\Delta U_i^1 = \rho_v(\bar{v} - \underline{v}) + (1 - \rho_v)(\bar{v} - \rho_p \underline{v})\tau_p p_i - h/\tau_h$$

The existence of discrimination on either the housing or the labor markets makes individuals from the majority experience larger gains from living in the center rather than in the suburbs, conditional on characteristics (summarized by p_i), compared to minority individuals. Applying the same reasoning as above, the housing slots in the center are allocated to the highest bidders. Denote N_0^c and N_1^c the number of individuals in each group that will live in the center, and p_0^* and p_1^* the thresholds beyond which these individuals choose to live in the center. The rent h is such that :

$$h = \rho_v(\bar{v} - \underline{v}) + (1 - \rho_v)(\bar{v} - \rho_p \underline{v})p_0^* \text{ , with } p_0^* = F_0^{-1}\left(\frac{N_0 - N_0^c}{N_0}\right) \tag{1}$$

$$= \tau_h \rho_v(\bar{v} - \underline{v}) + \tau_h \tau_p (1 - \rho_v)(\bar{v} - \rho_p \underline{v})p_1^* \text{ , with } p_1^* = F_1^{-1}\left(\frac{N_1 - N_1^c}{N_1}\right) \tag{2}$$

Combining equations (1) and (2) leads to:

$$p_1^* = \frac{p_0^*}{\tau_h \tau_p} + \frac{1 - \tau_h}{\tau_p \tau_h} \frac{\rho_v(\bar{v} - \underline{v})}{(\bar{v} - \rho_p \underline{v})(1 - \rho_v)} \tag{3}$$

Equation (3), together with the definitions of p_0^* and p_1^* and the constraint that $N_0^c + N_1^c = C$, pinpoints a value for N_0^c and N_1^c.[8]

If there exists labor-market discrimination but no housing discrimination, the cutoff in terms of actual (post-discrimination) probability to work is equal across groups, and sorting arises only because of the existence of a wedge between the pre-discrimination and the post-discrimination probability to work for minority individuals. If there exists housing discrimination, higher

[8]Note that, without further restriction on the distribution of characteristics in the minority population, it is a priori possible to have $N_1^c = 0$.

rents for minority workers shift up the probability to work required to get access to housing in the center. Provided that $\tau_h < 1$ or $\tau_p < 1$, the minimum value of p_i such that the individual chooses to settle in the center is higher in group 1 than in group 0. The existence of ethnic discrimination on either the housing or the labor market creates an ethnic asymmetry in the location choices. Workers from the minority group need better characteristics in order to choose to locate in the center. Therefore, conditional on X_i and u_i (*i.e.* conditional on p_i), the quality of the neighborhood in which workers settle is lower for the minority than for the majority. Given that u_i is assumed to be equally distributed across groups conditional on X_i, the previous statement remains true conditional on X_i only.

In the case of two locations (center vs. suburbs), it is easy to show that individuals from the minority have, on average, higher unobservables u_i than individuals from the majority, conditional on observables X_i and location a. Because $p_1^* > p_0^*$ and the distribution of $u_i|X_i$ is assumed identical across groups,

$$\mathbb{E}[u_i|p(X_i,u_i) > p_1^*, X_i] > \mathbb{E}[u_i|p(X_i,u_i) > p_0^*, X_i]$$
$$\mathbb{E}[u_i|p(X_i,u_i) < p_1^*, X_i] > \mathbb{E}[u_i|p(X_i,u_i) < p_0^*, X_i]$$

Intuitively, we might believe that a similar statement should hold for observables: a higher cutoff for minority individuals would induce minority individuals to be better on average in each neighborhood. However, because the distributions of observables may strikingly differ across groups, no such statement can easily be made. To go further, we develop a particular case in appendix A.2. in which we impose linear dependence between the variables to get the analysis tractable. Using this linear model, we find that, for some values of the parameters, minority individuals may have lower observables than the majority population, even conditional on location. A necessary condition (in that linear model) for this to occur is that the marginal distributions of observables should be such that minority individuals have much lower observables than majority individuals.

4 Set identification and estimation of the bounds

This section shows in a general setting that, when location has a causal effect on employment and when ethnic groups have different sorting behaviors conditional on their characteristics, the causal impact of the ethnic group on employment is only partially identified. Interestingly for applied purposes, identification does not rely on functional assumptions or on the ability to measure neighborhood quality. An estimation method for the bounds is then presented. This method, based on matching, only gives a limited role to parametric assumptions and is especially appealing when the number of observations per spatial unit is small, as the incidental parameter problem is accounted for.

4.1 Assumptions and identification issues

The notation used in this section is adapted from the Rubin framework (Rubin, 1974). The outcome Y, in our case the employment status, is a random variable that takes two potential values $Y(0)$ and $Y(1)$ depending on the group the individual belongs to. For a given individual, only the realization of $Y(0)$ (resp. $Y(1)$) is observed if the individual belongs to group 0 (resp. 1). The quantity of interest of this paper is the ethnic gap $\mathbb{E}(Y(1) - Y(0)|g = 1)$. The average employment rates in groups 0 and 1 are the direct empirical counterparts of $\mathbb{E}(Y(0)|g = 0)$ and $\mathbb{E}(Y(1)|g = 1)$. The difficult part is the identification of the counterfactual $\mathbb{E}(Y(0)|g = 1)$. As is well known from the evaluation literature, no direct empirical counterpart is available for this quantity without further assumptions about how unobervables differ across groups.

We first make an assumption about the form of the model. The outcome is assumed to depend only on observable characteristics X, unobservable characteristics u and the cluster of residence a.

Assumption 1. [Ignorability on observables and unobservables].

$$\forall (X, u, a), \mathbb{E}(Y(0)|X, u, a, g = 0) = \mathbb{E}(Y(0)|X, u, a, g = 1)$$

Even though this assumption looks like a traditional ignorability assumption, it is much weaker, as one conditions on both observables and unobservables.

The second assumption, about the distribution of unobservables conditional on observables, is probably the most restrictive one. $\Phi_{.}(.)$ denotes cdf and $\phi_{.}(.)$ pdf.

Assumption 2. [Conditional equidistribution of unobservables].

$$\forall (X, u), \Phi_{u|X,g=0}(u|X, g = 0) = \Phi_{u|X,g=1}(u|X, g = 1)$$

Conditional on our set of individual characteristics, unobservable determinants u of employment have to be distributed in the same way across groups. This assumption, which is frequently omitted in the decomposition literature (Fortin, Lemieux, and Firpo, 2011), is required to isolate the component of the ethnic emloyment rate gap that is due to ethnicity. One can imagine many stories why this assumption may be violated and some of them are discussed in subsection 5.3.

The preceding assumptions are not sufficient to identify the quantity of interest. One needs to make another assumption about the influence of location, or about sorting. Most studies performing ethnic-gap decompositions ignore the influence of location. What they implicitly assume is that the influence of location on employment, conditional on observables and unobservables, is of second order.

Assumption 3. [No conditional influence of location].

$$\forall (X, u, a), \mathbb{E}(Y(0)|X, u, a, g) = \mathbb{E}(Y(0)|X, u, g)$$

In this case, one is able to proceed to decomposition in a classical non-spatial way. Under assumptions 1-3, $\mathbb{E}(Y(0)|g = 1)$ can be showed to be equal to $\overline{Y}$ with:

$$\overline{Y} \doteq \int \mathbb{E}(Y(0)|X, g = 0) d\Phi_{X|g=1}(X|g = 1)$$

If the influence of location cannot be ignored, another possibility is to assume that spatial sorting is symmetric. Denote the probability to be located in neighborhood a conditional on X, u and g as $\mathbb{P}[a|X, u, g]$. When sorting is symmetric, $\mathbb{P}[a|X, u, g]$ does not depend on ethnicity either, conditional on observable and unobservable traits. Moreover,

Assumption 4. [Symmetric spatial sorting].

$$\forall (X, u, a), \mathbb{P}[a|X, u, g = 0] = \mathbb{P}[a|X, u, g = 1]$$

In the proof (in appendix A.3.) we also show that this assumption implies that the distribution of the unobservables conditional on individual characteristics and location does not depend on ethnicity. Under assumptions 1-2, 4, we have:

$$\mathbb{E}(Y(0)|g = 1) = \underline{Y} = \overline{Y}$$

with:

$$\underline{Y} \doteq \int \mathbb{E}(Y(0)|X, a, g = 0) d\Phi_{X,a|g=1}(X, a|g = 1)$$

This leads to an important practical conclusion. When sorting is symmetric, bounds collapse, and regressions will provide the same result whether location is controlled for or not.

In this paper, assumptions 3 and 4 are considered to be unacceptable, as the influence of location on employment cannot be ignored and spatial sorting is a priori asymmetric, as evidenced in the previous section. Following the deductions of the theoretical framework, we rather assume that having access to the best location is harder for the minority.

Assumption 5. [Asymmetric spatial sorting]. *Individuals from the minority group are less likely, conditional on X and u to be located in areas that cause higher employment.*

$$\forall (X, u), \mathbb{E}(Y(0)|X, u, g = 0) \geq \mathbb{E}(Y(0)|X, u, g = 1)$$

Unobservables are, conditional on X and a, higher in group 1 than in group 0.

$$\forall (X, a), \mathbb{E}(Y(0)|X, a, g = 0) \leq \mathbb{E}(Y(0)|X, a, g = 1)$$

This assumption, which is much weaker than assumption 4, leads to the main identification result of this paper.

Proposition 1. *Under assumptions 1-2 and 5,* $\mathbb{E}(Y(0)|g=1)$ *admits bounds that can be identified.*

$$\underline{Y} \leq \mathbb{E}(Y(0)|g=1) \leq \overline{Y}$$

The proof is provided in appendix A.4.

Note that, under assumptions 1 and 2, we may test assumption 4 against assumption 5. If the null hypothesis $\underline{Y} = \overline{Y}$ is rejected, so is the assumption of symmetric spatial sorting. The sign of $\overline{Y} - \underline{Y}$ provides the direction of the asymmetry of the sorting pattern. However, as illustrated more clearly in the linear case, the magnitude of the difference cannot be directly used to infer the extent of location effects (see appendix A.2).

4.2 Estimation

This section is dedicated to the estimation of the bounds $\underline{Y}$ and $\overline{Y}$. The following assumption aims at simplifying the empirical analysis by ruling out the curse of dimensionality.

Assumption 6. [Single index]. *The influence of observables can be captured by a single index* $s(X) = X\theta \in \mathbb{R}$.

$$\exists \theta \in \mathbb{R}^K \ s.t. \ \forall (X, a), \mathbb{E}(Y(0)|X, a, g=0) = \mathbb{E}(Y(0)|X\theta, a, g=0)$$

The estimation proceeds in two steps: first, estimate θ in the majority group; second, estimate the bounds based on the observations for both groups.

$\mathbb{E}(Y|X, a)$ is estimated by conditional logit.[9]

$$\mathbb{E}(Y|s(X), a) = \Lambda(X\theta + \vartheta_a)$$

The influence of residential location is assumed to be restricted to additive fixed effects ϑ, which shift the intercept without interacting with observable characteristics. Because the incidental parameter problem can be solved in this way, Λ is assumed to be the logistic cdf. Note that, under the specification assumptions, this estimation allows one to recover unbiased estimates for θ, but not for ϑ.

[9]The conditional logit has been recoded in `R` to account for the fact that some clusters are relatively large. To speed up computations, the denominator of the likelihood is computed in `C` and interfaced with `R`, using the package `Rcpp` (Eddelbuettel and François, 2011). The code of the estimation function is available upon request.

Once $s(.)$ has been estimated for group 0, we can proceed to the estimation of $\underline{Y}$ and $\overline{Y}$, by matching individuals from the minority with those from the majority. Kernel matching is used here, with a uniform (caliper matching) or an Epanechnikov kernel $K(.)$. A bandwidth parameter ν must be chosen beforehand. G_0 (G_1) is the set of majority (minority) individuals. For a pair of individuals $(i, j) \in G_0 \times G_1$, we define the weights $w_{ij} = K([s(X_j) - s(X_i)]/\nu)$.

For $\underline{Y}$, the algorithm goes as follows.

1. Consider an individual $i \in G_1$, with characteristics X_i and location a_i.
2. Define the set $\underline{\mathcal{J}(i)} = \{j \in G_0, a_i = a_j \text{ and } |s(X_j) - s(X_i)| < \nu\}$ of the G_0 individuals living in a_i such that weights are strictly positive.
3. Compute and store the quantity

$$\underline{y(i)} = \frac{\sum_{j \in \underline{\mathcal{J}(i)}} w_{ij} Y_j}{\sum_{j \in \underline{\mathcal{J}(i)}} w_{ij}}$$

and carry on for the next individual in G_1.

For $\overline{Y}$, the algorithm is simpler.

1. Consider an individual $i \in G_1$, with characteristics X_i.
2. Define the set $\overline{\mathcal{J}(i)} = \{j \in G_0, |s(X_j) - s(X_i)| < \nu\}$ of the G_0 individuals such that weights are strictly positive.
3. Compute and store the quantity

$$\overline{y(i)} = \frac{\sum_{j \in \overline{\mathcal{J}(i)}} w_{ij} Y_j}{\sum_{j \in \overline{\mathcal{J}(i)}} w_{ij}}$$

and carry on for the next individual in G_1.

The estimators $\widehat{\underline{Y}} \doteq 1/N_1 \sum_{i \in G_1} \underline{y(i)}$ and $\widehat{\overline{Y}} \doteq 1/N_1 \sum_{i \in G_1} \overline{y(i)}$ are considered as empirical counterparts of the bounds $\underline{Y}$ and $\overline{Y}$ of the quantity of interest. The technical conditions under which these estimators converge to $\underline{Y}$ and $\overline{Y}$ are not detailed here but they can be directly adapted from Heckman, Ichimura, and Todd (1998).

In order to account for the fact that there are two steps in the estimation, inference is performed by bootstrap.[10] At each iteration, we draw individuals with replacement and perform the estimation. Standard errors and confidence intervals are computed using 100 iterations.

[10] Abadie and Imbens (2008) prove that bootstrap fails to provide valid inference for matching procedures with a fixed number of neighbors. For kernel matching, however, they conjecture that bootstrap provide valid inference.

4.3 Support issues

This matching approach is similar to the one adopted by Nopo (2008) and is subject to the same problems. The main issue, also detailed in Fortin, Lemieux, and Firpo (2011), is the potential lack of commun support across groups. Because there are few observations per spatial unit, some spatial units contain only minority individuals. In these cases, $\underline{\mathcal{J}(i)}$ will be empty. Even when the spatial unit contains individuals of both groups, $\underline{\mathcal{J}(i)}$ or $\overline{\mathcal{J}(i)}$ can be empty if no individual of the majority group has characteristics which are close enough to the ones of the minority individual. Note that the single-index assumption helps alleviate the latter issue.

Support issues are problematic when they do not occur randomly. Let $\tilde{G}_1 = \{i \in G_1, \underline{\mathcal{J}(i)} \neq \emptyset\}$, with $N_{\tilde{G}_1} = Card(\tilde{G}_1)$. In practice, as in Nopo (2008), the estimator will be computed over the subsample $\tilde{G}_1$ for which there is no support issue, $\widehat{\underline{Y}} = 1/N_{\tilde{G}_1} \sum_{i \in \tilde{G}_1} \underline{y(i)}$. To assess the extent to which lack of support is really an issue, the percentage of cases with support problems $p_s = 1 - N_{\tilde{G}_1}/N_1$ as well as the relative gaps are reported:

$$\delta_y = 1 - \frac{\sum_{\tilde{G}_1} Y_i/N_{\tilde{G}_1}}{\sum_{G_1} Y_i/N_1}$$
$$\delta_s = 1 - \frac{\sum_{\tilde{G}_1} s(X_i)/N_{\tilde{G}_1}}{\sum_{G_1} s(X_i)/N_1}.$$

We can build another estimator for the lower bound that reduces the number of observations excluded by the bandwidth. When a control individual is not found for a given individual of interest, one can select as a control the nearest individual with a score lower than the individual of interest. The resulting estimator will still be a lower bound, though a looser one. Formally, this means that for an individual i of group 1, the control group $\underline{\mathcal{J}(i)}$ is replaced, if it is empty, by $\underline{\mathcal{J}'(i)} = \{j \in G_0 : s(X_j) = \max_k\{s(X_k) : a_k = a_i \text{ and } s(X_k) < s(X_i) - \nu\}\}$. We define:

$$\underline{y(i)}^* = \begin{cases} \underline{y(i)} & \text{if } \underline{\mathcal{J}(i)} \neq \emptyset \\ \underline{y'(i)} & \text{if } \underline{\mathcal{J}(i)} = \emptyset \text{ and } \underline{\mathcal{J}'(i)} \neq \emptyset \end{cases} \quad \text{with} \quad \underline{y'(i)} = \frac{\sum_{j \in \underline{\mathcal{J}'(i)}} Y_j}{Card(\underline{\mathcal{J}'(i)})}$$

The new estimator can be expressed as:

$$\widehat{\underline{Y}}^* = \frac{\sum_{i \in \tilde{G}_1^*} \underline{y(i)}^*}{Card(\tilde{G}_1^*)} \quad \text{with} \quad \tilde{G}_1^* = \{i \in G_1, \underline{\mathcal{J}(i)} \cup \underline{\mathcal{J}'(i)} \neq \emptyset\}$$

p_s^*, δ_y^* and δ_s^*, relating to the estimator $\widehat{\underline{Y}}^*$ can also be computed.

5 Results

5.1 Estimation of the employment equation

Table 3 reports the results of the estimation of the employment equation for the conditional logit model, in which fixed effects for geographic clusters are included.

[Insert here Table 3]

Most coefficients have the usual sign. Females are significantly less often in employment than males. The influence of age on employment displays an inverted U-shaped curve. Education is measured both in terms of level and field, relatively to the *Bac*, the degree obtained at the end of high school and required to enter university. Health degrees are associated with the highest probability of being employed. Higher levels of education and degrees with scientific or industrial majors seem to increase one's propensity to work. Having no diploma at all is associated with strongly lower probability of working.

5.2 Estimation of the bounds of the counterfactual employment level

Now that the employment equation has been estimated on population 0, we can estimate the bounds of $\mathbb{E}(Y(0)|g = 1)$, $\underline{Y}$ and $\overline{Y}$. Table 4 reports the estimates of the bounds. In columns 1 and 3, the matching kernel is uniform, while in columns 2 and 4, an Epanechnikov kernel is used. In columns 1 and 2, the bandwidth is set to .05. In columns 3 and 4, the bandwidth is set to .10.

[Insert here Table 4]

The employment gap between the two groups is on average equal to 20.7 percentage points, with sample means equal to 75.7% in group 0 and 55.0% in group 1. Restricting the sample to clusters in which both groups are present is necessary. This restriction does not affect much the sample means. The employment rate of majority in these mixed clusters is equal to 74.2%; for the minority group, it is unchanged: 55.0%. Within mixed clusters, the raw ethnic employment gap is equal to 19.2 percentage points.

Now, we can present the result to the main question. Some of the 19 percentage-point ethnic employment gap may be explained by the fact that individuals of the minority are on average less educated, younger and live in more distressed areas. What part of the employment gap can be attributed to ethnicity, rather than to differentials in observable traits or residential location? Under the assumptions detailed above, upper and lower bounds can be estimated: the counterfactual employment rate $\mathbb{E}(Y(0)|g = 1)$ is between 68% and 72% in all specifications presented in table 4. This means that there is still between 13 and 17 percentage points that are not due to observed characteristics or residential location.

In Table 4, we also check empirically that $\widehat{\overline{Y}}$ is higher than $\widehat{\underline{Y}}$. This inequality should be seen as a test of consistency of Assumption 5 (asymmetric sorting). If sorting were symmetric, all groups would have the same access to all neighborhoods, conditional on individual characteristics. In that case, the bounds should be equal to each other, which is not the case here. The asymmetry goes in the direction that we conjectured using summary statistics: conditional on age and education, French individuals with North African ancestry are less likely to be located in good neighborhoods than French with French parents.

Should we be concerned about support issues? In the first column, because the bandwidth of .05 is narrow, 10.1% of the minority individuals cannot be matched for the estimation of $\widehat{\underline{Y}}$. Matched individuals are on average 1.8 percentage points more likely to work than the average minority individual. The support issue is less severe when $\widehat{\underline{Y}}^*$ is used instead. Only 4.0% of the observations are not matched, with a gap between matched and average individuals lower than 1 percentage point. For the other two specifications, the support issue looks even less problematic, even for the estimator $\widehat{\underline{Y}}$. This is consistent with lower differences between the estimates of $\widehat{\underline{Y}}$ and $\widehat{\underline{Y}}^*$ in columns 1 and 2. Overall, all four columns tell the same story.

A final concern regarding common support is that, if controls and treated are too different, the analysis might rely too much on a small number of control observations; for instance if one observation is used as a control for many treated observations. To investigate this issue, we compute: (i) the number of control observations which are used at least once, (ii) the number of control observations j such that $\sum_i w_{ij} > 1$. In the case of the Epanechnikov kernel with bandwidth equal to .10, 63,491 observations are at least used once as controls (for 8432 treated observations). Among them, only 504 have a summed weight higher than 1. On these observations, $\sum_i w_{ij}$ has a mean of 1.8. Therefore, we expect our results to be reasonably robust to this issue.

5.3 Discussion

The preceding results show that the ethnic employment gap is hardly explained by differentials in individual characteristics or in residential location. The fact that individuals with North African parents are more likely to live in distressed areas than the majority group is not a sufficient explanation for their strikingly lower employment levels. The findings of this paper have important policy consequences. France has a long tradition of ignoring the ethnic dimension of inequalities, focusing on social or geographical dimensions. Public policies are designed accordingly: while fiscal incentives aim to achieve more economic equality on a social basis (for instance, subsidizing the hiring of low-skilled workforce; see Crépon and Desplatz (2001)) or a geographical one (subsidizing economic activity in distressed areas through enterprise-zone-like policies; see Givord, Rathelot, and Sillard (2013)), policy makers explicitly refuse to consider ethnicity among the possible criteria.[11] This paper suggests that some specific ethnically-targeted policy might be necessary to bridge ethnic gaps on the labor market.

Designing adequate policy interventions requires being able to identify the economic mechanisms at work. While one contribution of this paper is to rule out that differences in residential location are the key explanation for ethnic differentials on the labor market, the next important question is how to explain the massive residual ethnic gap? The leading explanations in the literature are discrimination (statistical or taste-based), the existence of ethnic-specific cultural traits, and ethnic differences in the quality of social networks. We provide two additional pieces of evidence that suggest that the importance of culture and networks should not be over-stated.

In the case of the US, Fernandez and Fogli (2009) argue that a significant part of the employment gap between women of foreign origin and native women can be explained by the female employment pattern in the woman's country of ancestry. If this phenomenon existed in France, it would cumulate with the other factors underlying the ethnic gaps: discrimination and differences in social networks have no reason to be *a priori* restricted to men. Therefore, the employment differential due to ethnicity should be higher for women than for men. The second column in Table 5 shows the estimated bounds of $\mathbb{E}(Y(0)|g=1)$ when the sample is restricted to women (both for the majority and the minority populations).[12] Female employment rates in both groups are 4 to 5 percentage points lower than in the whole population. While the total gap amounts to 19.7 percentage points, the gap that can be attributed to ethnicity is between 14 and 16 percentage points. These figures are very similar to the 13-17-percentage-point interval obtained on the whole sample. Since stratifying by gender does not change the results, we conjecture that cultural transmission from women's country of ancestry is not likely to be a crucial factor.

[Insert here Table 5]

[11] In a discourse dating from 2008, Nicolas Sarkozy declared "If the question of measuring inequalities and discriminations relating to national origin is open, the question of a voluntarist public action based on ethnic or religious criteria should be closed. [...] if we reduce all the social differentials, we will reduce at the same time all the ethnic, religious and cultural differentials." See Sarkozy (2008).

[12] For this column, the employment equation has been re-estimated on the subsample of women only.

Social networks are believed to play a prominent role in one's access to jobs (Montgomery, 1991; Topa, 2001; Calvo-Armengol and Jackson, 2004). By definition, immigrants have recently settled in their country of residence and may not have had the time to develop a deep social network.[13] Networks of lower quality might prove detrimental to their children. A first issue with this theory is that it hardly explains the differences of employment rates across ethnic minorities. Column 3 presents the decomposition for individuals of Southern European origin (instead of North Africa) and shows, in a nutshell, that there are no difference between this ethnic minority and the majority group.[14] One could argue that Southern European migrants arrived earlier than North African ones, but it is difficult to believe that a difference of ten or twenty years would provide such an impressive difference. Network theory also predicts a difference between the employment rates of individuals with two immigrant parents and those with only one immigrant parent. Column 3 shows the results when only individuals with two immigrant parents are considered as the minority group, while, in column 4, only those with an immigrant father and a non-immigrant mother are treated as the minority group. Columns 1 and 4 are very similar, which suggests that ethnic differentials in networks are not the main story.

Two warnings concerning the interpretation of the results should be explicated. First, consistent with the spatial-mismatch or the redlining stories, we assumed that the impact of residential location on employment was the same for both ethnic groups. If there exist some location-interacted-with-ethnicity effects, our analysis will count them as ethnically driven. For instance, say that discrimination is the key factor to explain ethnic gaps. We will not be able to distinguish whether discrimination is based on ethnicity only or on ethnicity interacted with location. These findings also relate to the social network literature in which local networks are shown to be, to some extent, ethnic-specific. Ethnic-related factors explaining labor market outcomes, and their interactions with location, have yet to be disentangled.

Second, education is assumed to be exogenous, while education acquisition may actually differ between the two ethnic groups. OECD (2012) presents the results of the PISA 2009 survey on 15-year-olds with immigrant parents. There are large discrepancies in the performance in reading, maths and science between children with immigrant parents and those with native parents. These differences tend to decrease when socio-economic background is accounted for, but remain sizable in most countries, France included. If this phenomenon results, for instance, from differentials in school quality or teachers' attitudes, it is possible that, conditional on

[13] There is an ethnic dimension in social networks, as evidenced by Topa (2001); Edin, Fredriksson, and Åslund (2003); Munshi (2003).

[14] Actually, in this case, $\widehat{\widehat{Y}}$ is slightly lower than $\underline{\widehat{Y}}$, which tends to show that individuals of Southern European origin are located in better neighborhoods than the majority, conditional on their characteristics.

their initial traits, education acquisition is more difficult for children from ethnic minorities. Thus, more "talent" would be needed to acquire the same level of education. In that case, our estimates of the part of the ethnic gaps that is inherently due to ethnicity would be biased downwards: unbiased ethnic gaps due to ethnicity would be even larger.

6 Conclusion

This paper makes two contributions. First, as discussed in the previous section, ethnic employment gaps should be mostly attributed to ethnicity, and not to differentials in residential location or observable characteristics. The second contribution is a methodological one, when the goal is to isolate the part of the ethnic gap that is due to ethnicity only when the outcome depends on both individual and neighborhood characteristics. Whether spatial sorting is symmetric across ethnic groups is shown to be crucial. If sorting is symmetric, valid estimates can be obtained with or without neighborhood quality controls. If sorting is asymmetric, inclusion or exclusion of neighborhood quality controls lead to two differents estimates and we show under some assumptions that these two estimates are bounds for the quantity of interest. The direction of the sorting, *i.e.* which group is more likely to locate in better neighborhoods, can be infered from these bounds.

Acknowledgements

I would like to thank Laurent Gobillon, the editor and two anonymous reviewers for their precious suggestions. I also thank Romain Aeberhardt, Nicoletta Berardi, Tatiana Coutto, Denis Fougère, Thomas Le Barbanchon, Thierry Magnac, Eric Maurin, David Neumark, Benoit Schmutz, Maxime To, Yves Zenou and participants of many seminars and conferences for useful comments and discussions, as well as Alex Olssen for his patient reading and his suggestions. This text is a deeply revised version of *Wrong origin or wrong neighborhood: Explaining lower labor market performance of French individuals of foreign origin*. Any opinions expressed here are those of the author and not of any institution.

References

Abadie, A., and G. W. Imbens (2008): "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, 76(6), 1537–1557.

Abowd, J. M., and M. Killingsworth (1984): "Do Minority/White Unemployment Differences Really Exist?," *Journal of Business & Economics Statistics*, 2, 64–72.

Aeberhardt, R., D. Fougère, J. Pouget, and R. Rathelot (2010): "Wages and Employment of French Workers with African Origin," *Journal of Population Economics*, 23(3), 881–905.

Bayer, P., S. L. Ross, and G. Topa (2008): "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes," *Journal of Political Economy*, 116(6), 1150–1196.

Bertrand, M., and S. Mullainathan (2003): "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 94(4), 159–190.

Black, D., N. Kolesnikova, S. Sanders, and L. Taylor (2013): "The Role of Location in Evaluating Racial Wage Disparity," *IZA Journal of Labor Economics*, 2(2), 1–18.

Calvo-Armengol, A., and M. O. Jackson (2004): "The Effects of Social Networks on Employment and Inequality," *American Economic Review*, 94(3), 426–454.

Crépon, B., and R. Desplatz (2001): "Une nouvelle évaluation des effets des allégements de charges sociales sur les bas salaires," *Economie et statistique*, 348, 1–24.

Cutler, D. M., and E. L. Glaeser (1997): "Are Ghettos Good or Bad?," *The Quarterly Journal of Economics*, 112(3), 827–72.

Duguet, E., N. Leandri, Y. L'Horty, and P. Petit (2010): "Are Young French Jobseekers of Ethnic Immigrant Origin Discriminated Against? A Controlled Experiment in the Paris Area," *Annales d'Economie et de Statistique*, 99-100, 187–215.

Eddelbuettel, D., and R. François (2011): "Rcpp: Seamless R and C++ Integration," *Journal of Statistical Software*, 40(8), 1–18.

Edin, P.-A., P. Fredriksson, and O. Åslund (2003): "Ethnic Enclaves and the Economic Success of Immigrants – Evidence from a Natural Experiment," *Quarterly Journal of Economics*, 118, 329–357.

Ellwood, D. (1986): "The Spatial Mismatch Hypothesis: Are There Jobs Missing in the Ghetto?," in *The Black Youth Employment Crisis*, ed. by Freeman, and Holzer, pp. 147–185. University of Chicago Press, Chicago.

Fairlie, R. W., and W. Sundstrom (1999): "The Emergence, Persistence and Recent Widening of the Racial Unemployment Gap," *Industrial and Labor Relations Review*, 52, 252–270.

Fernandez, R., and A. Fogli (2009): "Culture: An Empirical Investigation of Beliefs, Work, and Fertility," *American Economic Journal: Macroeconomics*, 1(1), 146–177.

Fortin, N., T. Lemieux, and S. Firpo (2011): "Decomposition Methods in Economics," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, vol. 4, chap. 1, pp. 1–102. Elsevier, Amsterdam.

Givord, P., R. Rathelot, and P. Sillard (2013): "Place-based tax exemptions and displacement effects: An evaluation of the Zones Franches Urbaines program," *Regional Science and Urban Economics*, 43(1), 151–163.

Gobillon, L., P. Rupert, and E. Wasmer (2014): "Ethnic Unemployment Rates and Frictional Markets," forthcoming in *Journal of Urban Economics*.

Gobillon, L., and H. Selod (2007): "The Effect of Segregation and Spatial Mismatch on Unemployment: Evidence from France," CEPR Working Paper 6198.

Gobillon, L., H. Selod, and Y. Zenou (2007): "The Mechanisms of Spatial Mismatch," *Urban Studies*, 44(12), 2401–2427.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65(2), 261–94.

Hellerstein, J. K., and D. Neumark (2012): "Employment in Black Urban Labor Markets: Problems and Solutions," in *The Oxford Handbook of the Economics of Poverty*. Oxford University Press.

Hellerstein, J. K., D. Neumark, and M. McInerney (2008): "Spatial mismatch or racial mismatch?," *Journal of Urban Economics*, 64(2), 464–479.

Ihlanfeldt, K. R. (2006): "A Primer on Spatial Mismatch within Urban Labor Markets," in *A Companion to Urban Economics*, ed. by R. Arnott, and D. McMillen, pp. 267–276. Blackwell Publishing.

Ihlanfeldt, K. R., and D. L. Sjoquist (1998): "The Spatial Mismatch Hypothesis: A Review of Recent Studies and their Implications for Welfare Reform," *Housing Policy Debate*, 8(4), 849–892.

Ioannides, Y. M. (2011): "Neighborhood Effects And Housing," in *Handbook of Social Economics*, ed. by A. B. Jess Benhabib, and M. O. Jackson, vol. 1, pp. 1281 – 1340. North-Holland.

Kain, J. F. (1968): "Housing Segregation, Negro Employment, and Metropolitan Decentralization," *Quarterly Journal of Economics*, 82(2), 175–197.

Montgomery, J. D. (1991): "Social Networks and Labor-Market Outcomes: Toward an Economic Analysis," *American Economic Review*, 81(5), 1407–18.

Munshi, K. (2003): "Networks In The Modern Economy: Mexican Migrants In The U.S. Labor Market," *The Quarterly Journal of Economics*, 118(2), 549–599.

Neal, D. A., and W. R. Johnson (1996): "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy*, 104(5), 869–95.

Nopo, H. (2008): "Matching as a Tool to Decompose Wage Gaps," *The Review of Economics and Statistics*, 90(2), 290–299.

OECD (2012): "Untapped Skills: Realising the Potential of Immigrant Students," Discussion paper, OECD.

R Core Team (2014): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rathelot, R. (2012): "Measuring Segregation When Units are Small: A Parametric Approach," *Journal of Business & Economic Statistics*, 30(4), 546–553.

Ritter, J. A., and L. J. Taylor (2011): "Racial Disparity in Unemployment," *Review of Economics and Statistics*, 93, 30–42.

Rubin, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Sarkozy, N. (2008): "Égalité des chances et diversité," Discours de M. le Président de la République, Palaiseau, École polytechnique, http://www.archives.elysee.fr/president/les-actualites/discours/2008/egalite-des-chances-et-diversite.5239.html, accessed Dec 16, 2012.

Topa, G. (2001): "Social Interactions, Local Spillovers and Unemployment," *Review of Economic Studies*, 68(2), 261–95.

Wickham, H. (2009): *ggplot2: elegant graphics for data analysis*. Springer New York.

Zenou, Y. (2002): "How Do Firms Redline workers?," *Journal of Urban Economics*, 52(3), 391–408.

——— (2009): *Urban Labor Economics*. Cambridge University Press.

Zenou, Y., and N. Boccard (2000): "Racial Discrimination and Redlining in Cities," *Journal of Urban Economics*, 48(2), 260–285.

Table 1: Summary Statistics

Variables	Populations		
	French	North Africa	Southern Europe
Labor market outcomes			
Employment	75.7	55.0	75.8
Unemployment	6.9	18.3	7.7
Inactivity	17.4	26.8	16.4
Wage (if employed): Q1	1170	1020	1147
Wage (if employed): median	1500	1314	1485
Wage (if employed): mean	1702	1423	1627
Wage (if employed): Q3	2000	1700	1950
Cultivator	1.6	0.0	0.6
Retail, Craft	4.4	2.5	5.3
Professionals	12.8	6.1	9.9
Technicians, Sales	20.1	14.8	20.0
Office worker	24.0	25.6	26.8
Blue collar	18.7	20.4	20.0
No occupation	18.4	30.6	17.3
Education			
Medicine doctorate	1.0	0.5	0.5
Master's degree and above	3.4	2.6	2.6
Master's Elite Universities	3.4	1.5	2.0
Univ.: Bac+4, Science-Industry	0.5	0.3	0.3
Univ.: Bac+4, other	2.6	2.1	2.4
Univ.: Bac+3, Science-Industry	0.8	0.7	0.6
Univ.: Bac+3, other	3.6	3.1	3.2
Univ.: Bac+2	1.9	1.5	1.7
Tech.: Bac+2, Industry	3.0	1.8	2.6
Tech.: Bac+2, other	6.2	6.9	6.3
Health: Bac+2	2.7	1.2	1.9
Bac: Humanities	5.7	7.0	6.1
Bac: Science	3.6	3.5	2.7
Bac: Technical, Industry	1.3	0.8	1.3
Bac: Technical, other	3.6	5.0	3.7
Bac: Vocational, Industry	3.2	2.3	2.9
Bac: Vocational, other	3.1	4.9	3.5
Bac-2: Vocational, Industry	14.8	10.3	16.0
Bac-2: Vocational, other	11.3	10.9	13.6
Lower Sec. Educ. Deg.	7.7	9.5	8.0
No diploma	16.6	23.7	18.0
Age			
Age, years: Q1	29.0	24.0	29.0
Age, years: median	40.0	30.0	39.0
Age, years: mean	39.5	31.6	38.8
Age, years: Q3	50.0	37.0	48.0
Demography and family			
Female	51.1	52.8	51.1
Couple	74.1	68.3	75.5
Working spouse	52.3	29.9	51.5
No child	58.3	49.5	55.8
1 child	19.5	23.0	21.5
2 children	16.1	17.4	17.5
3+ children	6.1	10.2	5.2
Youngest child less than 3	9.6	16.8	10.1
Nobs	220,802	8432	11,653

Source: Labor Force Survey 2005-2011 (Insee).
Note: All figures are proportions, expressed in percentage, except from the monthly wage (in euros) and the age (in years).
Reading note: 75.7% of French individuals with French parents are in employment. 55.0% of French individuals who have at least one parent born with the nationality of a North African country are in employment.

Table 2: Summary Statistics, for clusters

Variables	Clusters			
	All		Only mixed: majority with...	
			North Africa	Southern Europe
Proportion of minority group				
Proportion (North Afr): Q1	0	0	0.024	0
Proportion (North Afr): median	0	0.0099	0.05	0.013
Proportion (North Afr): mean	0.041	0.033	0.088	0.042
Proportion (North Afr): Q3	0.045	0.037	0.11	0.052
Proportion (Southern Eur): Q1	0	0.0089	0.013	0.032
Proportion (Southern Eur): median	0.032	0.032	0.043	0.058
Proportion (Southern Eur): mean	0.05	0.046	0.057	0.075
Proportion (Southern Eur): Q3	0.075	0.068	0.083	0.097
Employment rate				
Employment rate: Q1	0.67	0.69	0.64	0.68
Employment rate: median	0.76	0.76	0.74	0.76
Employment rate: mean	0.74	0.74	0.71	0.74
Employment rate: Q3	0.83	0.82	0.8	0.82
Number of individuals				
Nb ind: Q1	19	38	25	25
Nb ind: median	36	62	46	45
Nb ind: mean	44	65	51	51
Nb ind: Q3	62	88	71	70
Nb ind (majority): Q1	16	32	19	20
Nb ind (majority): median	31	54	37	38
Nb ind (majority): mean	38	58	42	43
Nb ind (majority): Q3	54	80	59	60
Nb ind (North Afr): Q1	0	0	1	0
Nb ind (North Afr): median	0	1	2	1
Nb ind (North Afr): mean	1.5	1.8	3.1	1.8
Nb ind (North Afr): Q3	2	2	4	2
Nb ind (Southern Eur): Q1	0	1	1	1
Nb ind (Southern Eur): median	1	2	2	2
Nb ind (Southern Eur): mean	2	2.8	2.8	3.1
Nb ind (Southern Eur): Q3	3	4	4	4
Number of clusters	5742	5742	2689	3812
Weights	No	Cluster size	No	No

Source: Labor Force Survey 2005-2011 (Insee).
Note: In columns 1, 3 and 4, statistics are unweighted. In column 2, statistics are weighted by the size of the clusters.
Reading note: The median cluster has an employment rate of 76%. Among the clusters mixing the majority population and the minority population with North African ancestry, the median employment rate is equal to 74%. From an individual point of view, the median size of a cluster is 62.

Table 3: Estimation results of the employment equation

	Cond. logit
Female	-1.40^{***} (0.25)
Experience	
Age (/10)	2.13^{***} (0.15)
Age (/10) squared	-2.52^{***} (0.20)
Education	
Bac: Humanities	Ref.
Medicine doctorate	0.47^{***} (0.24)
Master degree and above	0.40^{***} (0.05)
Elite university	0.45^{***} (0.05)
Univ.: Bac+4, Science-Industry	0.23^{***} (0.06)
Univ.: Bac+4, other	0.19^{***} (0.12)
Univ.: Bac+3, Science-Industry	0.13^{**} (0.06)
Univ.: Bac+3, other	0.13 (0.11)
Univ.: Bac+2	0.00 (0.07)
Tech.: Bac+2, Industry	0.37^{***} (0.05)
Tech.: Bac+2, other	0.40^{***} (0.05)
Health: Bac+2	0.48^{***} (0.09)
Bac: Science	-0.50^{***} (0.18)
Bac: Technical, Industry	0.09 (0.08)
Bac: Technical, other	0.09 (0.07)
Bac: Vocational, Industry	0.43^{***} (0.05)
Bac: Vocational, other	0.16^{**} (0.08)
Bac-2: Vocational, Industry	0.59^{**} (0.20)
Bac-2: Vocational, other	-0.01 (0.02)
Lower Sec. Educ. Deg.	-0.15 (0.10)
No diploma	-1.21^{***} (0.44)
Dummies for quarters	Yes
Nobs	237,039

Source: Labor Force Survey 2005-2011 (Insee).
Note: Conditional logit model with fixed effects for the clusters. Standard errors are given between parentheses. The estimation concerns the sample of majority individuals only.

Table 4: Actual and counterfactual employment probabilities: estimation results

	Specifications			
Kernel Bandwidth	Uniform .05	Epanechnikov .05	Uniform .10	Epanechnikov .10
$\mathbb{E}[Y(0)\|g=0]$	0.757 [0.755;0.760]	0.757 [0.755;0.760]	0.757 [0.754;0.761]	0.757 [0.754;0.760]
$\mathbb{E}[Y(0)\|g=0, a \in \mathcal{M}]$	0.743 [0.739;0.747]	0.743 [0.738;0.748]	0.743 [0.738;0.747]	0.743 [0.737;0.749]
$\widehat{\overline{Y}}$	0.715 [0.691;0.720]	0.714 [0.692;0.720]	0.718 [0.708;0.723]	0.717 [0.706;0.721]
$\widehat{\underline{Y}}$	0.697 [0.667;0.706]	0.695 [0.667;0.708]	0.688 [0.671;0.699]	0.688 [0.672;0.696]
$\widehat{\underline{Y}}^*$	0.683 [0.664;0.697]	0.681 [0.664;0.694]	0.685 [0.670;0.699]	0.685 [0.671;0.694]
$\mathbb{E}[Y(1)\|g=1, a \in \mathcal{M}]$	0.550 [0.537;0.565]	0.550 [0.535;0.563]	0.550 [0.536;0.561]	0.550 [0.539;0.562]
$\mathbb{E}[Y(1)\|g=1]$	0.550 [0.537;0.565]	0.550 [0.536;0.564]	0.550 [0.536;0.561]	0.550 [0.539;0.562]
p_s	0.101 [0.045;0.111]	0.101 [0.042;0.106]	0.043 [0.016;0.047]	0.043 [0.016;0.049]
δ_y	0.018 [0.012;0.021]	0.018 [0.011;0.019]	0.009 [0.005;0.010]	0.009 [0.005;0.011]
δ_s	0.019 [0.007;0.020]	0.019 [0.006;0.020]	0.009 [0.003;0.011]	0.009 [0.003;0.011]
p_s^*	0.040 [0.031;0.044]	0.040 [0.030;0.045]	0.027 [0.015;0.030]	0.027 [0.015;0.031]
δ_y^*	0.009 [0.007;0.014]	0.009 [0.008;0.015]	0.007 [0.004;0.009]	0.007 [0.005;0.009]
δ_s^*	0.013 [0.006;0.015]	0.014 [0.006;0.015]	0.009 [0.003;0.010]	0.009 [0.003;0.011]
Nobs minority	8432	8432	8432	8432

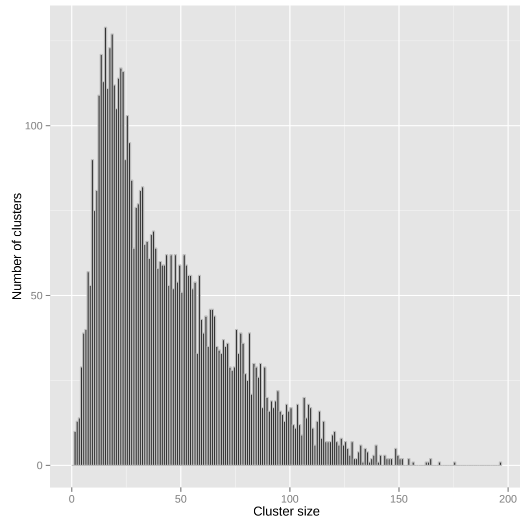
Source: Labor Force Survey 2005-2011 (Insee).
Note: In all specifications, the score is estimated by a conditional logit model (see Table 3). In this table, the estimates of the quantities introduced in Section 4 are reported for differents settings of the matching procedure. In columns 1 and 3, the matching kernel is a uniform, while in columns 2 and 4, an Epanechnikov kernel is used. In columns 1 and 2, the bandwidth is set to .05. In columns 3 and 4, the bandwidth is set to .10. The figures between square brackets are 95% confidence intervals, obtained by full bootstrap (with 100 iterations).

Table 5: Actual and counterfactual employment probabilities: other populations

Variables	Populations				
	Benchmark Population	Women only	Southern Europe	2 parents immigrants	Only father immigrant
$\mathbb{E}[Y(0)\|g=0]$	0.757 [0.754;0.760]	0.711 [0.707;0.717]	0.757 [0.753;0.759]	0.757 [0.755;0.760]	0.757 [0.755;0.760]
$\mathbb{E}[Y(0)\|g=0,a\in\mathcal{M}]$	0.743 [0.737;0.749]	0.693 [0.685;0.700]	0.759 [0.754;0.762]	0.736 [0.730;0.741]	0.738 [0.731;0.745]
$\widehat{\overline{Y}}$	0.718 [0.706;0.721]	0.664 [0.659;0.673]	0.759 [0.756;0.766]	0.719 [0.710;0.730]	0.713 [0.706;0.722]
$\widehat{\underline{Y}}$	0.688 [0.672;0.696]	0.650 [0.640;0.669]	0.763 [0.760;0.770]	0.685 [0.662;0.694]	0.696 [0.683;0.705]
$\widehat{\underline{Y}}^*$	0.685 [0.671;0.694]	0.646 [0.635;0.663]	0.762 [0.759;0.770]	0.682 [0.661;0.694]	0.694 [0.683;0.703]
$\mathbb{E}[Y(1)\|g=1,a\in\mathcal{M}]$	0.550 [0.539;0.562]	0.505 [0.490;0.521]	0.758 [0.749;0.766]	0.538 [0.523;0.553]	0.594 [0.574;0.612]
$\mathbb{E}[Y(1)\|g=1]$	0.550 [0.539;0.562]	0.504 [0.488;0.520]	0.758 [0.749;0.766]	0.538 [0.523;0.553]	0.594 [0.573;0.611]
p_s	0.043 [0.016;0.049]	0.073 [0.062;0.081]	0.015 [0.004;0.018]	0.015 [0.011;0.025]	0.028 [0.010;0.032]
δ_y	0.009 [0.005;0.011]	0.016 [0.012;0.022]	0.003 [0.001;0.004]	0.004 [0.003;0.008]	0.007 [0.003;0.010]
δ_s	0.009 [0.003;0.011]	0.015 [0.012;0.017]	0.005 [0.001;0.006]	0.002 [0.002;0.005]	0.007 [0.002;0.009]
p_s^*	0.027 [0.015;0.031]	0.053 [0.045;0.059]	0.011 [0.004;0.014]	0.015 [0.011;0.024]	0.020 [0.010;0.025]
δ_y^*	0.007 [0.005;0.009]	0.013 [0.009;0.017]	0.003 [0.001;0.003]	0.004 [0.003;0.008]	0.006 [0.002;0.009]
δ_s^*	0.009 [0.003;0.011]	0.015 [0.012;0.017]	0.004 [0.001;0.006]	0.003 [0.002;0.004]	0.007 [0.002;0.008]
Nobs minority	8432	4448	11,653	5801	1946

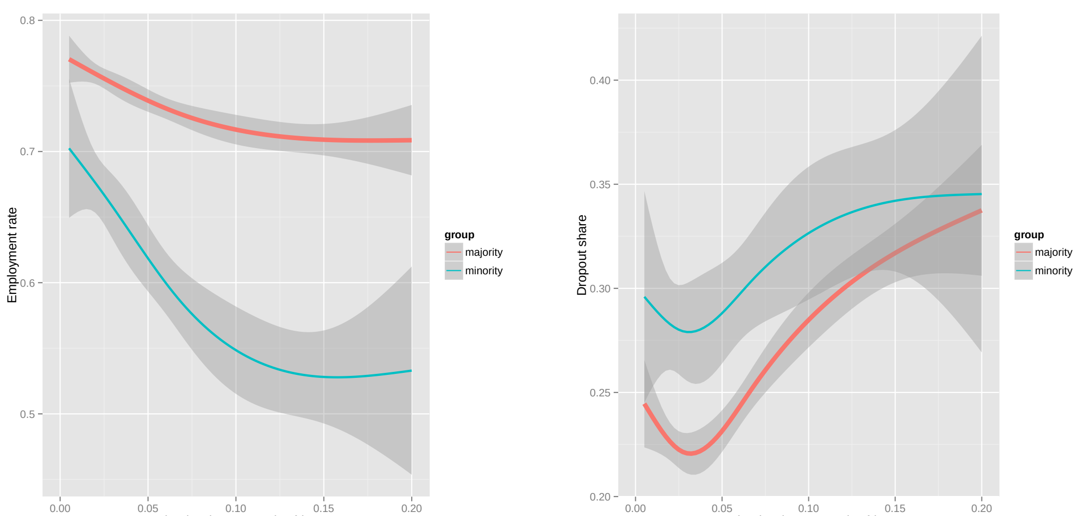
Source: Labor Force Survey 2005-2011 (Insee).
Note: In all specifications, the score is estimated by a conditional logit model (see Table 3). This table reports the estimates of the quantities introduced in Section 4 for several populations. In all cases, the matching is performed with an Epanechnikov kernel and a bandwidth set to .10.
The figures between square brackets are 95% confidence intervals, obtained by full bootstrap (with 100 iterations).

Figure 1: Size distribution of clusters



Source: Labor Force Survey 2005-2011 (Insee).
Reading note: There are 10 clusters with exactly 1 observation.

Figure 2: Employment rate and share of dropouts as function of the share of the minority group in the cluster



Source: Labor Force Survey 2005-2011 (Insee).
Note: The x-axis is the share in the cluster of French individuals with at least one parent from North Africa. Shades indicate 95% confidence interval. Kernel regressions have been estimated using the `geom_smooth` function in the `ggplot2` package of `R`; see R Core Team (2014) and Wickham (2009).
Reading note: The employment rate in the majority population in clusters in which 10% of the population are French individuals with North African parents is equal to 71%. In the same areas, the employment of the French individuals with at least one North African parent is equal to 55%.

Appendix.

A. 1. Distribution of employment rate and share of minority population across clusters

Figure 3 displays, in bold, the kernel density estimate of the distribution of the share per unit of French individuals with North African parents (left panel) and of the employment rate per unit (right panel).

[Insert here Figure 3]

Because clusters are of relatively small size, one would not expect these distributions to look like a Dirac, even in the case in which, in every unit, everyone had the same probability to belong to the minority group (or to be employed). The estimate should be compared to the simulated distribution – the thin line – obtained under the hypothesis that each cluster has the same intrinsic probability and drawing outcomes in a binomial distribution. Both panels of figure 3 show that the observed distributions are unambiguously more dispersed than the simulated ones. In the left panel, the number of areas in which there is no individual from the minority is almost twice as high as it would be in the homogeneity case. The upper tail of the observed distribution is also notably thicker than the one of the simulated distribution, which suggests that minority individuals tend to be concentrated in some areas. More evidence on the level of ethnic segregation in France, measured with the same data, is provided by Rathelot (2012). In the right panel, the observed distribution of the employment rate by cluster also displays more dispersion than the simulated one. Interestingly, the lower tail of the bold line, from .65 downwards, is higher than the thin line by many orders of magnitude. This stresses the existence of distressed areas, in which the employment rate is especially low, and corresponds to social segregation.

A.2. A linear model of sorting and employment

This subsection illustrates in a simple linear case that sorting on observables and unobservables do not need to go in the same direction. Simple expressions for the set estimates introduced in Section 4 are also provided.

Asymmetric sorting and within-cluster ethnic differentials in observables and unobservables

Thoughout this section, we relax the assumption that there are only two locations. Space is assumed to be continuous and a location is characterized by its quality q. The rest of the framework is simplified and linear relationships are assumed between variables. g denotes the individual ethnic group and is equal to one if the individual belongs to the minority group, zero otherwise. The observable characteristics that are relevant to employment are assumed to be summarized by a real-valued function $s(.)$, defined such that the probability of being employed is increasing in $x = s(X)$. x is assumed to be lower on average in the minority group. The quality of the neighborhood chosen by an individual increases in her characteristics x and u and, based on the prediction from the previous subsection, decreases in g. Finally, the latent variable y^ℓ determining the employment status is assumed to be increasing in x, u and q and decreasing in g. We combine these assumptions to form the following system:

$$x = x^* - \delta g \tag{4}$$

$$q = \beta(x + u) - \gamma g + u_q \tag{5}$$

$$y^\ell = \eta x + \mu u + \alpha q - \lambda g + u_e \tag{6}$$

We consider the data-generating process defined by equations (4)-(6), together with the assumptions that x^*, g, u, u_q, u_e are random variables that are independent from each other. Parameters α, β, γ, δ, λ, μ and η are assumed to be all positive. x^* represents the observable characteristics that the individual would have if he belonged to the majority group. In this world, the distribution of the observables x is identical in both groups up to a translation by δ. Given her characteristics x and u, an individual will live in a place of quality q, lower for the minority group. Finally, given x, u, g and q, individuals draw a latent for employment.

In this case, by assumption, individuals from the majority live in better neighborhoods than individuals from the minority. Conditional on x, the gap is equal to γ and, unconditional on x, the gap is even larger, equal to $\gamma + \beta\delta$, as individuals from the minority have on average lower characteristics x. Empirically testing this prediction is not straightforward, as there is no perfect measure of neighborhood quality. In order to provide tentative evidence, we estimate $s(X)$ at the individual level (and denote $\hat{s}(.)$ the estimate) and consider a measure of quality $\hat{q}_i$ the share of individuals j in the cluster where i lives who have $\hat{s}(X_j)$ higher than $\hat{s}^*_{90}$, which is the 90th percentile of the estimated $\hat{s}(X)$ in the whole population.[15] We regress $\hat{q}_i$ on $\hat{s}(X_i)$ and the ethnicity dummy g_i (in this case, having parents from North Africa). We restrict the sample to individuals with $\hat{s}(X_i) < s^*_{90}$ to avoid a direct correlation between the two variables. In this case, the estimated value for γ is positive and strongly significant, with a t-statistic higher than 20, even when errors are clustered to account for within-cluster correlation. With the dgp defined with (4)-(6), γ is consistently estimated by OLS.

Now, how should characteristics x and u vary with the ethnic group, conditional on the

[15] The details of the estimation procedure of $\hat{s}(X)$ are delayed to section 4.2.

neighborhood quality q? From equation (5), if $\beta \neq 0$:

$$u = \frac{1}{\beta}q + \frac{\gamma}{\beta}g - x - \frac{1}{\beta}u_q$$
$$x = \frac{1}{\beta}q + \frac{\gamma}{\beta}g - u - \frac{1}{\beta}u_q$$

Regressing u on g, controlling for q and x, will not yield γ/β because of the correlation between q and u_q. OLS regression would provide an attenuated estimator, but with the same sign as γ/β.

$$\mathbb{E}[u|x,q,g=1] - \mathbb{E}[u|x,q,g=0] = \frac{\gamma}{\beta}\frac{1}{1+\frac{\mathbb{V}(u_q)/\beta^2}{\mathbb{V}(u)}}$$

Regressing x on g, controlling for q will also yield a inconsistent estimator, which may not have the same sign as γ/β.

$$\mathbb{E}[x|q,g=1] - \mathbb{E}[x|q,g=0] = \frac{\gamma}{\beta}\frac{1-\frac{\delta\beta}{\gamma}\frac{\mathbb{V}(u_q)/\beta^2+\mathbb{V}(u)}{\mathbb{V}(x^*)}}{1+\frac{\mathbb{V}(u_q)/\beta^2+\mathbb{V}(u)}{\mathbb{V}(x^*)}}$$

Even when γ/β is positive and when neighborhood quality is controlled for, ethnic minorities may have on average lower observed characteristics than the majority. To be more precision, a sufficient and necessary condition for this issue to arise is:

$$\delta > \frac{\gamma}{\beta}\frac{\mathbb{V}(x^*)}{\mathbb{V}(u)+\mathbb{V}(u_q)/\beta^2}$$

that is, when the difference δ in the expectation of x across groups and the effect β of characteristics on neighborhood quality are large relative to the sorting asymmetry γ or, when the variance of the errors u_q or the unobservables u is large compared to the within-group variance of the observables x^*.

[Insert here Figure 4]

Figure 4 illustrates the issue qualitatively. Data are simulated following the dgp defined by equations (4)-(6). x^*, g, u, u_q are random variables distributed as:

$$\begin{aligned} &x^* \sim N(0,\sigma^2_{x^*}) && u \sim N(0,\sigma^2_u) \\ &u_q \sim N(0,\sigma^2_{u_q}) && g \sim Bernouilli(.5) \\ &\beta = .6; \gamma = .8; \delta = 1.7 && \sigma_{x^*} = 2; \sigma^2_u = 2.5; \sigma^2_{u_q} = 3. \end{aligned}$$

Characteristics (x in the left panel, u in the right panel) are on the x-axis and neighborhood quality q is on the y-axis. The figures show scatterplots of the observations, for both groups (majority group in red points and minority group in green triangles) as well as regression lines for q as a function of x-axis variable (plain lines) and the x-axis variable as a function of q (dotted lines). Green thin lines correspond to the minority group, red thick ones to the majority group. In the left panel, it is clear that the reverse regression may lead to a change of the sign of the ethnic difference in characteristics within neighborhoods. In the right panel, the difference in unobservables is attenuated by the reverse regression but the sign remains the same. If u_q and u were very small, however, the points would be all very close to the plain lines, and the dotted lines would coincide with the plain lines. In our sample, regressing the estimated $\hat{s}(X)$ on g, controlling for the share $\hat{q}$ of individuals living in the cluster with estimated $\hat{s}(X)$ higher than s^*_{90} yields a significantly negative estimate.[16] Conditional on neighborhood quality, individuals from the minority still have worse characteristics than those from the majority. This empirical finding, which may seem paradoxical in a deterministic framework, is well accounted for by our small stochastic linear model.

Set identification in the simple linear model

Now, consider that the econometrician wants to use equation (6) to learn about λ, which is the primary quantity of interest in this paper. Suppose, for simplicity, that she observes y^ℓ the outcome, x the observable characteristics, q the neighborhood quality and g the group. Regressing y^ℓ on x, q, and g will provide an estimate that controls for the existence of some sorting. In this case, the OLS estimator of λ converges to:

$$\begin{aligned}\lambda_{xq} &= \mathbb{E}(y^\ell|x,q,g=1) - \mathbb{E}(y^\ell|x,q,g=0) \\ &= \lambda + \mu\frac{\gamma}{\beta}\frac{1}{1+\frac{\mathbb{V}(u_q)/\beta^2}{\mathbb{V}(u)}}\end{aligned}$$

The econometrician may also forget about neighborhood quality or may have no information about residential location, as in the vast majority of empirical studies about ethnic gaps. She would regress y^ℓ on x and g only. In this case, the OLS estimator converges to:

$$\begin{aligned}\lambda_x &= \mathbb{E}(y^\ell|x,g=1) - \mathbb{E}(y^\ell|x,g=0) \\ &= \lambda - \alpha\gamma\end{aligned}$$

Under the assumptions made in subsection 3.2 and 3.3, λ_x and λ_{xq} are bounds for the true quantity λ: $\lambda_x < \lambda < \lambda_{xq}$. Note that there are two interesting special cases. If there is no causal effect of the neighborhood on employment, $\alpha = 0$, $\lambda_x = \lambda$ and q should be omitted from the regression. If sorting is symmetric, $\gamma = 0$, $\lambda_{xq} = \lambda_x = \lambda$, the estimator is consistent whether or not controls for neighborhood quality are included.

A.3. Decomposition under two simplifying assumptions

Assumption 1-3 Assumptions 1 and 3 together imply that:

$$\forall(X,u), \mathbb{E}(Y(0)|X,u,g=0) = \mathbb{E}(Y(0)|X,u,g=1) \tag{7}$$

[16] This result still holds when cluster fixed effects are introduced to account for neighborhood quality instead of $\hat{q}$.

Then, by applying Bayes' law,

$$\mathbb{E}(Y(0)|g=1) = \int \mathbb{E}(Y(0)|g=1,X,u)\phi_{u|X,g=1}(u|X,g=1)d\Phi_{X|g=1}(X|g=1)$$

Using equation (7),

$$\mathbb{E}(Y(0)|g=1) = \int \mathbb{E}(Y(0)|g=0,X,u)\phi_{u|X,g=1}(u|X,g=1)d\Phi_{X|g=1}(X|g=1)$$

Using assumption 2,

$$\mathbb{E}(Y(0)|g=1) = \int \mathbb{E}(Y(0)|g=0,X,u)\phi_{u|X,g=0}(u|X,g=0)d\Phi_{X|g=1}(X|g=1)$$

So that,

$$\mathbb{E}(Y(0)|g=1) = \int \mathbb{E}(Y(0)|g=0,X)d\Phi_{X|g=1}(X|g=1) = \overline{Y}$$

Assumption 1-2 and 4 Start by applying Bayes' law,

$$\phi_{u|X,a,g}(u|X,a,g) = \frac{\mathbb{P}[a|X,u,g]\phi_{u|X,g}(u|X,g)}{\mathbb{P}[a|X,g]}$$

or, developping the denominator,

$$\phi_{u|X,a,g}(u|X,a,g) = \frac{\mathbb{P}[a|X,u,g]\phi_{u|X,g}(u|X,g)}{\int \mathbb{P}[a|X,u,g]d\Phi_{u|X,g}(u|X,g)}$$

Thus, using assumptions 2 and 4,

$$\phi_{u|X,a,g=0}(u|X,a,g=0) = \phi_{u|X,a,g=1}(u|X,a,g=1) \tag{8}$$

Now, applying Bayes' law

$$\mathbb{E}(Y(0)|g=1) = \int \mathbb{E}(Y(0)|g=1,X,u,a)\phi_{u|X,a,g=1}(u|X,a,g=1)d\Phi_{X,a|g=1}(X,a|g=1)$$

Using equation (8) and assumption 1,

$$\mathbb{E}(Y(0)|g=1) = \int \mathbb{E}(Y(0)|g=0,X,u,a)\phi_{u|X,a,g=0}(u|X,a,g=0)d\Phi_{X,a|g=1}(X,a|g=1)$$

So that

$$\mathbb{E}(Y(0)|g=1) = \int \mathbb{E}(Y(0)|g=0,X,a)d\Phi_{X,a|g=1}(X,a|g=1) = \underline{Y}$$

which provides one equality. We also have:

$$\mathbb{E}(Y(0)|g=1,X,u) = \int \mathbb{E}(Y(0)|g=1,X,u,a)\mathbb{P}[a|X,u,g=1]$$

Using assumptions 1 and 4, we obtain equation (7). As proved in the previous paragraph, equation (7) together with assumption 2 leads to:

$$\mathbb{E}(Y(0)|g=1) = \overline{Y}$$

Therefore, under assumptions 1, 2 and 4,

$$\mathbb{E}(Y(0)|g=1) = \underline{Y} = \overline{Y}$$

A.4. Proof of Proposition 1

Using Bayes' law,

$$\mathbb{E}(Y(0)|g=1) = \int \mathbb{E}(Y(0)|g=1, X, u)\phi_{u|X,g=1}(u|X, g=1)d\Phi_{X|g=1}(X|g=1)$$

Using assumption 5,

$$\mathbb{E}(Y(0)|g=1) \leq \int \mathbb{E}(Y(0)|g=0, X, u)\phi_{u|X,g=1}(u|X, g=1)d\Phi_{X|g=1}(X|g=1)$$

Using assumption 2,

$$\mathbb{E}(Y(0)|g=1) \leq \int \mathbb{E}(Y(0)|g=0, X, u)\phi_{u|X,g=0}(u|X, g=0)d\Phi_{X|g=1}(X|g=1)$$

So that,

$$\mathbb{E}(Y(0)|g=1) \leq \int \mathbb{E}(Y(0)|g=0, X)d\Phi_{X|g=1}(X|g=1)$$

which shows the first inequality.

Applying Bayes' law,

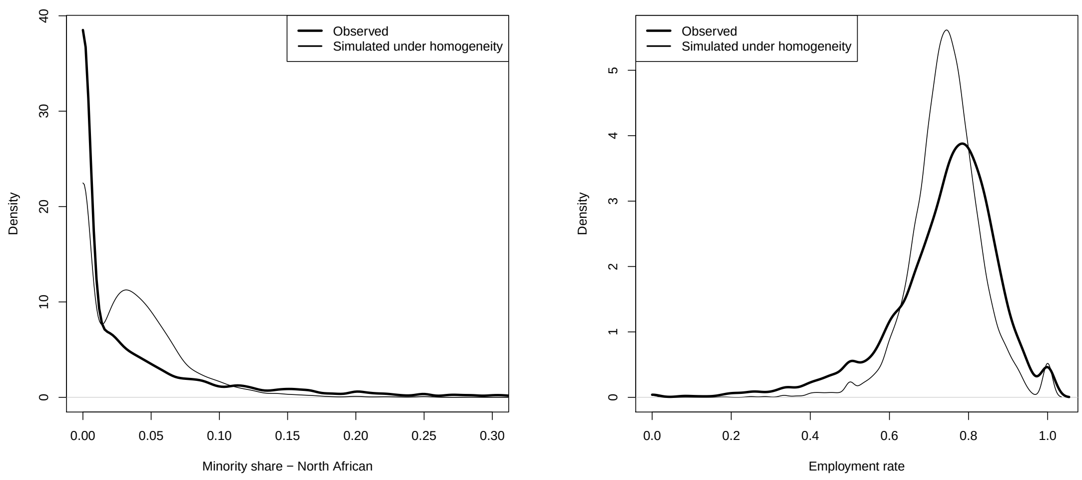
$$\mathbb{E}(Y(0)|g=1) = \int \mathbb{E}(Y(0)|g=1, X, a)d\Phi_{X,a|g=1}(X, a|g=1)$$

Using assumption 5,

$$\mathbb{E}(Y(0)|g=1) \geq \int \mathbb{E}(Y(0)|g=0, X, a)d\Phi_{X,a|g=1}(X, a|g=1)$$

which shows the second inequality.

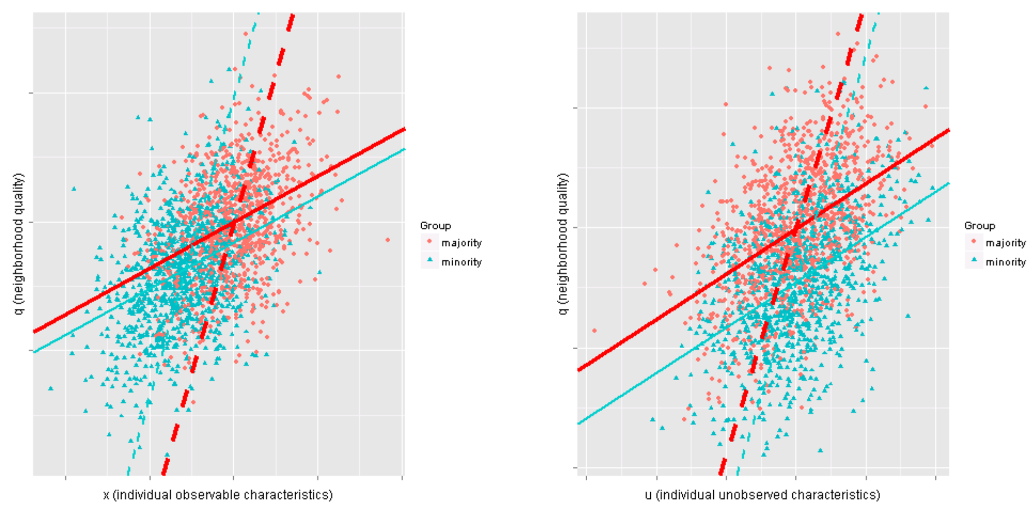
Figure 3: Distributions of the share of the minorities (left panel) and of the employment rate (right panel) by cluster: observed and simulated under an homogeneity hypothesis



Source: Labor Force Survey 2005-2011 (Insee).

Note: The left panel displays the distribution of French individuals with at least one parent of North African origin; the right panel displays the distribution of the employment rate across units.

Figure 4: Simulations from the linear model: neighborhood quality, observable and unobservable characteristics



Source: simulations.

Note: In the scatterplots, each point corresponds to an individual. Plain lines correspond to regression lines of the variable on the y-axis (neighborhood quality q) on the variable on the x-axis (x or u). Dotted lines correspond to reverse regressions: variable on the x-axis on variable on the y-axis.