UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TROIS ESSAIS SUR LA SÉGRÉGATION

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN ÉCONOMIQUE

PAR

OUALID MOUSSOUNI

AVRIL 2020

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

THREE ESSAYS ON SEGREGATION

THESIS

PRESENTED

AS PARTIAL REQUIREMENT

OF DOCTORAL OF PHILOSOPHY IN ECONOMICS

BY

OUALID MOUSSOUNI

APRIL 2020

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TROIS ESSAIS SUR LA SÉGRÉGATION

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN ÉCONOMIQUE

PAR

OUALID MOUSSOUNI

AVRIL 2020

REMERCIEMENTS

Je remercie mon directeur de thèse, qui est pour moi presque un directeur parfait (presque). J'ai appris, j'apprends et j'apprendrais toujours de lui, MERCI Kris, merci pour le temps que tu as consacré pour moi. Je veux bien remercier d'autres personnes, mais je ne sais même pas qui va lire cette section. Je vais sûrement prendre soin de les remercier en personne, leur dire combien ils m'ont aidé, et on ira boire un coup, pour ceux qui boivent. [1] Sinon, je tiens à me remercier beaucoup, it was not easy on me.

1. Idéalement, ils vont payer l'addition pour me féliciter de cet accomplissement.

TABLE DES MATIÈRES

LIST OF FIGURES

LIST OF TABLES

RÉSUMÉ

Cette thèse s'interroge sur les questions de ségrégation et continue une longue tradition en science économique (particulièrement en économie géographique et urbaine), mais aussi en sociologie. Cette tradition de l'analyse comporte généralement trois volets : (1) mesures de ségrégation, (2) analyses des causes de la ségrégation et (3) de ses conséquences. Cette thèse est composée de trois chapitres et explore ces trois volets.

Le premier chapitre, intitulé « *Distance-based segregation measures* », couvre le volet concernant les mesures de la ségrégation. Nous adaptons des mesures de concentration des firmes afin de quantifier la ségrégation de certaines populations. Ce chapitre propose une mesure innovante de la ségrégation et répond aux critiques sur les mesures indicielles souvent utilisées dans la littérature. Dans un cadre unifié, nous combinons les dimensions intra-groupe et inter-groupes, et évaluons les distributions spatiales par race et revenus en utilisant des micro-données géographiques. Nos mesures sont insensibles aux découpages géographiques, permettent de faire des comparaisons spatiales et temporelles, et de simuler des contrefactuelles afin de tester la significativité statistique des distributions observées. Qui plus est, cette méthode permet d'isoler de manière distincte l'effet de la race et celui de revenu sur la ségrégation. Les résultats de ce chapitre présentent un portrait original sur la métropole de New-York. Il montre comment la race pousse les individus à se concentrer spatialement avec leur propre groupe et à se distancer des autres, et comment la pauvreté accentue encore plus ces effets.

Le deuxième chapitre, intitulé « *What matters for choosing your neighbors ? Evidence from Canadian metropolitan areas* », couvre le volet concernant les causes de la ségrégation. Afin d'analyser les déterminants de choix de localisation des individus, nous explorons la dimension inter-groupe de la ségrégation en calculant des mesures de co-agglomération de différentes ethnies dans les six grandes métropoles Canadiennes. Ce chapitre présente un portrait multidimensionnel de la ségrégation urbaine et permet de comprendre comment les individus choisissent leurs voisins. Notamment, ces choix sont influencés par différentes dimensions linguistiques, culturelles, religieuses, politiques et génétiques. Nous montrons comment ces dimensions de similarité sont positivement associées avec la co-localisation des groupes ethniques. En d'autres termes, nos résultats révèlent que des mesures de proximité (linguistiques, culturelles, religieuses, etc.) s'avèrent dans la détermination du choix de localisation des individus d'origines différentes.

Le troisième chapitre, intitulé « *Race and firm location : who moves where ?* », couvre le volet concernant les conséquences de la ségrégation. Ce chapitre présente une analyse approfondie de la distance entre les individus et les emplois en explorant l'hétérogénéité de la race et des revenus.

L'idée est que certaines minorités ethniques sont géographiquement loin de leur lieu d'emploi et que cela pourrait avoir des répercussions sur leur situation sur le marché du travail. Ce phénomène, communément appelé « spatial mismatch hypothesis », est donc vu comme une cause possible du chômage et des revenus faibles de certains groupes comme les Afro-Américains, les Hispaniques et autres groupes minoritaires. Ce chapitre teste cette hypothèse en fournissant une nouvelle mesure de co-agglomération entre des firmes de certaines industries et des individus de certains groupes. Il montre l'existence de certaines régularités sur l'effet de la race et de la pauvreté sur cette déconnexion géographique ainsi que sur la décentralisation. Nos résultats préliminaires suggèrent qu'entre 1990 et 2010, les Blancs sont plus proches, tandis que les Noires, Hispaniques et Asiatiques sont plus loin des emplois. De plus, la pauvreté accentue encore plus cet éloignement. Enfin, en comparant le degré de déconnexion physique, les Blancs et Asiatiques semblent être plus proches de leurs employeurs potentiels que les Noirs et Hispaniques.

Mots-clé : Ségrégation, race et ethnie, homophilie, déconnexion physique, décentralisation.

ABSTRACT

My research focuses on understanding the location of people and firms. More precisely, I study three different themes. First, I provide new measures of the spatial concentration of individuals (i.e., segregation) that overcome specific issues with existing measures. Second, I use these new measures to look at the potential causes that lead individuals of a given ethnic origin to co-locate with other ethnies. Third, combining both previous points, I explore the joint spatial distribution of individuals and firms to provide new insights on the spatial mismatch hypothesis.

In the first chapter, "*Distance-based segregation measures*", we apply point-pattern based measures of geographic concentration—usually used to assess the extent and statistical significance of the spatial clustering of firms or plants—to the measurement of segregation. Our measures of 'excess segregation' satisfy a number of desirable properties and allow to assess the geographic distribution of groups using spatially fine-grained data. They allow for statistical testing of the observed patterns against various reference distributions and can be applied to the measurement of segregation (within groups) and isolation (between groups). We use them to also partly disentangle segregation by race from the geographic concentration of poverty.

In the second chapter, "*What matters for choosing your neighbors? Evidence from Canadian metropolitan areas*", we adapt the previous measures and combine two fields of science, i.e., geography and psychology, and answer an urban economics question. In geography, The First Law of Geography states that "everything is related to everything else, but near things are more related than distant things." In psychology, The Principle of Homophily posits that "similarity breeds connection." Thus, a corollary of this two starting points could be that "near things are more similar than distant things." We test that proposition using spatially fine-grained data on thousands of colocation patterns of ethnic groups in the six largest Canadian metropolitan areas. The geographic patterns reveal that groups that are more similar along various non-spatial dimensions— language, culture, religion, genetics, and historico-political relationships—colocate more. These results are robust to numerous controls and provide a quantitative glimpse of the 'deep roots' of homophily.

In the third chapter, "*Race and firm location: who moves where?*", we provide a measure for, and empirical application of, the spatial mismatch hypothesis in

order to better understand patterns of firm and individual location choice. We first explore how the dynamic of decentralization in New York area affected different groups in a different magnitude between 1990 and 2010. We then test the spatial physical disconnection between individuals and their potential employers. The article shows a robust empirical regularities of the effect of race and poverty on the spatial mismatch. For instance, we find that White shifted towards jobs while Black, Hispanic and Asian shifted away from jobs. Within each group, the shift between jobs and poor individuals is even more pronounced.

Keywords: Segregation, race and ethnicity, homophily, physical disconnection, decentralization.

INTRODUCTION

En 2014, Piketty et Goldhammer montre que la rémunération du capital étant historiquement supérieure à la croissance économique serait à l'origine des inégalités dans les pays développés. Ce phénomène est toujours au cœur des débats académiques et politiques actuels. Dans cette lignée, les économistes urbains se sont aussi intéressés à la distribution spatiale des revenus au sein du même pays. De manière générale, ils trouvent que les villes les plus densément peuplées ont tendance à être plus inégales que les villes moins denses.

Dans cette thèse, nous nous intéressons à des phénomènes de concentration spatiale qui s'opèrent dans ces villes denses, et regardons comment la race et la pauvreté affectent ces distributions spatiales. Même si cette thèse ne fournit pas de lien causal entre ces phénomènes et les inégalités de revenus, il y a une littérature abondante qui montre que les groupes raciaux ségrégés, pauvres et peu éduqués ont tendance à cumuler des difficultés d'insertion sur le marché de travail, et donc de leur niveau de revenu. Cependant, étudier les conséquences de la concentration spatiale des individus, ainsi que les causes, requiert d'abord une mesure adéquate pour mieux capter et saisir son amplitude. C'est pourquoi cette thèse s'est penchée, dans un premier temps, à développer de nouvelles mesures qui surmontent la faiblesse des mesures existantes.

En effet, dans le premier chapitre, nous nous concentrons sur l'évaluation de la ségrégation en captant ses différents aspects et dimensions. Nous développons des nouveaux outils pour montrer l'étendu de la ségrégation et comment la race et les revenues façonnent ce phénomène. On définit la ségrégation comme un phénomène

de concentration spatiale d'individus ayant des caractéristiques communes. Dans cette thèse, on s'intéresse à la race et au revenu comme caractéristiques poussant les groupes à se concentrer spatialement. Par ailleurs, cette concentration peut prendre différentes dimensions. On définit la dimension *inter-groupe* lorsqu'on s'intéresse à comment un groupe est co-localisé avec un autre groupe (par exemple : Noirs-Blancs), et la dimension *intra-groupe* lorsqu'on s'intéresse à comment un groupe et co-localisé avec ses semblables (par exemple : Noirs-Noirs).

Mesurer la ségrégation a donné lieu à une littérature abondante. La première génération d'indices s'est focalisée sur deux groupes ou plus (Bell, 1954). Ces indices sont généralement basés sur des unités géographiques et administratives. Mesurer ainsi la ségrégation peut être erroné puisque la position relative des unités n'est pas prise en compte, ce problème est connu sous le *Modifiable Areal Unit Problem (MAUP)*. Le deuxième problème de ces indices est leur testabilité. En général, ces indices ne sont pas testables ou ils sont testables contre des contrefactuels utopiques.[2] Il est alors important de mesurer la magnitude de la ségrégation en comparant des distributions observées avec des distributions contrefactuelles adéquates. Pour remédier partiellement au MAUP, Reardon et O'Sullivan (2004) et Feitosa *et al.* (2007) proposent des indices plus spatiaux mais sans résoudre le problème de testabilité. Cela a donné naissance à une troisième génération d'indices qui est à la fois spatiale et testable (Mele (2013), Echenique et Fryer (2007)).

Nous nous inscrivons dans cette lignée et proposons des mesures basées sur des calculs de distances bilatérales. Dans un cadre unifié, nous adaptons des mesures de concentration de firmes, développées par Duranton et Overman (2005), pour

2. La droite à 45 degrés dans l'indice de Gini reflète une égalité parfaite et jamais atteignable, cela pourrait fausser l'amplitude des résultats.

quantifier de manière continue la ségrégation. Nos mesures satisfont la majorité des propriétés désirables, permettent de mesurer la dimension intra-groupe et inter-groupe de la ségrégation, mais aussi de distinguer l'effet de la race de celui de revenu sur le choix de localisation des individus. Avec des contrefactuels plus réalistes, nous explorons ces différentes dimensions en analysant comment la race peut amplifier la ségrégation par les revenus et vice-versa.

Nous utilisons des catégories raciales officielles dont la terminologie est définie par le recensement américain et explorons des données sur la métropole de New York. Par sa densité, cette ville amplifie les phénomènes urbains ce qui la rend intéressante à étudier. Pour la dimension intra-groupe, nos résultats montrent clairement l'existence de la ségrégation par race mais aussi par les revenus. On trouve que les Noirs, les Hispaniques et les Asiatiques sont significativement ségrégés et le revenu amplifie davantage la ségrégation. Cette dernière est plus importante pour les Noirs alors que la ségrégation par les revenus est plus importante pour les Asiatiques. Pour la dimension inter-groupe, on trouve que les groupes ont généralement tendance à se distancier entre eux. Cet effet est encore plus prononcé pour les pairs Asiatiques- Noirs et Asiatiques-Hispaniques.

Dans le deuxième chapitre, nous nous focalisons sur les origines de la co-agglomération des groupes ethniques dans les grandes villes canadiennes (dimension inter-groupe de la ségrégation). Nous voulons comprendre comment et pourquoi les individus choisissent d'être proches spatialement les uns des autres. Ce chapitre combine deux domaines de science géographique et de psychologie humaine pour répondre à une question d'économie urbaine.

En effet, la Première Loi de la Géographie de Tobler (1970) stipule que *toutes les choses sont reliées entre elles, mais les choses proches sont plus reliées que celles qui sont distantes*. Le Principe de l'Homophilie, en Sociologie et Psycholo-

gie sociale, de McPherson *et al.* (2001) stipule que *les similarités renforcent et engendrent les connexions humaines*. Par conséquent, une question naturelle découle de ces deux affirmations : est-ce que ces connexions humaines vont pousser les groupes ethniques à se localiser proches les uns des autres. Si oui, un corollaire de la Première Loi de la Géographie stipulerait alors que *les choses proches entre elles sont plus similaires que celles qui sont distantes*. Théoriquement, les modèles de Schelling (1969, 1971) montrent que les préférences génèrent des clusters spatiaux, c'est-à-dire, de la ségrégation. Cependant, empiriquement, on n'a pas beaucoup d'évidences et la littérature s'est penchée davantage sur les conséquences de la ségrégation que sur les causes.

Dans ce chapitre, on teste empiriquement ce corollaire en utilisant les données du recensement canadien. Ces données sont riches, elles contiennent de l'information sur les ethnies et fournissent un portrait détaillé sur les groupes raciaux. Le but de ce chapitre est d'essayer de comprendre les origines de la ségrégation intergroupe en essayant de comprendre mieux ces préférences, de les décortiquer, et d'explorer comment elles se transforment dans les choix de co-localisation des individus. Répondre à ces questions est primordial pour mieux élaborer des politiques publiques efficaces favorisant la diversité.

Nos résultats montrent que nos variables non spatiales ont un effet sur nos variables spatiales. En effet, les *préférences homophiliques* qui se traduisent par des distances linguistiques, culturelles, religieuses, génétiques et historico-politiques, ont un effet statistiquement et économiquement significatif sur la colocalisation des individus de différentes origines ethniques. En d'autres mots, les personnes qui sont similaires par ces caractéristiques ont tendance à vouloir être proches les uns des autres. L'homophilie a un effet sur le choix des voisins avec qui nous vivons, et on déduit donc un corollaire qui stipulerait que *les choses proches entre elles sont plus similaires que celles qui sont distantes*.

Nos résultats sont robustes à une série de variables de contrôle économiques et géographiques, et une batterie de tests et de mesures alternatives. Ils sont valides pour les six grandes villes canadiennes, mais avec une hétérogénéité et un gradient est-ouest. Par exemple, la langue a plus d'effets sur la co-localisation à Montréal et Ottawa que Toronto et l'ouest du pays. Nous avons aussi contrôlé pour l'hétérogénéité des agents et les contraintes de choix de localisation en faisant des sous échantillons de populations pauvres, riches, locataires et propriétaires. Nos conclusions tiennent toujours quantitativement et statistiquement, ce qui confirme alors le corollaire de la Première Loi de la Géographie.

Dans le troisième chapitre, nous regardons la distribution jointe firme-individu. Dans un premier temps, nous explorons la décentralisation de l'activité économique dans la métropole de New York et regardons comment la race et le revenu ont un effet sur ce phénomène. Ensuite, dans un deuxième temps, nous testons la déconnexion physique entre les individus de différents groupes et leurs employeurs potentiels, hypothèse appelée aussi: *Spatial Mismatch Hypothesis* (SMH).

Pour ce faire, nous avons besoin de données sur les individus par race ainsi que sur les firmes par industries. Pour les individus, on utilise de l'information sur la composition démographique des quartiers obtenue grâce au recensement Américain. Par ailleurs, les emplois sont obtenus avec des données de NETS : *National Establishment Time Series*, qui incluent jusqu'à 1.5 millions d'établissements et jusqu'à 11 millions d'emplois. Ces données contiennent de l'information sur l'emploi, localisation physique et classification industrielle des firmes. Les données de recensement et celles de NETS sont utilisées pour étudier la décentralisation ainsi que la déconnexion physique entre les individus et leurs employeurs potentiels. Enfin, pour définir ces employeurs potentiels, on utilise aussi des données du CPS : *Current Population Survey*, qui nous renseignent sur la distribution nationale de l'emploi par race et industrie.

Ce troisième chapitre apporte trois principales contributions. Premièrement, nous fournissons une méthodologie novatrice pour mesurer la déconnexion spatiale. Nous adoptons des mesures continues de concentration de firmes, développées par Duranton et Overman (2005, 2008), et les étendons pour capter la distance physique qui sépare les travailleurs de leur employeur potentiel. Cette mesure nous permet de tester différents contrefactuels et de surpasser les faiblesses des mesures existantes. Deuxièmement, la littérature s'est longtemps focalisée sur les Noirs et Blancs. Dans ce papier, on explore aussi d'autres groupes comme les Asiatiques, les Hispaniques, mais aussi les pauvres de chaque groupe. Troisièmement, nous allons au-delà de la littérature courante sur le SMH. On raffine les emplois totaux et définit des opportunités comme des employeurs potentiels pour chaque groupe en utilisant la distribution nationale des emplois par race et revenu.

Nos résultats montrent que les emplois se sont décentralisés de la métropole de New York, pendant que la population totale a tendance à être stable géographiquement entre 1990 et 2010. Cependant, quand nous regardons la population pauvre, on trouve qu'elle s'est significativement décentralisée. Aussi, nous trouvons de l'hétérogénéité raciale dans la décentralisation des individus. En effet, les Blancs se sont rapprochés davantage du centre-ville alors que les Noirs, Hispaniques et Asiatiques ont tendance à se localiser loin du centre-ville. Par ailleurs, en appliquant nos mesures de SMH, deux points importants en ressortent. Premièrement, entre 1990 et 2010, la déconnexion spatiale a augmenté et les individus sont davantage loin de leur lieu de travail. Deuxièmement, en analysant l'effet de la race, on trouve que les Blancs et les Asiatiques sont plus proches de leurs opportunités que les Noirs et les Hispaniques. Ce classement est similaire à celui du chômage par race aux États-Unis, ce qui suggère un lien potentiellement causal entre cette déconnexion physique et l'employabilité de certains groupes.

Le reste de la thèse est organisé comme suit : dans la prochaine section, nous

détaillons nos mesures de ségrégation et discutons de ses applications. Ensuite, dans la section suivante, nous nous focalisons sur les causes de la ségrégation inter-groupe. Dans la troisième section, nous nous concentrons sur la distribution jointe firme-individu, et explorons la décentralisation ainsi que la déconnexion physique. Enfin, nous terminons la thèse par une conclusion générale.

CHAPITRE I

DISTANCE-BASED SEGREGATION MEASURES

Abstract

We apply point-pattern based measures of geographic concentration—usually used to assess the extent and statistical significance of the spatial clustering of firms or plants—to the measurement of segregation. Our measures of 'excess segregation' satisfy a number of desirable properties and allow to assess the geographic distribution of groups using spatially fine-grained data. They allow for statistical testing of the observed patterns against various reference distributions and can be applied to the measurement of segregation (within groups) and isolation (between groups). We use them to also partly disentangle segregation by race from the geographic concentration of poverty.

Keywords : Segregation ; point-pattern based concentration measures ; clustering ; statistical significance tests.

JEL Classification : R23.

1.1 Introduction

Segregation by race and income are widespread and pervasive phenomena within cities across the world, with well-documented negative consequences for the segregated populations—more crime, worse role models and peer networks, and less social mobility, to name but a few. Hence, addressing the problem of segregation is an important policy issue. However, to do so requires first to appropriately measure segregation, and second to better understand its underlying causes. We contribute in this paper to the measurement of segregation and develop new tools for that purpose. We also show how these tools can be used to refine the measurement along various lines and to sort out—at least partly—the strength of some of the underlying causes.

Measuring segregation is a complex endeavour with a long history. The voluminous extant literature has developed a wide range of different indices. First generation indices look at segregation between two groups (Bell, 1954) or between several groups (Morgan, 1975). They are all area based and computed using the distribution of racial shares across different administrative geographic units. James et Taeuber (1985) and Massey et Denton (1988) provide good surveys. As is well known, such indices are inherently non-geographic—since the relative position of the areal units does not matter—and sensitive to the geographic scale chosen for the analysis—a problem known as the Modifiable Areal Unit Problem (MAUP).[1]

1. Conventional indices suffer from their dependence on the geographical partition of the study area. Boundaries are often established at the convenience of administration services, and are arbitrary and not based on any socioeconomic concept of neighborhood. The shape and scale of these units can lead to different assessments of segregation, even if the spatial distribution of individuals remains constant. As a result, the comparability across time (dynamics of segregation) and space (comparison of two cities or countries) is problematic.

Furthermore, existing indices usually do not provide any criterion by which we can assess their statistical significance. In a nutshell, it is usually unclear whether the segregation we observe could be due to 'chance' only.[2] To overcome the first issue and to make the indices more spatial, Reardon et O'Sullivan (2004) and Feitosa *et al.* (2007) propose second generation indices that extend the classical measures in Massey et Denton (1988) to account for spatial characteristics. These approaches are still location specific since they deal at best with an ad hoc partition of space into 'local neighborhoods'. To overcome the MAUP and to tackle the issue of statistical significance, third generation indices—truly spatial indices—have been proposed. Using a spatial Poisson point process, Mele (2013) develops an alternative family of indices that allow to estimate a continuous spatial density for a given racial group and to then compare it to a counterfactual distribution. Using data on friendship networks in the context of school segregation, Echenique et Fryer (2007) exploit the structure of social networks and propose an index based on social interactions that can be disaggregated to the individual level. It is hence a measure specific to an individual which has the advantage to avoid issues related to areal units. However, it is more suited to aspatial contexts, i.e., segregation in schools or firms.

Despite the progress made, existing indices do not to our knowledge allow to disentangle the different underlying causes for segregation. The two most important

---

2. Classical indices of inequality and segregation also often consider utopian benchmarks. Think, for example, about the 45-degree line for the Gini index. The unrealistic nature of this benchmark—perfect equality—is problematic when it comes to interpretations, e.g., to what extent a city is segregated or not. It is thus crucial to test whether there is a significant difference between the empirical distribution of a given group and a 'realistic' benchmark, keeping in mind that some degree of observed unevenness is not necessarily segregation per se. As in dartboard games, throwing darts on the board—even with a random hand—could result in darts clustering in one section (Ellison et Glaeser, 1997).

causes may be race per se ('homophily') and income ('poverty'). Is the observed pattern driven by sorting along racial lines, or is is sorting along income more generally ?[3] A better measure of the magnitude of segregation by race on top of segregation by poverty would appear to be useful.

We develop new tools that allow us to measure the extent of racial segregation in cities. Contrary to previous measures—which are mainly based on the distribution of racial shares across areal units—our measures build on the distribution of bilateral distances across individuals. To this end, we adapt the continuous distance-based measures of firm concentration, pioneered by Duranton et Overman (2005), to the measurement of segregation. These measures satisfy many desirable properties and allow us to think about two dimensions of segregation—evenness-clustering and exposure-isolation—within a unified framework. Crucially, our measures are truly spatial and allow for statistical testing of the observed patterns of segregation by simulating random distributions within appropriately chosen benchmarks. They also allow us—by comparing the observed distributions to those benchmarks using a case-control design—to partly disentangle segregation by race from segregation by poverty.

We illustrate our measures using New York core-based statistical area (NYCBSA) census data. The scale, density, and diversity of the largest metropolitan area in the U. S. provides an ideal setting to look at segregation by race and by poverty. Since we do not have geocoded data at the level of each individual, we make use of census block data to compute a variety of versions of our measures and to compare them against a rich set of benchmarks. In particular, we estimate measures of

3. Sorting along income is what Cutler *et al.* (1999) refer to as 'decentralized racism'. In our analysis, we can control for 'decentralized racism' using income, but we are unable to disentangle homophily from 'collective action racism'. The latter played historically a large role but seems to become less important in more recent years.

segregation by race on top of segregation by poverty ; segration by poverty on top of segregation by race ; as well as exposure of a group to (or isolation of a group from) another group, conditional on segregation patterns within the own race. Finally, we also construct more complicated 'estimated' benchmarks, where we predict a counterfactual distribution of people across the NYCBSA, based on observable block-level characteristics.

Previewing our main empirical results, we find clear patterns of segregation by both income and race. Our results reveal that Black, Hispanic, and Asian are segregated, with income amplifying even further the magnitude of their segregation. Within 5 kilometers distance, there is a 5.6 percentage point excess segregation of Black compared to the overall population distribution. In other words, a pair of African-Americans drawn at random in the NYCBSA are 5.6 percentage points more likely to live within 5 kilometers distance than are a pair of New Yorkers draw at random in general. The figures for Hispanic and Asian are 3.1 and 2.8 percentage points, respectively. On top of segregation by race, we find that there are 3.2, 4.2, and 6.3 percentage points of excess segregation by income for Black, Hispanic, and Asian with 5 kilometers distance, respectively. This suggests that segregation by race is strongest for Black, while segregation by poverty (conditional on segregation by race) is strongest for Asian. Turning to measures of the pairwise exposure between groups, our results show that groups tend to isolate themselves from other groups. This effect is again especially strong for poor Asian with respect to the other groups : they are 5.2 and 4.2 percentage points less exposed to poor Black and to poor Hispanic, conditional on segregation by race. In other words, even when conditioning on observed patterns of racial segregation, poor Asian tend to cluster strongly (and significantly) together.

The remainder of the paper is organized as follows. Section 1.2 first provides an overview of geographic concentration measures and then discusses the dimen-

sions of segregation. We also review the properties an ideal measure of segregation should satisfy according to the consensus in the literature. Section 1.3 lays out our methodology and shows that point-pattern based measures of geographic concentration satisfy most properties of an ideal index of segregation. Section 1.4 discusses details of the empirical implementation, while Section 1.5 provides an application to the measurement of segregation by race and poverty in the NYCBSA. Last, Section 1.7 concludes. We relegate technical details and additional results to a set of appendices.

1.2 Measures of geographic concentration and segregation

Segregation and the geographic concentration of firms are both spatial phenomena that share similarities and hence can be approached drawing on similar sets of tools. On the one hand, when studying the geographic concentration of industries the researcher is interested in the spatial proximity of plants or workers sharing some common characteristics (e.g., belonging to a specific industry or being exporters) or interacting along relevant dimensions (e.g., trading inputs or exchanging knowledge). On the other hand, when studying segregation the researcher is interested in the geographic concentration of individuals sharing some common characteristics (e.g., race, ethnic background or poverty status) or the geographic exposure of individuals to others with different characteristics. Our approach to measuring segregation hence naturally draws on techniques developed previously in the literature concerned with the geographic concentration of industries (agglomeration) and the geographic concentration of industry pairs (coagglomeration). The parallel for individuals is to look at the spatial distribution of people belonging to the same group (within group) or the colocation—exposure—of some groups to other groups (between groups).

1.2.1 Geographic concentration

There are many ways to measure the geographic concentration of economic activity and the coagglomeration of industries (see, e.g., Chapter 8 in Combes *et al.* 2008). The most well-known measure is probably the Ellison-Glaeser index (Ellison et Glaeser 1997 ; Ellison *et al.* 2010).[4] This index can be used to either assess the geographic concentration of a single industry (agglomeration) or the colocation of industry pairs (coagglomeration). The Ellison-Glaeser index satisfies a number of desirable properties. Firstly, it is comparable across industries. Secondly, it is defined against a well-defined benchmark. Thirdly, it controls for 'lumpiness', i.e., the fact that plants are indivisible units so that the geographic concentration of industries necessarily partly reflects the concentration of workers within individual firms. Nevertheless, it also has at least two serious drawbacks. First, it is computed for predetermined spatial units and therefore suffers from the well know Modifiable Areal Unit Problem (MAUP). Second, it is essentially aspatial in nature—permuting the position of individual spatial units does not change the value of the index. Last, statistical tests of the significance of the observed location patterns are rarely carried out in practice, though simulation approaches to doing so have been developed (Cassey et Smith, 2014). We will return to these three points later since they obviously also affect many segregation measures.

To cope with these three problems, the literature has gradually moved away from area-based measures and has adopted point-pattern based measures borrowed from spatial statistics. Duranton et Overman (2005) have proposed a new way to measure the geographic concentration of industries and the coagglomeration

4. See Maurel et Sedillot (1999) for a closely related index which they applied to French data.

of industry pairs.[5] Their index exploits the microgeographic location patterns of individual plants and, therefore, obviates the need for area-based measures. The index is comparable across industries, offers flexibility in defining the benchmark, allows for a clear interpretation of the degree of concentration, allows to test the statistical significance of the observed location patterns and entirely solves the MAUP problem because it does not rely on any spatial units.

The segregation indices we propose below are based on the Duranton et Overman (2005) approach to measuring geographic concentration. They can be linked to the different dimensions of segregation highlighted in the literature (Massey et Denton, 1988) and satisfy most of the desirable properties of good segregation indices (James et Taeuber, 1985; Reardon et O'Sullivan, 2004). More importantly, these measure will also allow us to partly separate the effects of racial segregation from those of sorting along income.

1.2.2 Segregation

How to measure segregation is an old question that dates back to at least Bell (1954). There was little consensus until the 1980s about what properties segregation measures should satisfy and which of the numerous measures researchers should prioritize. This state-of-affairs was largely due to "the absence of a clear set of criteria, derived from a comprehensive definition of segregation" (James et Taeuber, 1985, p.2). In two important contributions, James et Taeuber (1985) and Massey et Denton (1988) comprehensively delineated the different dimensions of segregation and established a clear set of criteria that segregation indices should satisfy—drawing largely from the literature on inequality measures.

---

5. See Marcon et Puech (2017) for more discussion on the distance based approach to measure concentration.

Dimensions of segregation. In their seminal contribution, Massey et Denton (1988) analyzed 20 segregation indices to highlight five distinct dimensions of segregation :

(D1) *Evenness :* the proportional spatial distribution of a given group across geographic units (e.g., census blocks), i.e., "the degree to which groups are distributed proportionally across areal units in a city" (Ibid., p.309) ;

(D2) *Exposure :* the possible interactions between two groups that differ along some dimension (say race or income), i.e., "the extent to which members of different groups share common residential areas within cities" (Ibid., pp.309–310) ;

(D3) *Clustering :* how a given group of people is closely packed in space, i.e., "the degree to which minority areas are located adjacent to each other" (Ibid., p.310) ;

(D4) *Centralization :* the degree to which a group is located near the city center ; and

(D5) *Concentration :* the relative amount of physical space occupied by a given group. [6]

For our analysis, only evenness, exposure, and clustering—**(D1)** to **(D3)**—will be of interest. Massey et Denton (1988) have shown these to be the most important dimensions that explain best the patterns in the data. Concentration is clearly less important in their empirical findings, and it is also harder to distinguish as a

---

6. The most widely used segregation measures in the literature are : (i) the dissimilarity, Gini and entropy indices for the dimension of *evenness* ; (ii) the correlation ratio, interaction, and isolation indices for the dimension of *exposure* ; (iii) the absolute clustering, spatial proximity, and distance-decay interaction indices for the dimension of *clustering* ; (iv) the relative and absolute centralization indices for the dimension of *centralization* ; and (v) the delta, absolute, and relative concentration indices for the dimension of *concentration*.

separate dimension of the segregation phenomenon. Centrality has conceptually little meaning. It is not clear that centrality is a dimension of segregation per se, and it seems specific to the structure of U.S. cities (e.g., measuring segregation by centrality would make little sense in most European cities such as Paris, where segregated neighborhoods with low socio-economic status are mostly found in the 'banlieus défavorisées' far away from the 'center').

Although clustering will be important in our subsequent analysis, we do not consider it as a separate dimenion of segregation in what follows. There are two reasons for that. First, as argued by Reardon et O'Sullivan (2004), being unevenly distributed is a reflection of clustering in itself, so that evenness and clustering should be collapsed into an evenness-clustering dimension.[7] Also, not having individuals from other groups nearby (i.e., no exposure) is the flip-side of isolation, so that exposure and isolation should be combined into an exposure-isolation dimension. Second, contrary to most existing measures of segregation, our measures are fully spatial and account for the geographic distribution of all individuals. Hence, they naturally encapsulate the clustering dimension,

To summarize, we will be interested in essentially two dimensions of segregation : *evenness-clustering* (henceforth, **E/C**) ; and *exposure-isolation* (henceforth, **E/I**).

Properties of segregation measures. In their seminal contribution, James et Taeuber (1985) postulated four basic properties that a good measure of segregation should verify :

(P1) *Organizational equivalence* : if an area is split into several sub-units that have the same group proportions than the original area, the index should

7. Reardon et O'Sullivan (2004) further argue that centralization and concentration are included in the evenness-clustering dimension as subcategories.

be unchanged ;

(P2) *Size invariance* : the index should be unchanged if we multiply the size of each group within the geographic area we analyze by a constant ;

(P3) *Compositional invariance* : the index should be unchanged if we multiply the size of each group in each areal unit by a constant factor ;

(P4) *Transfer principle* : moving individuals between areas should decrease the segregation measure if their group proportion is higher in the sending than in the receiving area.

The foregoing properties were developed for aspatial indices, i.e., indices that consider the racial composition of areas but that disregard how these areas are located relative to one another in space. Reardon et O'Sullivan (2004) added two desirable properties specific to spatial segregation indices : [8]

(P5) *Arbitrary boundary independence* : changing spatial boundaries should not change the measure, i.e., there is no MAUP ;

(P6) *Scale interpretability* : the measure should vary between zero and one and also be able to capture 'hyper-integrated' cases (in which case it takes

---

8. Reardon et O'Sullivan (2004) further propose generalizations of several properties to the spatial context, taking into account how units are organized in space : **(P1')** *location equivalence* (organizational equivalence) ; **(P2')** *population density invariance* (size invariance) ; **(P3')** *composition invariance* (compositional invariance) ; and **(P4')** *transfers and exchanges* (transfer principle). Properties **(P1)**–**(P4)** (or equivalently **(P1')**–**(P4')**) are desirable for 'first generation' indices that do not consider more than two groups at the same time. Reardon et Firebaugh (2002) adapted these properties to the case of multigroup indices. They also split the transfer principle **(P4)** into 'transfer' and 'exchange', and added two new criteria : *additive organizational decomposability* and *additive group decomposability* (further adapted by Reardon et O'Sullivan 2004 to the spatial context as *additive spatial decomposability*). We disregard in what follows these latter two properties. The former requires spatial areas, which we do not have ; while the latter applies to multigroup indices, which we do not consider in this paper.

negative values).

Finally, several more recent contributions have added new items to the list of desirable properties that segregation measures should satisfy :

(P7) *Disaggregation* : indices should disaggregate in order to allow for an analysis of segregation at the individual level ;

(P8) *Significance tests* : a measure of segregation should allow for statistical testing, i.e., the researcher should be able to assess how likely the observed segregation is compared to some well-defined benchmark.

The Spectral Segregation Index (SSI) of Echenique et Fryer (2007)—which builds on a network representation of social interactions—satisfies the property of disaggregation.[9] The same holds true for the local segregation index of Feitosa *et al.* (2007). These indices also allow for statistical testing using permutation tests for random assignment of characteristics to locations. Last, Mele (2013) proposes a measure of segregation based on spatial point processes. His Poisson index of segregation can be disaggregated to the individual level and is measured as deviation from a baseline distribution with 'random labeling'.

To summarize, good measures of segregation should satisfy properties **(P1)**–**(P8)**. In what follows, we will propose measures that subsume well the evenness-clustering and exposure-isolation dimensions while satisfying most of the proper-

9. Echenique et Fryer (2007) do not directly test the significance of their SSI. However, they compute a 'baseline SSI' for a set of 1,000 artificial cities with 100 households each to assess the deviation of the empirical SSI from that baseline. Their index is well suited to an analysis where social networks are known, but it is less-well suited to the spatial context where the network must be approximated by either distance or nearest neighbors. They also posit that "an individual is more segregated the more segregated are the agents with whom she interacts" (Ibid. p.441), which is a specific instance of the clustering dimension (**D3**).

ties **(P1)** to **(P8)** either exactly or approximately.

1.3 Measuring segregation using distance-based methods

The key difference between our measures and those in the literature is a change in the statistical unit of the analysis. It is not, as most segregation measures, an area-based but a point-based measure : instead of using the proportions of individuals of different types in an area we will use the geographic distances between pairs of individuals. [10]

1.3.1 K-densities and their properties

Consider a group of N_x individuals who share some common characteristic x (e.g., ethnic origin or poverty status). Let $\widetilde{d}_{ij}$ denote the distance between individuals i and j, which we assume to be symmetric ($\widetilde{d}_{ij} = \widetilde{d}_{ji}$). We want to measure the segregation of the N_x individuals and assess its statistical significance compared to some benchmark distribution. To do so, we use the K-densities pionneered by Duranton et Overman (2005) (henceforth, DO). The probability density function (PDF) of the distribution of bilateral distances within group x is estimated as follows :

$$\widehat{k}^{xx}(d) = \frac{1}{\widetilde{h} N_x (N_x - 1)/2} \sum_{i=1}^{N_x - 1} \sum_{j=i+1}^{N_x} f\left(\frac{d - \widetilde{d}_{ij}}{\widetilde{h}} \right), \tag{1.1}$$

where d is the distance at which the K-density is evaluated, f is a Gaussian kernel, and $\widetilde{h}$ the bandwidth set following Silverman's rule-of-thumb. Observe that the measure (1.1) easily extends to the case of two groups. Assume that there are

10. Our measure is close to that of Echenique et Fryer (2007), but uses the geographic distance between individuals instead of some network distance. We compute the measure using bilateral distances between pairs of individuals belonging to same group or pairs belonging to different groups.

N_x individuals who share some common characteristic x and N_y individuals who share some common characteristic y. Then the probability density function (PDF) of the distribution of bilateral distances between groups x and y is estimated as follows :

$$\widehat{k}^{xy}(d) = \frac{1}{\widetilde{h} N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} f\left(\frac{d - \widetilde{d}_{ij}}{\widetilde{h}}\right). \tag{1.2}$$

Assume that the exact locations of individuals are not observed, but that we know their distribution across census blocks. In other words, instead of observing N_x and N_y individual locations we observe a distribution of counts n_i^x and n_i^y across n_x and n_y blocks, with $\sum_i n_i^x = N^x$ and $\sum_i n_i^y = N^y$. We can adapt the foregoing measures by assuming that individuals are assigned to the block centroid to obtain weighted versions of (1.1) and (1.2) as follows :

$$\widehat{k}_W^{xx}(d) = \frac{2}{h \sum_{i=1}^{n_x} \sum_{j=i}^{n_x} n_i^x n_j^x} \sum_{i=1}^{n_x} \sum_{j=i}^{n_x} n_i^x n_j^x f\left(\frac{d - d_{ij}}{h}\right), \tag{1.3}$$

and

$$\widehat{k}_W^{xy}(d) = \frac{1}{h \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} n_i^x n_j^y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} n_i^x n_j^y f\left(\frac{d - d_{ij}}{h}\right), \tag{1.4}$$

where d_{ij} is the distance between the block centroids, and h is the bandwidth set following Silverman's rule-of-thumb. [11]

The weighted K-densities (1.3) and (1.4) thus describe the distribution of bilateral distances between individuals under the assumption that within blocks individuals are assigned to the centroid. In the case of (1.3), for example, this amounts to replacing

$$\frac{1}{\widetilde{h} N_x (N_x - 1)/2} f\left(\frac{d - \widetilde{d}_{ij}}{\widetilde{h}}\right) \quad \text{with} \quad \frac{n_i^x n_j^x}{h \sum_{i=1}^{n_x} \sum_{j=i}^{n_x} n_i^x n_j^x} f\left(\frac{d - d_{ij}}{h}\right).$$

---

11. In (1.1), we do not count the distance of an individual to himself. However, we do so in (1.3). Since there are n_i^x individuals in block i, and since we want to count the distances between individuals within the block, we make those enter into our computations.

It should be clear that the approximation above induces measurement error. We do not think that this is a serious problem. When working with small spatial units—such as census blocks in the US or dissemination areas in Canada—all individuals are 'close' to the centroid anyway. [12] Furthermore, the kernel smoothing that we apply to the distribution is precisely designed to take into account this type of measurement error in distances (see Duranton et Overman 2005, p.1083 for a discussion). Last, it is also unclear whether geographic distance is the correct measure of 'social interactions' that are important for segregation (see, e.g., Echenique et Fryer 2007, who note this but also use geographic distance to infer social interactions). Kernel smoothing again helps here.

Additional measures of segregation—both within and between groups—can be obtained by considering the cumulative distribution function (CDF) of the K-densities :

$$\widehat{K}^{xx}(d) = \int_0^d \widehat{k}^{xx}(i)\mathrm{d}i \quad \text{and} \quad \widehat{K}_W^{xx}(d) = \int_0^d \widehat{k}_W^{xx}(i)\mathrm{d}i. \tag{1.5}$$

and

$$\widehat{K}^{xy}(d) = \int_0^d \widehat{k}^{xy}(i)\mathrm{d}i \quad \text{and} \quad \widehat{K}_W^{xy}(d) = \int_0^d \widehat{k}_W^{xy}(i)\mathrm{d}i. \tag{1.6}$$

The cumulative distribution (1.5) at distance d provides a measure of the share of pairs within group x that is located less than d from each other. For example, a value of 0.33 at $d = 10$ kilometers means that one third of the individuals with characteristic x are located less than 10 kilometers from each other. Analogously, the cumulative (1.6) at distance d provides a measure of the share of pairs formed by individuals from groups x and y that is located at less than d from each other.

12. In Canada, the centroids of the dissemination areas—called 'representative points'—are population weighted, which further reduces measurement error. Observe also that in downtown areas with a higher frequency of tall buildings, the centroid is likely to be the exact location of a given building.

Properties of K-densities. Let us start with a few basic observations regarding the properties of the measures (1.1)–(1.6) :

(P1) *Organizational equivalence* : this is trivial for the indices (1.1), (1.2), and their associated CDFs since they do not rely on any spatial subdivisions. However, the weighted versions of the indices—(1.3), (1.4), and their associated CDFs—do change with the subdivisions. However, given small spatial units and large sample sizes, and given that the indices are kernel-smoothed over the whole distribution, the changes are not likely to be substantial (as they can be for area-based indices).

(P2) *Size invariance* : the index should be unchanged if we multiply the size of each group within the geographic area we analyze by a constant. This is trivially satisfied because the measures (1.1) and (1.3) are distribution functions, whereas (1.5) is a cumulative distribution.

(P3) *Compositional invariance* : the index should be unchanged if we multiply the size of each group in each areal unit by a constant factor. Again, this is trivially satisfied because the measures (1.1) and (1.3) are distribution functions, whereas (1.5) is a cumulative distribution.

(P4) *Transfer principle* : moving an individual between areas r and s (i.e., changing n_r and n_s to $n_r - 1$ and $n_s + 1$) decreases the segregation measure if the group size is larger in the sending than in the receiving area and if the move is from a 'more accessible area' to a 'less accessible area'. To understand why moving individuals from accessible and large places to less accessible and small places decreases the measure of segregation at short distances, note that such a move decreases the number of zero distances within areas, whereas it increases (on average) the distances with the other areas. Because our measures are distributions, this means that the density at long distances must increase in response to having less density at

short distances. It then follows that the cumulative $\widehat{K}_W(n_r, n_s)$ first-order stochastically dominates $\widehat{K}_W(n_r - 1, n_s + 1)$. In this sense, the transfer principle is satisfied. [13]

(P5) *Arbitrary boundary independence* : since there are no spatial boundaries, changing them cannot influence our measures. In other words, there is no MAUP for our ideal index (1.1) which is free from any spatial scale. However, this is not true for the operational version (1.3) as a consequence of the 'lumpy approximation' we have to make given the data availability. This holds true for all other measures that have been proposed in the literature (see, e.g., Echenique et Fryer 2007 and Mele 2013) since we do not observe the exact distribution of the whole population. However, working at small geographic scales and using weighted centroids should minimize the effects of geographical boundaries. [14]

(P6) *Scale interpretability* : because (1.3) and (1.5) are distribution functions and cumulative distributions, they vary between zero and one. As we explain later—after talking about the benchmark—our operational measure can also capture 'hyper-integrated' cases where it takes negative values.

(P7) *Disaggregation* : our measures can be disaggregated to allow for an ana-

---

13. One important remark is in order to qualify this statement. One can show that the transfer principle holds true for the raw distribution of the bilateral distances. It is, however, not clear how kernel smoothing may affect this property. The reason is that the kernel is not monotone with respect to $d_{ij} - d$ and that the bandwidths before and after the transfer usually differ (since the distribution of distances has changed). We conjecture that these effects are small enough for large enough samples so that moving units in a way postulated in the proposition satisfies the transfer principle.

14. Another problem are 'edge effects' in the sense that the study area is bounded. Points close to the border may suffer from mis-measurement because potential interactions with points outside the study area are discarded.

lysis of segregation at the individual level. By definition, (1.3) and (1.5) can be expressed in individual form for person p as follows :

$$\widehat{k}_W^p(d) = \frac{1}{h \sum_{i=1}^{n} \sum_{j=1}^{n} n_i n_j} \sum_{j=1}^{n} n_j f\left(\frac{d - d_{p(i)j}}{h}\right), \text{and} \tag{1.7}$$

$$\widehat{K}_W^p(d) = \int_0^d \widehat{k}_W^p(i)\mathrm{d}i, \tag{1.8}$$

where $d_{p(i)j}$ is the distance of person p's block i and block j. We can standardize (1.7) and (1.8) to measure the relative contribution of each individual to the global measure at distance d.

Measures of segregation aim to capture the deviation of the geographic distribution of a group from some underlying benchmark. We hence have to think about what is this benchmark. In the next section, we explain how we can test whether or not the geographic distribution of some minority group follows that of some reference population or whether there are statistically significant differences. In other words, our measures satisfy :

(P8) *Significance tests* : The magnitude of our measures of segregation can be interpreted with respect to some well-defined benchmark and allow for statistical testing.

1.3.2 From K-densities to measures of segregation

The K-densities presented thus far do not compare the distribution of groups x and y to some underlying benchmark. They are *absolute measures* that state how closely packed individuals with characteristic x are in space, or how close individuals with characteristic x are from individuals with characteristic y. To obtain *relative measures*—as required for the analysis of segregation—a comparison with an appropriate benchmark is required. The question of interest is whether the dis-

tribution of some group (e.g., African Americans) follows that of some reference population. We now explain how we construct proper measures of segregation from the K-densities and how they satisfy our last property of significance tests. We then also show that the 'evenness-clustering' **(E/C)** and 'exposure-isolation' **(E/I)** dimensions of segregation can be approached in a unified way using these measures.

Statistical testing. Similar to Duranton et Overman (2005), we compare the K-densities with an appropriately defined benchmark distribution. To fix ideas, consider the ideal case in which we would have access to all individual locations, so that we can estimate (1.1) or (1.2) and compute the associated cumulative distribution (1.5) or (1.6). Assume that we look at the location of individuals with characteristic x in a metropolitan area. Let $\mathcal{X}$ denote the set of locations with individuals of type x, with $|\mathcal{X}| = N_x$. Let $\mathcal{C}$, where $\mathcal{X} \subset \mathcal{C}$, denote a set of control locations, with $|\mathcal{C}| = N_c$. In other words, we consider a larger set of $N_c > N_x$ locations that includes locations in which individuals do not have the characteristic x.

To perform our case-control design, we first estimate the observed K-density $\widehat{K}^{xx}(d)$ for all $(i,j) \in \mathcal{X}^2$ and for some range of distances d. We then draw a random set of N_x individuals from $\mathcal{C}$, and estimate the K-density $\widetilde{K}^{xx}(d)$ for the bilateral distances. This may be viewed as a counterfactual situation where all indiviuals with characteristic x would be randomly reshuffled across a set $\mathcal{C}$ of control locations. We repeat this exercise 1,000 times.[15] The distribution of the counterfactual K-densities is then used to derive upper and lower bounds, $\overline{K}^{xx}(d)$

15. We could use more than 1,000 replications, depending on the precision and the computation time. The results are not very sensitive to this when we run estimations on large samples, but computing costs can be extremely high—even with fast computer codes—as we explain later.

and $\underline{K}^{xx}(d)$, of confidence bands at every distance d. The latter can be used for statistical testing of the significance of segregation patterns : if the K-density lies above the upper bound of the confidence band ($\widehat{K}^{xx}(d) > \overline{K}^{xx}(d)$), individuals with characteristic x are significantly more concentrated ($\mathbf{E/C}$ dimension) than the set of control individuals at distance d ; and if the K-density lies below the lower bound of the confidence band ($\widehat{K}^{xx}(d) < \underline{K}^{xx}(d)$), they are significantly more dispersed at distance d. Similarly, if $\widehat{K}^{xy}(d) > \overline{K}^{xy}(d)$, individuals with characteristics x and y are significantly more exposed ($\mathbf{E/I}$ dimension) than the control ; and if $\widehat{K}^{xy}(d) < \underline{K}^{xy}(d)$, individuals with characteristics x and y are significantly more isolated from one another than the control. Using the distribution of the counterfactual K-densities, we then define two-sided 90% confidence intervals. For each d, we rank the values of the simulated K-densities in ascending order and keep the distribution between the 5th and the 95th percentile. [16]

When implementing this procedure for the measures (1.3) based on area centroids, we need to tackle a number of computational challenges. We will spell out the technical details in the next section. There, we will also explain how we select the benchmarks that we consider.

Subsuming dimensions of segregation. How do our measures (1.3) and (1.5) relate to the dimensions of segregation ? As explained before, we consider only two dimensions : (i) evenness-clustering ($\mathbf{E/C}$) ; and (ii) exposure-isolation ($\mathbf{E/I}$). The basic idea underlying the construction our measures of segregation is to look

16. Because of multiple hypothesis testing, we adjust the confidence bands using a standard Bonferroni procedure. We do not correct for spatial correlation in the confidence bands as in Duranton et Overman (2005). In our application, sample sizes are so large that this does not make any substantive difference. We also use linear interpolation to determine the appropriate percentile values when required.

at the gap between the empirical distributions and the confidence bands, i.e., deviations from the benchmark. Starting with the **E/C** dimension for a group x, if $\widehat{K}_W^{xx}(d) > \overline{K}^{xx}(d)$ then group x is overrepresented at distance d compared to the benchmark distribution. In other words, group x is unevenly distributed, i.e., it is *relatively more concentrated* than the benchmark accepting some level of statistical risk as embodied by the confidence band. Let

$$\text{ec}^{xx}(d) \equiv \max\left\{\widehat{K}_W^{xx}(d) - \overline{K}^{xx}(d),\, 0\right\} \tag{1.9}$$

be our measure of *excess concentration* of group x at distance d. Conversely, if $\widehat{K}_W^{xx}(d) < \underline{K}^{xx}(d)$, then group x is underrepresented at distance d compared to the benchmark distribution. Group x is unevenly distributed, i.e., it is *relatively more dispersed* than the benchmark (it is 'hyper-integrated' in the terminology of Reardon et O'Sullivan 2004). As argued before, a good segregation measure should be able to capture that case and should take negative values. Let

$$\text{ed}^{xx}(d) \equiv \min\left\{\widehat{K}_W^{xx}(d) - \underline{K}^{xx}(d),\, 0\right\} \tag{1.10}$$

be our measure of *excess dispersion* of group x at distance d. Observe that (1.9) and (1.10) capture the dimension of evenness-clustering (**E/C**). They capture the over- or underrepresentation of group x compared to some benchmark, and they take into account the whole distribution of group x across space (i.e., they correct for clustering). We equate evenness with *randomness* : if $\text{ec}^{xx}(d) \in \left[\underline{K}^{xx}(d), \overline{K}^{xx}(d)\right]$ and $\text{ed}^{xx}(d) \in \left[\underline{K}^{xx}(d), \overline{K}^{xx}(d)\right]$ then the observed distribution of bilateral distances between members of group x is not statistically distinguishable at distance d from a distribution where members of x would be randomly distributed across $\mathcal{C}$.

The measures (1.9) and (1.10) naturally extend to the two-group case to capture the dimension of exposure-isolation (**E/I**). Consider groups x and y, and assume

that $\widehat{K}_W^{xy}(d) > \overline{K}^{xy}(d)$. Then, groups x and y are jointly overrepresented at distance d compared to the benchmark distribution. In other words, groups x and y are unevenly distributed, i.e., they are *relatively more exposed* to each other than the benchmark. Let

$$\text{ee}^{xy}(d) \equiv \max\left\{\widehat{K}_W^{xy}(d) - \overline{K}^{xy}(d),\, 0\right\} \tag{1.11}$$

be our measure of *excess exposure* of group x to group y (or, by symmetry, of group y to group x) at distance d. Conversely, if $\widehat{K}_W^{xy}(d) < \underline{K}^{xy}(d)$, then groups x and y are jointly underrepresented at distance d compared to the benchmark distribution. In other words, groups x and y are unevenly distributed, i.e., they is *relatively more isolated* from each other than the benchmark. Let

$$\text{ei}^{xy}(d) \equiv \min\left\{\widehat{K}_W^{xy}(d) - \underline{K}^{xy}(d),\, 0\right\} \tag{1.12}$$

be our measure of *excess isolation* of group x from group y (or, by symmetry, of group y from group x) at distance d.

The relative measures defined above compare the empirical frequency distribution of bilateral distances with simulated frequency distributions drawn at random from some underlying benchmark. We pick up 'excess segregation' (or 'excess integration', 'excess isolation', or 'excess exposure') at distance d when we reject the null hypothesis of randomness. How can we measure the magnitude of segregation ? We consider the cumulative of the 'excess' over all distances $d \leq \overline{d}$, formally given by :

$$\text{EC}^{xx}(\overline{d}) \equiv \int_0^{\overline{d}} \text{ec}^{xx}(i)\,\text{di}, \quad \text{ED}^{xx}(\overline{d}) \equiv \int_0^{\overline{d}} \text{ed}^{xx}(i)\,\text{di} \tag{1.13}$$

and

$$\text{EE}^{xy}(\overline{d}) \equiv \int_0^{\overline{d}} \text{ee}^{xy}(i)\,\text{di}, \qquad \text{EI}^{xy}(\overline{d}) \equiv \int_0^{\overline{d}} \text{ei}^{xy}(i)\,\text{di}. \tag{1.14}$$

These are the measures of segregation that we use in what follows. [17]

---

17. Because $\widehat{K}_W^{xx}$ and $\widehat{K}_W^{xy}$ are distributions, they necessarily sum to one. The same holds true

To summarize, the difference between the K-density of group x and the confidence band provides a natural measure of 'unevenness'. Furthermore, the difference between the K-density of two groups x and y and the confidence band provides a natural measure of 'exposure' of the two groups to each other. Last, the cumulatives of our measures up to distance d provide natural metric for the degree of 'excess segretation' or 'excess exposure'. All our measures have a simple probabilistic interpretation. For example, if $\text{EC}^{xx}(500m) = 0.1$, this means that two members of group x drawn at random have a 10% higher chance of being less than 500m from one another than if they would be drawn from the benchmark distribution, and accepting a statistical risk of error of 10%.

1.3.3 Discussion and limitations

Our measures of segregation have several advantages. Firstly, they capture—within a unified framework—the most important dimensions of segregation, subsuming the evenness-clustering and the exposure-isolation dimensions (**(D1)**–**(D3)**). While measures of evenness and of isolation usually differ on a conceptual basis in the literature, they can be viewed in a similar way using our approach. Secondly,

---

for the simulated counterfactual K-densities $\widetilde{K}_W^{xx}$ and $\widetilde{K}_W^{xy}$. Hence, if group x is overrepresented at distance $d \leq \overline{d}$, it must be underrepresented at some distances beyond $\overline{d}$. The same holds true if group x is underrepresented at $d \leq \overline{d}$: it must then be overrepresented at some distances beyond $\overline{d}$. Duranton et Overman (2005) suggest to consider that there is excessive geographic concentration if the K-density exceeds the upper bound of the confidence band. They also suggest to consider that there is excessive geographic dispersion if the K-density lies below the lower bound of the confidence band at least once, and never exceeds the upper bound (over the distance range they consider). We compute separately the different components (1.13) and (1.14). When some groups are overrepresented over some distances but underrepresented over others, we do not compute the 'net effect' but consider both over- and underrepresentation separately.

our measures satisfy a number of desirable properties. They are, in particular, invariant to composition and size (**(P2)** and **(P3)**), easy to interpret **(P6)**, can be disaggregated in needed **(P7)**, and allow for statistical testing **(P8)**. When computed from individual data, they also obviate the need for observational equivalence and are naturally independent from boundaries (**(P1)** and **(P5)**). As explained before, when such data is not available, we still need to rely on spatial subdivisions, but the way the measures are computed makes those subdivisions less crucial than for more traditional area-based measures. Last, our measures also intuitively satisfy some form of transfer principle **(P4)**, although a formal proof of the latter is hard to establish because of the kernel smoothing.

As we show in the empirical application that follows, a final substantive advantage is that our measures allow for great flexibility in constructing the benchmarks against which we test deviations of either group x or groups x and y. This flexibility allows us to address a number of questions that are usually hard to deal with, e.g., how to separate segregation by race from segregation by poverty.

Despite their appealing properties, our point-pattern based measures of segregation also have a number of limitations. First, they are 'global measures' in the sense that they depend at each point on the whole distribution of all observations. Hence, these measures have both a local and a global component. This induces potentially a confusion between local agglomeration and global agglomeration. Imagine, e.g., an economy with a large number of widely spaced small clusters. In such an economy, there are many short distances between points (within clusters), but there are even more large distances between points (between clusters). Being a density distribution over all distances, the existence of many larger distances between clusters reduces the relative contribution of the shorter distances within clusters. If we think that the phenomenon under consideration is essentially one that depends on short distances, this poses a potential problem (see, e.g., Mori

et Smith 2015; Billings et Johnson 2016). A direct consequence of that point is that the decomposability of the DO measure to the local level does *not* identify local clusters in a narrow spatial sense. It does identify the locations that are the most exposed to the characteristic of interest we are looking at, where exposure is measured relative to the whole kernel-smoothed distribution of distances. This implies, in particular, the more centrally located points—in a geographical sense—appear more exposed. We need to keep this in mind when interpreting our results.

Second, the DO measures are known to be somewhat sensitive to sample size. For example, when industries are smaller and we use fewer bootstrap replications the "DO test is asymptotically consistent, but systematically upward-biased in small samples." (Barlet *et al.*, 2013, p.349). While we are aware of this problem and acknowledge it, we work with 1,000 replications. Furthermore, we show that for basically all of our samples, the confidence bands of the DO test are very close to the empirical distribution of the counterfactual location universe (which suggests that our results are really 'asymptotic'). Consequently, we do not think that this is a very important issue in our setting. The benefits of the continuous measure and the 'case-control' approach outweigh, in our view, these disadvantages.

1.4 Empirical implementation

1.4.1 Data

Ideally, we would require microdata where each individual can be precisely geolocated. Unfortunately, such data are not available.[18] However, as explained above,

18. Geo-referenced firm-level data are becoming increasingly available. The same is, unfortunately, not true for personal data. For example, the Integrated Public-Use Microdata Series (IPUMS) provides the county of residence as the finest geographic unit, which is way too coarse

our method can be adapted to cope with spatially fine-grained data aggregated at the level of small areas such as census blocks. In what follows, we use publicly available data at the smallest geographic level, i.e., the census block. As we do not observe the within-block distribution, we assume implicitly that everyone is located at the census block centroid. The measurement error induced by this assumption is small and basically random, and becomes even less relevant in the presence of kernel smoothing as explained before.

We use the 2010 Decennial U.S. Census data for the New York Core Based Statistical Area (henceforth, NYCBSA), extracted from the National Historical Geographic Information System (Ruggles *et al.*, 2016). The NYCBSA consists of 25 counties with 240,291 Census blocks in 2010. We drop all blocks with zero population—essentially water blocks, large administrative buildings, and other empty blocks—which leaves us with 178,179 populated blocks. We extract a variety of socioeconomic information at the block and the block-group levels : total population by race (block) ; latitude and longitude coordinates (block) ; aggregate income by race (block group) ; educational attainment by race (block group) ; and a variety of control variables at both levels that we will explain in more detail later in this section. Data at the block group level are broken down to the block level using population weights.

To compute (1.3) and (1.4) requires the population count of each racial group. The U.S. Census provides two major categories : (1) Hispanic or Latino ; and (2) not Hispanic or Latino. Each category is divided into seven subgroups. [19] We

---

for our analysis. Hence, we would need access to the restricted-access confidential census data. The same holds true for other countries such as Canada, where detailed personal information is accessible but at a larger geographic scale than that required for our analysis.

19. White alone ; Black or African American alone ; American Indian and Alaska Native alone ;

aggregate the racial composition of blocks to obtain four major categories : (i) White alone, not Hispanic or Latino ("White" for short) ; (ii) Black or African American alone, not Hispanic or Latino ("Black" for short) ; (iii) Asian alone, not Hispanic or Latino ("Asian" for short) ; and (iv) all of the second major category of the census : Hispanic or Latino ("Hispanic" for short). White is the majority group, followed in order by Hispanic, Black, and Asian. [20]

TABLE 1.1: Racial composition of blocks.

	Average composition (%, all incomes)			
Neighborhood	White	Black	Hispanic	Asian
White	0.74	0.05	0.12	0.07
Black	0.27	0.41	0.24	0.06
Hispanic	0.45	0.14	0.31	0.07
Asian	0.51	0.06	0.13	0.27
	Average composition (%, poor only)			
White	0.66	0.18	0.35	0.41
Black	0.08	0.42	0.19	0.07
Hispanic	0.18	0.34	0.38	0.18
Asian	0.06	0.04	0.06	0.30
NYCBSA	0.65	0.11	0.15	0.07

Notes : We use 2010 census data at the block level. A block is defined as White, Black, Hispanic, or Asian if more than 10% of its population belongs to that race. A block is defined as poor if the total income of the race in the block belongs to the bottom quartile of the city-wide distribution for that race.

Asian alone; Native Hawaiian and Other Pacific Islander alone; Some other race alone; Two or more races.

20. Table 1.3 in Appendix A provides descriptive statistics by race across all income levels (some variables of this table will be explained and used later).

Table 1.1 shows the general pattern of the spatial distribution of races across census blocks. We define a census block as 'White', 'Black', 'Hispanic', or 'Asian' if more than 10% of its population belongs to that race. As shown in the top panel, White blocks are, on average, 74% White, but only 5%, 12%, and 7% Black, Hispanic, and Asian, respectively. Whites also form a majority of 65% in the NYCBSA, which is why we take them as the reference group. Black blocks are, on average, 41% Black compared to 11% Black for the the whole NYCBSA. This figures suggest that there is excess clustering by race, since the average White, Black, Hispanic, and Asian block has a 9%, 30%, 16%, and 20% larger population share compared to the metro area as a whole

As can be seen from the bottom panel of Table 1.1, the patterns change slightly as one focuses on 'poor blocks by race'.[21] In particular, White blocks display less excess clustering, whereas Black blocks, Asian blocks and, especially, Hispanic blocks display more excess clustering. This suggests that poverty and segregation should be disentangled to measure the 'pure effect' of segregation by race on top of segregation by poverty.[22]

We use additional data to construct benchmark distributions, which we will explain below once we have explained the benchmarks themselves.

---

21. We use aggregate income and education by race—reported at the block-group level—and break them down to the block level using population counts by race. Income is defined for persons 15 years and over as incomes received on a regular basis before payments of any type of tax, for the last 12 months prior to the survey. It is reported in inflation-ajusted dollars for the release year. This includes wages, salary, commissions, bonuses or tips from all jobs. It also includes self-employment income from own nonfarm or farm businesses. Aggregate income provides the total income at the block group. Dividing by the racial count provides per capita income. We drop block groups where income is negative (e.g., prisons, other group quarters).

22. Table 1.4 in Appendix A provides descriptive statistics by race for poor only.

1.4.2 Benchmark distributions

The benchmark distribution against which we assess the extent and statistical significance of segregation is of paramount importance. The basic idea is to create counterfactual 'random spatial distributions' from which we assess departures of the empirical distributions.[23] In an ideal world, we would define a benchmark as the distribution that should occur in the absence of any type of sorting (either by income or by race). For example, we could look at the observed distribution of African Americans in New York and compare it to the distribution of the overall population in the NYCBSA. By randomly reassigning all African Americans across the census blocks of New York, we would obtain a counterfactual distribution that would be due to randomness only. Repeating this process a large number of times—say 1,000 times—we can then think about the expected distribution that we would observe if African Americans were located purely randomly, independently of any consideration of sorting along race or income. In a similar manner, we could look at the distribution of the poor African Americans, taking as the benchmark the observed distribution of all African Americans. Repeating the process of random assignment, we would then obtain a counterfactual distribution of the poor *conditional* on the observed distribution of African Americans in New York. Since the benchmark already controls for the observed segregation along racial lines, we can then measure how much more the poor African Americans are concentrated.

One important point of counterfactual distributions is that (despite being counterfactual) they have to be realistic in the sense that the locations considered are ' feasible locations' for the individuals whose geographic concentration we want

23. See, e.g., Klier et McMillen (2008) and Carrillo et Rothbaum (2016) for alternative approaches to constructing counterfactual spatial distributions.

to assess. To take a crude example, one is unlikely to live on Park Avenue in NYC or the Champs Élysées in Paris if one is poor. Hence, blocks on Park Avenue or the Champs Élysées should not be part of the benchmark used for poor people. To illustrate this point with the example from the foregoing paragraph, when thinking about counterfactual distributions of poor African Americans in the NYCBS, we may want to use a benchmark consisting of all 'poor blocks', irrespective of their racial composition. The underlying idea is that the poor can only possibly choose to live in poor blocks (where housing is of lower quality and cheaper, where there are more rentals, etc.), and thus their counterfactual distribution should be computed by considering only the blocks that are potentially feasible to them. We stress this point because existing measures of segregation often propose counterfactuals that are far from being meaningful. Yet, the departure of the actual distribution from a 'realistic counterfactual distribution' is crucial for meaningfully assessing the magnitude of segregation. [24]

As explained above, the definition of the benchmark matters. From a technical perspective, how we compute the counterfactual distributions using that benchmark matters too. In that respect, adapting the methodology by Duranton et Overman (2005) to the measurement of segregation is technically challenging. First, computing K-densities is very time consuming, especially for larger samples. Even when using recent approximation techniques to computing K-densities (Scholl et Brenner, 2015), given our sample sizes this still represents a large computational burden. Second, we need to take into account that populations in census blocks are a priori 'divisible units'. If there are ten Asian people in a block, those ten people should be *independently* reshuffled among the benchmark. This is different

24. A perfectly integrated distribution is hardly ever a realistic benchmark. For example, the benchmark of a perfectly egalitarian distributions—underlying, e.g., the Gini index—is basically meaningless.

from what is done for the case of firms, where the size of the firm (as measured by employment) is taken as indivisible. In other words, whereas the ten Asian in the block are reshuffled independently, for a firm the ten employees would be reshuffled *jointly* as the firm is considered indivisible (Duranton et Overman, 2005). This fundamental difference implies that the permutations need to be done differently in our case. While DO reshuffle all plants in an industry among all locations where we observe plants—keeping the size of the plant constant—we need to reshuffle individual people across the feasible blocks.[25]

---

25. To fix ideas, consider Blacks in the NYCBSA. There are 178,179 census blocks with strictly positive population, and the total population is 19,567,410. There are 85,368 blocks with at least one African American person, and the total African American population is 3,430,080. To measure segregation, we can compute the K-densitues and their cumulative distributions ((1.9)–(1.10) and (1.13)) for African Americans and compare them to the counterfactual distributions that would prevail if Blacks were randomly distributed following the total population. Technically, this requires to randomly permute the 3,430,080 Blacks across all blocks in the NYCBSA, where each block has as many possible locations as there is total observed population there. For example, a block with a total population of 1,000 people in 2010 has 1,000 possible locations for blacks in the counterfactual. In other words, we hold the total population of each block constant (which reflects, e.g., the amount of available housing). In this benchmark, we "expand" our blocks by their observed total population to obtain 19,567,410 locations and permute the 3,430,080 blacks across those locations. Doing so makes sure that : (i) the census geography is respected ; (ii) that there can be no more blacks in each block than there is total population in 2010 in that block ; and (iii) the indivisible unit is the individual, i.e., we do not jointly reshuffle the entire block (which would make little sense). Once this permutation at the individual level is done, we reaggregate up the randomly reshuffled Blacks to the block level and compute the counterfactual K-densities using our weighted epressions. We repeat this procedure 1,000 times and derive the confidence bands from that series. As should be clear, expanding the data to almost 20 million observations, permuting, re-aggregating, and computing the kernel-smoothed K-densities 1,000 times is a computationally very demanding procedure.

Simple benchmarks. Let us first explain what we call 'simple benchmarks'. These benchmarks use the traditional case-control methodology from Duranton et Overman (2005). There are 'cases', e.g., the location of poor Black. We want to assess how concentrated these cases are compared to the 'controls' (e.g., the location of all Black or the location of all poor). We call these benchmarks 'simple' because we can easily observe the controls in our data. By choosing appropriately the controls, we can then assess the deviation of the cases from several different benchmarks, which allows us to explore different dimension and aspects of segregation. The method is fairly flexible and rich in terms of interpretations, which is one of the desirable features of that type of approach.

Predicted benchmarks. In simple benchmarks, each block is drawn with a probability proportional to the number of people located there (either the overall population, or a specific race). This implies that a block with good amenities—e.g., access to parks or public transit—is as likely to be chosen as a block that has no amenities. To control for this aspect, we construct predicted benchmarks. These benchmarks are used to construct counterfactual distributions based on 'locational fundamentals'.

More precisely, we estimate count models that give us a predicted population distribution by race and income level. We cannot use a standard model because of overdispersion.[26] We thus need a model that deals with that overdispersion (an excessive number of zeros). To this end, we estimate a zero-inflated Poisson (ZIP)

26. Conditional on set of explanatory variables X, a Poisson model requires that the conditional mean and variance are equal. This is known as the equidispersion property. This assumption is violated in our data, and a standard Poisson model might not perform well in predicting the observed proportion of zeros. This is due to the fact that not each group or income level is located everywhere.

model :

$$f(y_i) = \begin{cases} f_1(0) + [1 - f_1(0)]f_2(0) & \text{if} \quad y_i = 0 \\ [1 - f_1(0)]f_2(y_i) & \text{if} \quad y_i \geqslant 1 \end{cases} \tag{1.15}$$

where y_i is the dependent population count variable. As can be seen, the ZIP model has two components that correspond to two zero-generating processes. The idea is that the zeros are explained by one process, $f_1(\beta X)$, whereas the rest of the sample is explained by another process, $f_2(\beta Z)$. The first process is governed by a binary distribution that generates structural zeros, whereas the second process, is governed by a standard Poisson distribution that generates counts, some of which may be zero. If $f_1(0) = 0$, the model is only governed by $f_2(\beta Z)$, i.e, the standard Poisson count process. In our application, we parameterize $f_1(0)$ as a binomial probit model with the same set of variables for f_1 and f_2, i.e, $X = Z$. These variables are independent of race and income segregation and used to predict the counts that are related to the 'pure' characteristics of the block. We estimate (1.15) separately for each race , and use it to predict a benchmark population distribution. This gives us counterfactual population sizes for each block, which can be used as new weights to compute a counterfactual distribution (using again random permutations within that benchmark).

Table 1.2 shows our results and the set of explanatory variables that we use to predict the count of each category. These include the count of housing units, total population, and various 'amenities' (distance to nearest subway, landmark, waterfront). We also control for the racial population from the 2000 Census since there is strong persistence in the location choices of the groups ; as well as a measure of centrality, given by the distance to Wall Street. See Tables 1.3 and 1.4 in the appendix for descriptive statistics.

One variable that deserves to mentioned in more detail is that of 'employment opportunities'. To construct the latter, we use NETS data on the exact geographic

TABLE 1.2: Zero-inflated Poisson regressions

Dependent Variable :	All population count			
	Black	White	Hispanic	Asian
Total population	0.001^a	0.001^a	0.002^a	0.001^a
	(3.e-06)	(2.e-06)	(2.e-06)	(4.e-06)
Housing units	-6.e-04^b	9.e-05	-0.002^a	-2.e-04
	(7.e-06)	(3.e-06)	(5.e-06)	(9.e-06)
Average distance to employment opportunities	-0.013^a	0.013^a	-0.028^a	0.042^a
	(3.e-04)	(9.e-05)	(2.e-04)	(6.e-04)
Employment opportunities (1km)	-4.e-06^a	2.e-05^a	-8.e-06^a	6.e-06^a
	(2.e-07)	(1.e-07)	(1.e-07)	(9.e-08) ‘
Employment opportunities (5km)	-7.e-06^a	4.e-06^a	-8.e-07^a	3.e-07
	(3.e-08)	(2.e-08)	(1.e-08)	(2.e-08)
Share of owners	-1.095^a	0.298^a	-1.913^a	-0.541^a
	(0.002)	(0.001)	(0.002)	(0.003)
Number of subway entrances	-0.159^a	0.028	0.032	0.019
	(0.003)	(0.002)	(0.002)	(0.003)
Distance to nearest subway	-2.e-05^a	-2.e-06^b	-2.e-05^a	2.e-05^a
	(3.e-07)	(8.e-08)	(3.e-07)	(4.e-07)
Distance to nearest park	7.e-06	-3.e-05^a	-9.e-05^a	-5.e-06
	(7.e-07)	(2.e-07)	(5.e-07)	(7.e-07)
Distance to waterfront	2.e-04^a	-5.e-05^a	2.e-05^a	8.e-05^a
	(5.e-07)	(3.e-07)	(5.e-07)	(8.e-07)
Distance to CBD (Wall Street)	0.004^b	-0.009^a	0.001	-0.036^a
	(2.e-04)	(5.e-05)	(1.e-04)	(4.e-04)
Racial population in 2000	1.e-04^a	3.e-04^a	4.e-04^a	5.e-04^b
	(4.e-07)	(5.e-07)	(9.e-07)	(2.e-06)
County fixed effects	Yes	Yes	Yes	Yes
Observations	177,492	177,492	177,492	177,492
Number of zeros	92,644	7,346	47,253	87,784
corr$(Y, \widehat{Y})$	0.68	0.26	0.66	0.33

Notes : ^{a}, ^{b}, and ^{c} denote coefficients significant at 1%, 5%, and 10%, respectively. The dependent variable is the count by race for all population at the block level in the 2010 Census. Appendix B provides details and explains also the construction of the ‘employment opportunities’ variables and the geographic controls.

distribution of all establishments in NYCBSA and use MORG data to estimate which types of industries are the top employers of different racial groups, depending on income and on education. Thus, knowing which firms are likely to hire individuals of a particular racial group and income/education level, we can compute a measure of accessibility to employment opportunities, which we include into our regressions. We provide details, as well as summary statistics on top-employers by groups, in Appendix B.

As shown in Table 1.2, the correlations between the predicted and observed counts are reasonably high and for all categories except for White. This is partially due to the lesser degree of zero inflation. Indeed, we have 52% , 27% , and 49% of blocks with zero Black, Hispanic, and Asian, respectively, but only 4% for White.

1.5 Results

Figure 1.11 in Appendix A shows that, unsurprisingly, there is substantial segregation by race in the NYCBSA. The extent of this segregation can be measured more formally by traditional measures that capture the within and between group dimensions—such as the dissimilarity and the exposure/isolation indices—which we compute for Black and White and for different levels of geographic aggregation (block, block group, tract, county, NYC, and NYCBAS) in Table 1.5 in Appendix A. Unsurprisingly, Black are unevenly distributed and are less exposed to White (the majority group) than to themselves. In Hunterdon and Nassau counties, Black are the most unevenly distributed compared to the overall population. Similarly, in Bronx, Kings, and Essex counties, Black are the most isolated from White. Nevertheless, as mentioned previously, these statements require some caution. First, the measures in Table 1.5 carry no information on statistical significance of the observed patterns. Hence—though given the extent of segregation seen from

Figure 1.11 this is highly unlikely—these numbers could be due to random clustering. Second, the measures we provide suffer from the MAUP. Table 1.5 shows that the magnitude of segregation automatically decreases when the geographic units considered are bigger. Last, as explained before, there is no geographical dimension in those measures. Any random permutation of populations across geographic units gives the same value for the segregation indices because the relative position of the units does not matter.

To deal with these problems, we now turn to our point pattern-based measures and discard areal units. We firstly present our baseline results for **E–C** and **E–I** by race and poverty status using as our benchmarks the 'simple' observed distributions. We secondly disentangle the effect of race from the effect of poverty and vice versa. We finally show that our results are robust to the use of more sophisticated predicted benchmarks. [27]

1.5.1 Evenness-clustering by race : simple benchmarks

We compute the K-density PDFs, CDFs, and excess segregation measures that capture departures of the observed distributions from the benchmarks. For the **E–C** dimension, we discard the majority group (White) in our analysis and focus only on Black, Hispanic, and Asian. We use White as the majority group for the **E–I** dimension and report exposure to and isolation from White for the other groups.

---

27. We have four races (White, Black, Asian, and Hispanic), two dimensions (**E–C** and **E–I**), and a several different benchmarks (overall racial distribution, poverty status, and estimated benchmarks). Providing results for all cases is infeasible—because computation times are substantial—and would result in too much information. We hence restrict ourselves to sets of results that we think illustrate the key aspects of our approach and that are relevant to both researcher and policy makers. Also, we relegate results where we decompose the K-density to measure the exposure of individual blocks to segregation to Appendix D.

We also compute results of exposure-isolation between non-majority groups, for example, exposure-isolation between Black and Hispanic.

We start by looking at the **E–C** dimension for Black, Hispanic, and Asian, taking as the benchmark the observed distribution of the overall population across census blocks in the NYCBSA. Figure 1.1 shows that Black, Hispanic, and Asian (red line) are segregated compared to our baseline benchmark (the dashed black confidence band). This holds true for distances below about 30 kilometers, as shown in panel (b) which depicts the excess clustering of the groups measured as the difference between the observed distribution and the upper bound of the confidence band. In words, the observed distribution is significantly less evenly distributed than a distribution generated by randomly permuting racial populations across all possible locations in the NYCBSA, holding constant the geography and the overall population distribution.

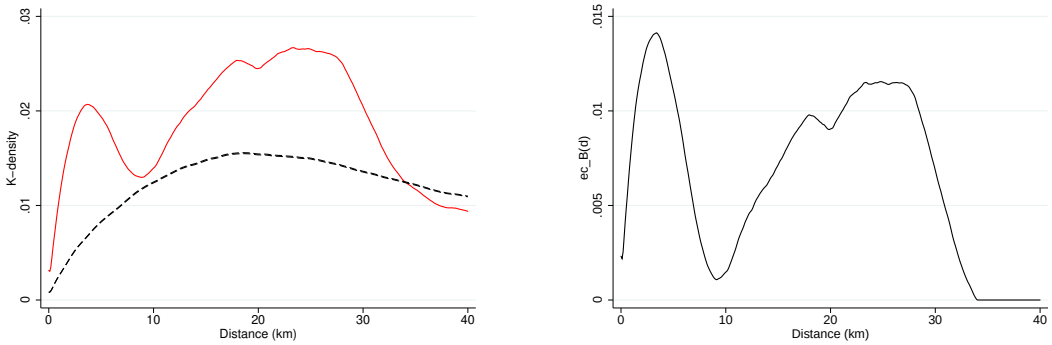
Figure 1.2 depicts simultaneously the K-density PDFs for Black, Hispanic, Asian, and the overall population. Two results are worth noting. First, as shown, the three groups are more clustered than the overall population.[28] Second, the observed distribution of overall population (in blue) looks very similar to the confidence bands in panel (a) of Figure 1.1. This suggests that with large samples, using the observed distribution of overall population as the benchmark provides results that are virtually identical to those obtained using 1,000 random permutations of the racial groups across all possible locations. We show in Appendix C, using numerical simulations for a 'toy city', that this need however not be the case when samples are smaller and when spatial units are unevenly spaced and/or of uneven

28. Black have the highest excess clustering at short distances, with almost 0.014 at 5 kilometers distance. This is followed by roughly the same share of 0.07 and 0.06 for Hispanic and Asian at that same distance. Hence, Black are about twice as clustered at 5 kilometers than the two other groups.

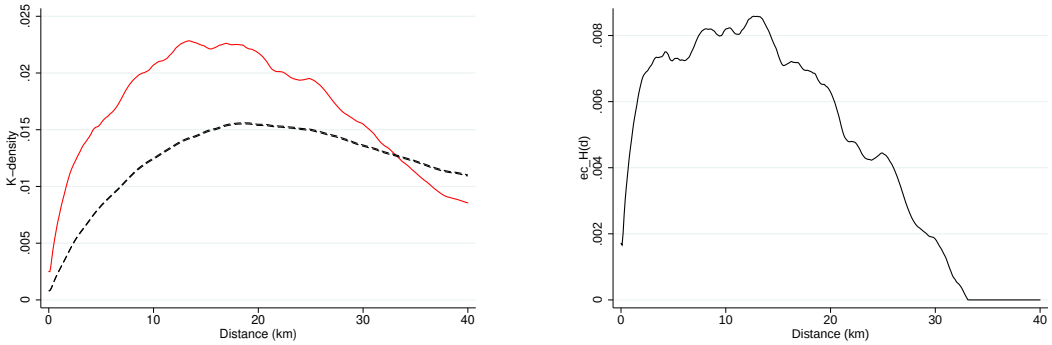
FIGURE 1.1: Evenness-clustering for Black, Hispanic, and Asian.

(a) PDF and confidence band. (b) Excess concentration.

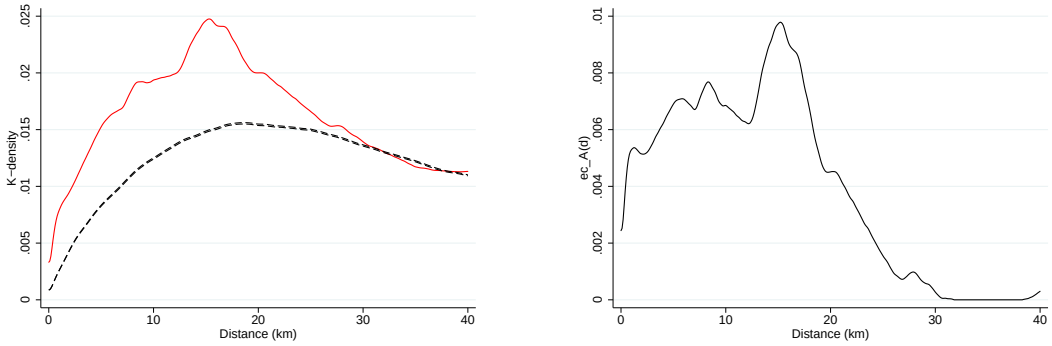
[1] Black.



[2] Hispanic.

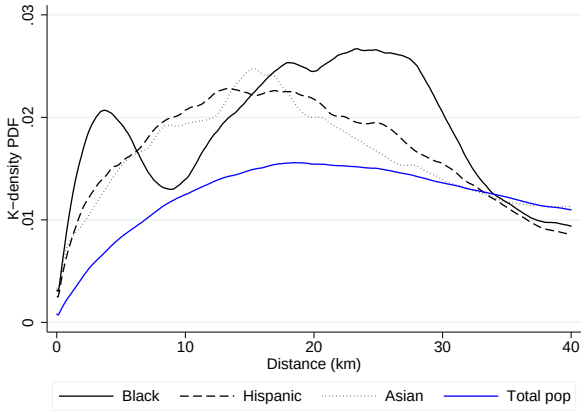


[3] Asian.



Notes : Panel (a) depicts the K-density PDFs (red lines) and the confidence bands for a counterfactual distribution based on the benchmark of the overall population distribution (dashed lines) for each distance d. Panel (b) depicts the distribution in excess of the upper bound of the confidence band. All confidence bands are the 90% confidence interval computed from 1,000 replications using random permutations of the racial population across the benchmark.

FIGURE 1.2: K-density PDFs for Black, Hispanic, Asian, and total population.

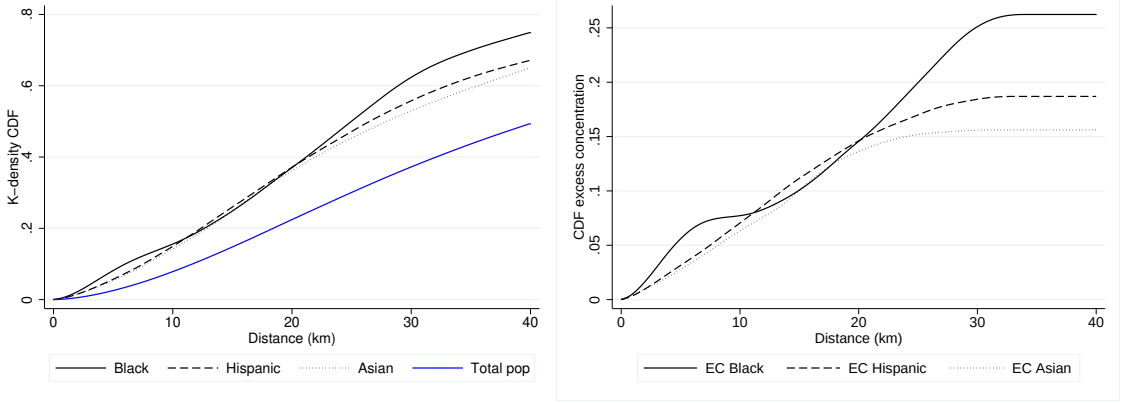


size. In that case, random permutations may deliver counterfactual distributions that look somewhat different from the observed overall population distribution.

We next compute the K-density CDFs, which provide measures of the magnitude of segregation that can be compared to more traditional indices (Reardon et O'Sullivan, 2004). Figure 1.3 shows the CDFs for the three groups and for the total population. For distances below 10 kilometers, we find that the previous rankings still hold. Black is the most segregated group, followed by Hispanic and Asian. To understand how to interpret panel (a) of Figure 1.3, consider the case of Black at a distance of 10 kilometers. The CDF is 0.159 at that distance. This means that if we draw at random two Black in the NYCBSA, there is a probability of 15.9% that the two live at a distance of less than 10 kilometers from each other. This is much larger than the corresponding probability for two New Yorkers drawn at random from the overall population, which would have only about a 7.86% probability to live less than 10 kilometers from each other. Hence, Black are geographically more concentrated than the benchmark, and the excess concentration is 7.72% at 10 kilometers distance, as shown in panel (b) of Figure 1.3. Table 1.6 in Appendix A provides the exact numbers for the CDFs and their excess compared to the benchmark for our three groups of interest.

FIGURE 1.3: K-density CDFs and excess concentration.

(a) CDFs. (b) Excess concentration.

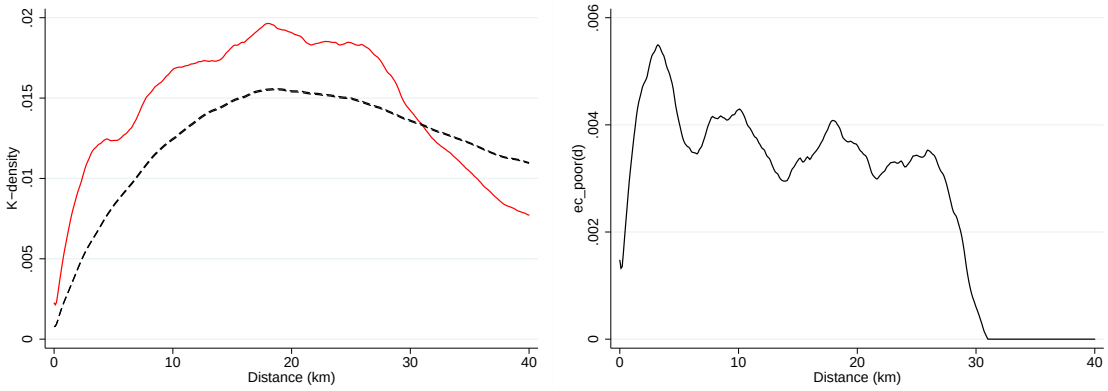


1.5.2 Evenness-clustering by poverty : simple benchmarks

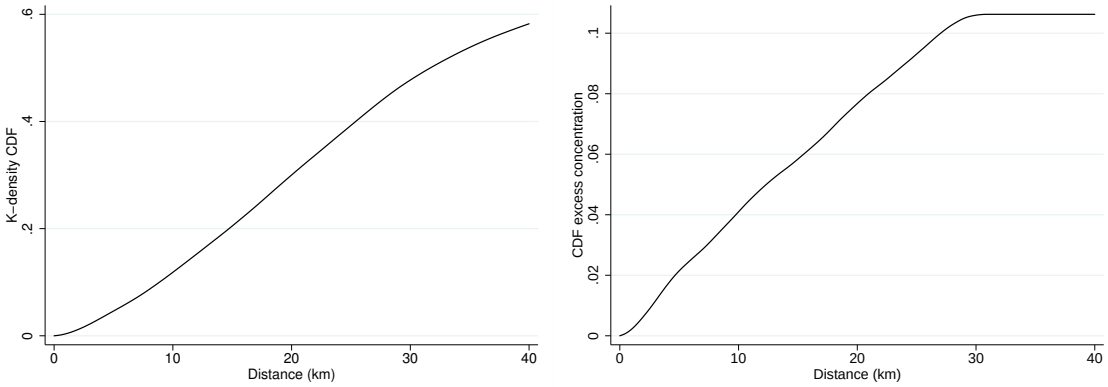
We next explore how clustered are the poor. To this end, we define poor blocks as blocks belonging to the bottom 10% of the income distribution in the NYCBSA. Panel (a) of Figure 1.4 shows that there is clustering of poverty for all short distances across all racial groups. The black dashed lines correspond to the confidence band derived from 1,000 counterfactuals using the overall population as the benchmark. It corresponds to what we should observe if there was no sorting along income, whereas the red line is the observed distribution of poverty. The gap between the former and the latter provides a measure of the magnitude of sorting along income (or income segregation). The magnitude of that gap is depicted in panel (b). At 10 kilometers distance, there are 4.1% more bilateral distances among poor in the NYCBSA than among the population in general. This shows that poverty is geographically concentrated, but less than race : the excess concentration of Black, Hispanic, and Asian at 10 kilometers distance is 7.72%, 7.04%, and 6.32%, respectively (see Table 1.6).

FIGURE 1.4: Evenness and excess concentration of poverty.

(a) K-density PDF and excess concentration.



(b) K-density CDF and excess concentration.



Poverty conditional on race

Since the census provides income by race, we can look at segregation by poverty conditional on race. Doing so allows us to partly disentangle the geographic concentration of poverty from the geographic concentration of race. To do so, we first keep all locations with at least one individual of group x (e.g., Black). We refer to this as the location universe of group x or the 'control group'. Then, within each group, we keep the subsample of the bottom quartile of the blocks in group x's city-wide income distribution. We refer to this as the 'case group'.[29] We then compute the observed distribution of cases (i.e., the distribution of poor individuals of a given group) and create 1,000 counterfactual distributions where we randomly reshuffle the cases among the controls. This allows to gauge how much more (or less) poor members of group x are concentrated in space compared to group x in general. Observe that by doing so we already control for racial segregation since the benchmark is the geographic distribution of group x. Hence, this allows to answer the question of how much more the poor of group x are segregated conditional on the geographic distribution of group x.

As Figure 1.5 shows, the poor of each race are substantially more clustered than the race in general. For all three groups, there are clear patterns of income segregation conditional on racial segregation. Observe that our previous ranking does not hold anymore. Indeed, the geographic concentration of poverty is more pronounced for Asian than for Black or Hispanic. In other words, once segregation by race is controlled for, Asian is the most segregated group by income, more so than Black or Hispanic.

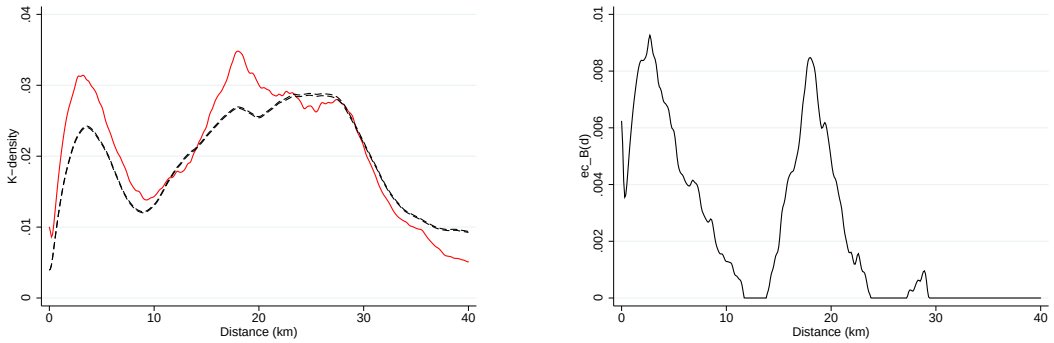
---

29. We also run the estimations by defining 'poor' as the bottom 10 % in the city-wide income distribution. This yields smaller samples.

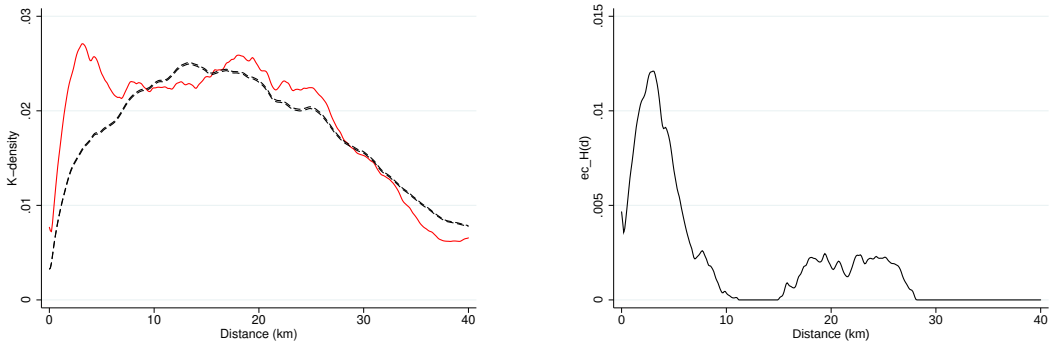
FIGURE 1.5: Evenness-clustering of poverty, conditional on clustering by race.

(a) PDF and confidence band. (b) Excess concentration.

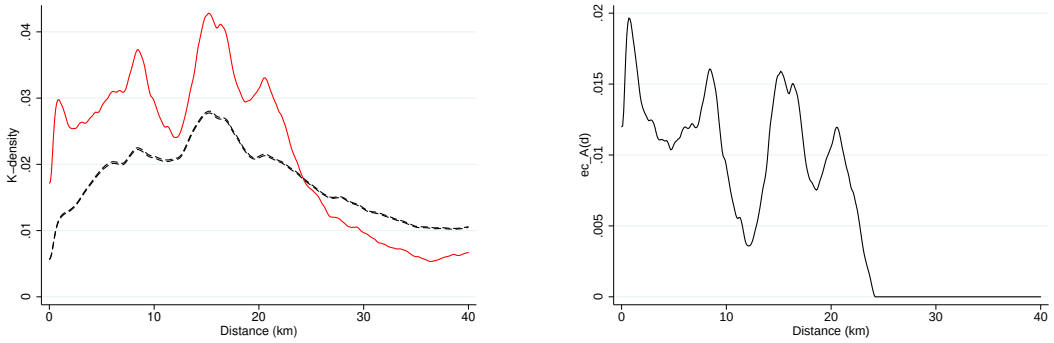
[1] Black.



[2] Hispanic.



[3] Asian.



Notes : Panel (a) depicts the K-density PDFs (red lines) and the confidence bands for a counterfactual distribution based on the benchmark of the overall population distribution (dashed lines) for each distance d. Panel (b) depicts the distribution in excess of the upper bound of the confidence band. All confidence bands are the 90% confidence interval computed from 1,000 replications using random permutations of the racial population across the benchmark.'Poor' are defined as the census blocks in the bottom quartile of the city-wide income distribution of the race under consideration.

Race conditional on poverty

The exercise in the foregoing subsection can be 'inverted' to look at the effects of race conditional on poverty, which provides a complementary angle to look at the question of segregation. As shown before, for large samples there is virtually no difference between using the confidence bands derived from 1,000 counterfactual distributions and the empirical distribution of the benchmark. Here, we hence take the observed distribution of the poor blocks (bottom quartile) as our benchmark, and look at how much more or less the poor of each race are segregated on top of poverty.

FIGURE 1.6: Evenness and excess concentration, conditional on poverty.

(a) PDFs. (b) Excess concentration (CDF).

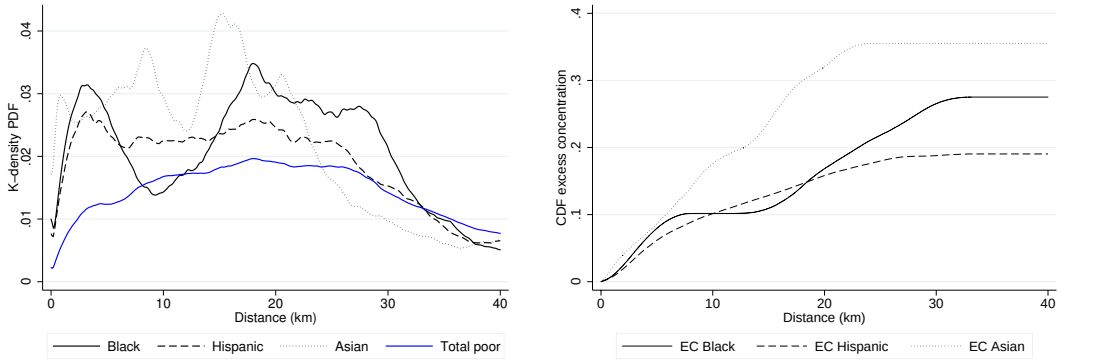


Figure 1.6 depicts the PDFs for the poor of the three groups as well as the observed overall concentration of the poor (across all races). While this figure looks similar to Figure 1.5, its interpretation is different. In this case, the underlying benchmark is the overall distribution of the poor and not the overall distribution of the race. Hence, what we measure is how much more or less a poor racial group is concentrated than poverty in general, which allows to answer the question of how much more the poor of group x are segregated than poverty itself.[30] As shown,

30. Since we measure the observed distribution of poor in group x in both Figures 1.5 and

conditional on the distribution of poverty, the poor Black, Hispanic, and Asian experience more geographic concentration. In other words, race pushes people to locate closer from one another, even when controlling for segregation by income. Furthermore, as shown by panel (b) of Figure 1.6, Asian poor are strikingly more clustered than the poor of the two other groups, more than twice at about 10 kilometers distance. Hence, segregation along racial lines seems especially prevalent among Asian.

Table 1.7 in Appendix A summarizes the magnitude of segregation by race on top of segregation by income. As shown, whereas 1.6% of bilateral distances between residents of poor blocks are below 2.5 kilometers in the NYCBSA, the corresponding number for poor Black is 5.1%. Hence, poor Black are excessively clustered with 3.4 percentage points more of bilateral distances below 2.5 kilometers. The corresponding figures are an excess clustering of 2.7 percentage points for Hispanic, and 4.9 percentage points for Asian. As Table 1.7 shows, Asian poor are particularly more clustered than poor in general, followed by Black and Hispanic poor with similar degrees of excess clustering.

1.5.3 Exposure-isolation : simple benchmark

We now turn to intergroup contacts, i.e., we explore the between-group dimension of segregation. To this end, we select two groups, say Black and White and compute the observed distribution of bilateral distances between all pairs from the two groups. We then compare this to a counterfactual benchmark where we randomly permute the populations of the two groups among all locations with

---

1.6, the depicted PDFs are the same. However, the underlying benchmarks differ, and so do the measures of excess segregation (which are computed for each different benchmark).

residents of either of the two groups.[31] In words, this means that we measure the exposure to (or isolation from) members of one group to members of the other group, conditional on the overall spatial distribution of the two groups.[32]

Figure 1.7 depicts the results for all pairs of groups. As shown, all groups are isolated from each other in the NYCBSA at short distances—all observed distributions are below the confidence bands of the counterfactual random allocations, meaning that groups are more dispersed from one another than what a random distribution would predict (conditional on group sizes and the geographic distribution of the two groups). As further shown in Figure 1.7, the magnitudes of isolation differ across pairs of groups. While White-Asian (panel (f)) and Black-Hispanic (panel (b)) are only slightly isolated from each other at very short distances (see panel (f)), Black-Asian (panel (c)) and Hispanic-Asian (panel (e)) are strongly isolated from each other. Also, Black-White and Hispanic-White are isolated, though quantiatively a bit less strongly than Asian.

1.5.4 Exposure-isolation : poverty conditional on race

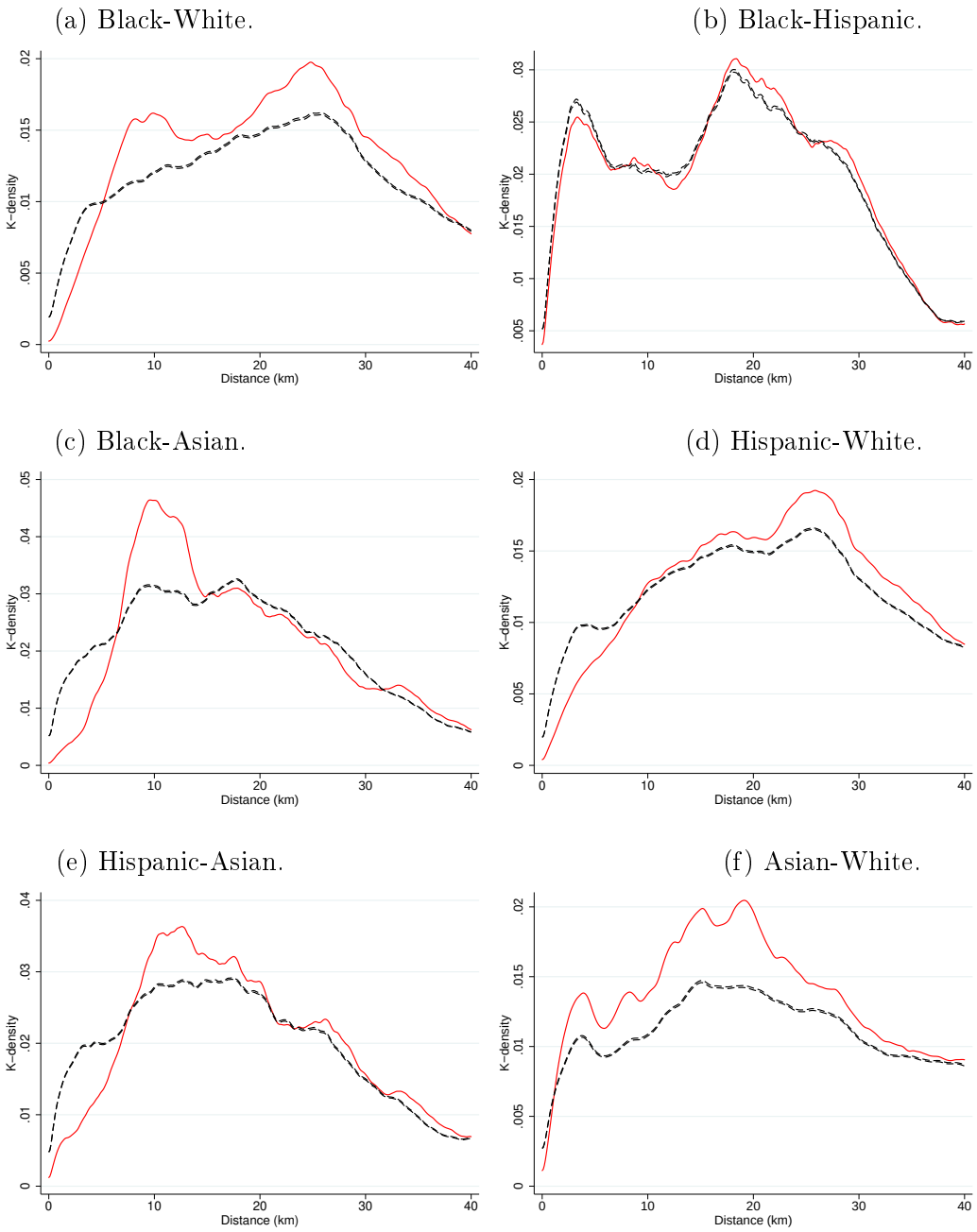
To limit the number of cases, we compute and discuss only the exposure of poor of groups x to poor of group y conditional on the geographic distribution of groups x and y. In other words, we focus on the effects of poverty conditional on race,

---

31. The controls are hence all census blocks with at least one resident of either group x or group y. We also did the computations using a stricter criterion requiring at least 10% of the block's population to belong to either x or y. The results are fairly similar and available upon request.

32. Two groups can be strongly concentrated geographically but not much exposed to each other. Hence, not controlling for the spatial distribution of the groups will make strongly clustered groups appear exposed to each other though they might be very segregated.

FIGURE 1.7: Exposure-isolation for all pairs of groups.



Notes : The figure depicts the K-density PDFs (red lines) and the confidence bands (dashed lines) for a counterfactual distribution based on the benchmark of the population distribution of the two groups. All confidence bands are the 90% confidence interval computed from 1,000 replications using random permutations of the racial population of the two groups across the benchmark.

which we think is the most interesting case.

Figure 1.8 shows that, even when controlling for racial segregation, poverty is spatially stratified across racial lines for all the six pairs. For short distances, the observed distribution of bilateral distances between poor members of two different groups lies below the confidence bands, meaning that pairs of poor from different groups are less exposed to each other than if they were randomly reallocated across all blocks with members of either group. Hence, there is a sizeable amount of segregation by income even conditional on segregation by race. This effect is more pronounced for poor Asian. They tend be far from other groups, especially Black and Hispanic. Table 1.8 in Appendix A shows that for distances below about 5 kilometers, the strongest isolation is for the pair Black-Asian, with 5.22 percentage points less exposure than predicted by a random allocation.

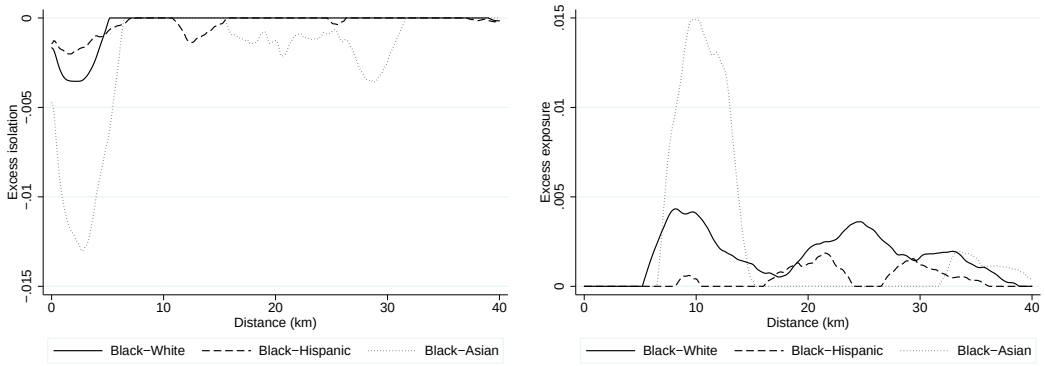
1.5.5 Predicted benchmarks

We finally provide some results using more complicated 'predicted benchmarks'. Recall that in the previous sections, we only used 'simple benchmarks' based on observed distributions of populations with certain characteristics (race, income). There was no 'weighting' of any sort taking into account that different populations may value different block-level characteristics differently. As explained in Section 1.4.2, we can define more complicated benchmarks that use a counterfactual population distribution predicted using a number of observable block-level characteristics. The idea is that, as in Ellison *et al.* (2010) or Klier et McMillen (2008), we want to predict a counterfactual distribution of population based on the locations effectively chosen by that population as a function of locational fundamentals. By doing so, we generate a predicted distribution that will serve as the benchmark.

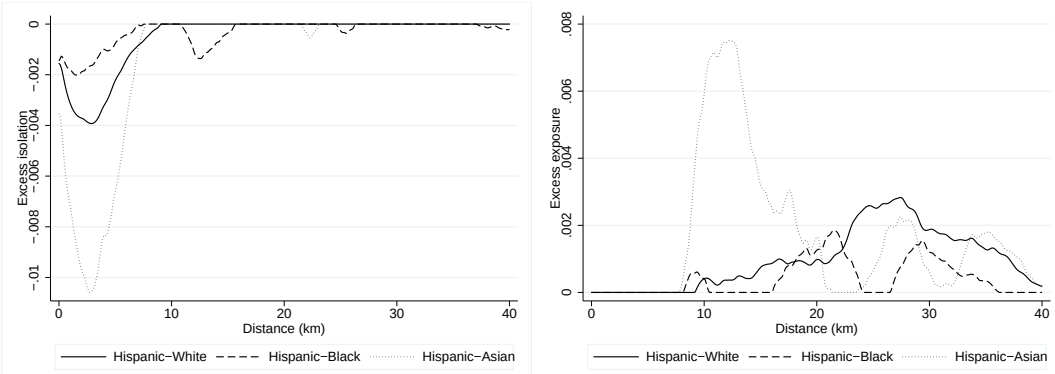
FIGURE 1.8: PDF of excess **E–I** for poor (conditional on racial segregation).

(a) Excess isolation. (b) Excess exposure.

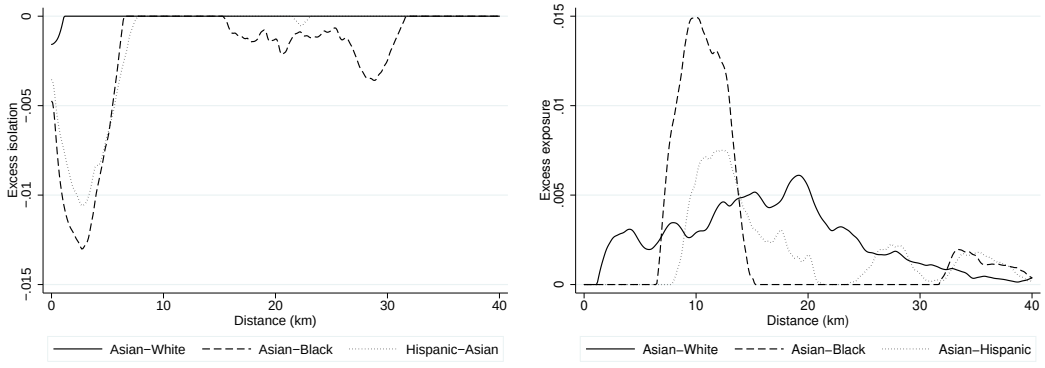
[1] Black.



[2] Hispanic.



[3] Asian.



Notes : Panel (a) depicts excess isolation, whereas panel (b) depicts excess exposure.

For the ease of comparison with our previous results—and since the computations are very time-consuming—we again focus only on the three minority groups (Black, Hispanic, and Asian) and explore the **E**–**C** dimension. Figure 1.9 shows the results for Black (panel [1]), Hispanic (panel [2]), and Asian (panel [3]), respectively. As shown, the magnitude of segregation depends to some extent on the benchmark against which we compare the observed distribution. More precisely, the confidence bands based on the predicted benchmark (blue dashed lines) lie in general below the confidence bands based on the simple benchmark (black dashed lines) for all three groups. Hence, we potentially underestimate segregation when using simple benchmarks compared to the baseline benchmark. These differences suggest that based on block-level characteristics, the benchmark distribution of the groups should be more dispersed than what it would be using the observed distribution, which implies that we slightly underestimate the amount of segregation on top of those characteristics.

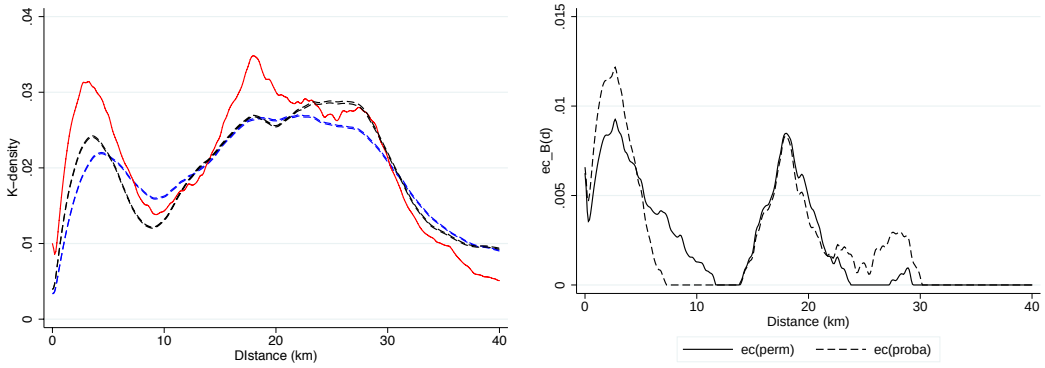
1.6 Appendix

A. Additional tables, figures, and results

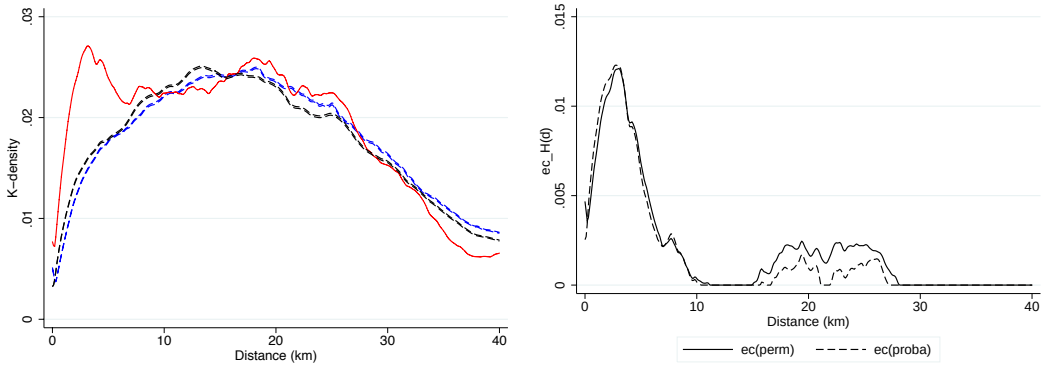
FIGURE 1.9: Comparing simple and estimated benchmarks, poverty conditional on race.

(a) PDFs and confidence bands. (b) Excess clustering.

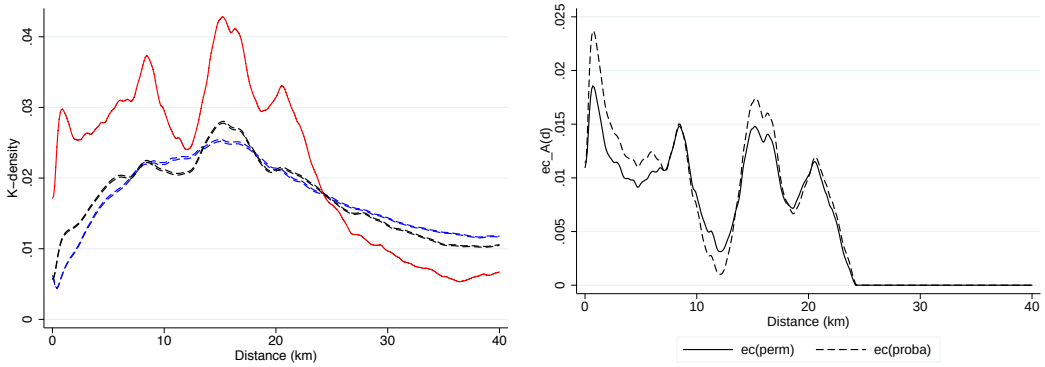
[1] Black.



[2] Hispanic.



[3] Asian.



Notes : Panel (a) depicts the PDFs (in red), as well as the confidence bands generated using simple benchmarks (random permutations ; in black) and estimated benchmarks (counterfactual distributions, based on block-level characteristics ; in blue). Panel (b) depicts excess exposure for the two cases.

TABLE 1.3: Descriptive statistics, NYCBSA 2010 (block level, all income).

	White		Black		Hispanic		Asian		All	
Variables	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Population count	61	95	72	135	53	108	26	64	109	179
Income per capita	45,602	39,391	40,575	168,781	36,218	135,011	63,787	161,195	38,448	21,713
Income share	0.61	0.22	0.24	0.20	0.21	0.17	0.20	0.15	–	–
Owner share	0.84	0.21	0.42	0.33	0.17	0.16	0.22	0.19	0.72	0.31
Housing units	39	74	62	100	55	88	55	98	43	80
Land occupied (square km)	0.11	0.51	0.04	0.23	0.05	0.28	0.06	0.22	0.10	0.48
Water occupied (square km)	0.0005	0.0085	0.0001	0.0036	0.0002	0.0074	0.0002	0.0046	0 .0005	0.008
Average distance to employment	0.05	0.02	0.04	0.02	0.05	0.02	0.04	0.01	0.05	0.02
Average distance to employment (weighted)	0.05	0.02	0.04	0.02	0.05	0.02	0.04	0.01	0.05	0.02
Employment opportunities (1km)	272	1901	796	2091	993	3799	1019	6753	–	–
Employment opportunities (2-5km)	1416	7188	4424	7477	5253	13416	4997	21739	–	–
Employment opportunities (5km)	5089	16240	15768	21507	21691	42881	16612	43431	–	–
Firm opportunities (1km)	36	139	38	58	124	240	93	435	–	–
Firm opportunities (2-5km)	183	549	205	245	628	903	442	1364	–	–
Firm opportunities (5km)	632	1349	709	759	2332	3010	1441	2938	–	–
Number of subway entrances	0.01	0.11	0.02	0.15	0.01	0.14	0.02	0.17	0.01	0.12
Distance to nearest subway	5.01	7.13	2.52	4.29	3.02	4.91	2.80	3.77	4.58	6.84
Distance to nearest landmark	2.03	2.25	1.31	1.50	1.55	1.82	1.65	1.83	1.91	2.17
Distance to nearest waterfront	1.08	1.10	1.39	1.17	1.23	1.16	1.35	1.12	1.14	1.13
Distance to CBD (Wall Street)	50.51	31.57	36.19	26.78	40.19	29.64	33.01	21.31	47.63	31.43
Total Population (Millions)	9.70		3.43		4.42		1.89		19.5	
Population share	0.49		0.17		0.22		0.09		–	
Income share	0.56		0.13		0.15		0.14		–	
Number of blocks (non-zero pop.)	156,198		43,433		76,336		40,198		178,179	

Notes : Some of the cross variables by race are available in the Census. For others, we define a variable-race in a block where there is at least 10 % of that race. We keep only blocks with at least one resident. All distances are given in kilometers. See Appendix B for a description of how we construct the 'employment opportunities' and 'firm opportunities' variables, as well as the block-level geographic controls.

TABLE 1.4: Descriptive statistics, NYCBSA 2010 (block level, poor only).

	White		Black		Hispanic		Asian		All	
Variables	mean	std	mean	std	mean	std	mean	std	mean	std
Population count	64.52	105.37	105.50	181.00	84.03	150.22	65.51	114.71	181.39	254.20
Income per capita	23709	5882	10560	3843	8248	2969	13946	5576	941113	3098502
Income share	0.65	0.27	0.32	0.29	0.20	0.22	0.16	0.16	–	–
Owner share	0.73	0.27	0.37	0.31	0.20	0.18	0.27	0.21	0.49	0.33
Housing units	44.45	71.41	87.55	121.26	69.92	100.55	73.63	99.54	65.24	94.64
Land occupied (square km)	0.11	0.61	0.04	0.19	0.05	0.27	0.04	0.13	0.06	0.42
Water occupied (square km)	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01
Average distance to employment	0.05	0.02	0.04	0.02	0.05	0.02	0.04	0.01	0.05	0.02
Average distance to employment (weighted)	0.05	0.02	0.04	0.02	0.04	0.02	0.04	0.01	0.04	0.02
Employment opportunities (1km)	205	373	962	1378	609	846	373	1402	–	–
Employment opportunities (2-5km)	1132	1530	5249	7567	3234	3785	2063	4779	–	–
Employment opportunities (5km)	4395	6224	19245	26384	11908	13693	8234	15061	–	–
Firm opportunities (1km)	29	40	130	150	95	89	47	89	–	–
Firm opportunities (2-5km)	148	188	703	827	476	425	249	342	–	–
Firm opportunities (5km)	543	728	2444	2740	1669	1531	880	1051	–	–
Number of subway entrances	0.01	0.10	0.02	0.17	0.02	0.15	0.02	0.18	0.02	0.15
Distance to nearest subway	6.29	9.12	1.92	4.00	2.59	4.62	2.01	2.91	3.43	6.62
Distance to nearest landmark	1.98	2.28	1.13	1.26	1.32	1.59	1.36	1.64	1.43	1.78
Distance to nearest waterfront	1.16	1.15	1.37	1.16	1.22	1.09	1.47	1.11	1.34	1.16
Distance to CBD (Wall Street)	52.35	36.03	33.10	27.67	38.41	30.48	25.04	17.24	38.47	32.42
Total Population (Millions)	2.34		1.12		1.61		0.68		8.07	
Population share	0.24		0.32		0.36		0.35		0.41	
Income share [1]	0.35		0.08		0.09		0.06		0.08	
Number of blocks	36353		10650		19188		10353		44515	

Notes : Some of the cross variables by race are available in the Census. For others, we define a variable-race in a block where there is at least 10 % of that race. We keep only blocks with at least one resident. All distances are given in kilometers. See Appendix B for a description of how we construct the 'employment opportunities' and 'firm opportunities' variables, as well as the block-level geographic controls.[1]share of poor income within each race.

TABLE 1.5: Dissimilarity and exposure indices, NYCBSA, 2010 Census.

		Block			Block group			Tract		
County	State	Even	Expo	units	Even	Expo	units	Even	Expo	units
Essex	NJ	0.43	0.12	8751	0.41	0.13	671	0.40	0.14	210
Hudson	NJ	0.50	0.33	4286	0.48	0.36	445	0.46	0.39	166
Hunterdon	NJ	0.67	0.67	3512	0.45	0.86	79	0.41	0.92	26
Middlesex	NJ	0.41	0.56	12730	0.30	0.65	523	0.27	0.69	175
Monmouth	NJ	0.63	0.47	14943	0.55	0.56	469	0.52	0.59	144
Morris	NJ	0.57	0.75	9411	0.38	0.84	295	0.35	0.86	100
Ocean	NJ	0.61	0.76	16067	0.43	0.88	378	0.38	0.92	126
Passaic	NJ	0.52	0.25	6736	0.48	0.28	365	0.46	0.29	100
Somerset	NJ	0.54	0.53	5595	0.48	0.59	181	0.48	0.62	68
Sussex	NJ	0.64	0.88	4475	0.30	0.96	108	0.25	0.97	41
Union	NJ	0.46	0.28	7563	0.42	0.32	417	0.40	0.35	108
Dutchess	NY	0.51	0.57	6548	0.44	0.64	248	0.42	0.67	79
Nassau	NY	0.67	0.28	22127	0.64	0.32	1143	0.62	0.35	284
Orange	NY	0.48	0.58	10034	0.38	0.66	276	0.34	0.70	79
Putnam	NY	0.52	0.89	2423	0.22	0.96	69	0.16	0.97	19
Rockland	NY	0.55	0.41	4764	0.48	0.50	203	0.45	0.54	65
Suffolk	NY	0.61	0.45	28892	0.52	0.54	999	0.49	0.58	323
Westchester	NY	0.53	0.35	15170	0.47	0.41	704	0.46	0.44	223
Pike	PA	0.64	0.67	3736	0.49	0.78	43	0.48	0.80	18
New York	NY	0.50	0.29	3950	0.49	0.30	1170	0.47	0.33	288
Bronx	NY	0.24	0.09	5498	0.23	0.09	1154	0.21	0.10	339
Kings	NY	0.49	0.12	9764	0.49	0.12	2085	0.48	0.13	761
Queens	NY	0.59	0.15	14858	0.58	0.16	1746	0.57	0.17	669
Richmond	NY	0.62	0.36	5078	0.59	0.41	339	0.57	0.44	111
Five boroughs	NYC	0.49	0.14	39148	0.48	0.15	6494	0.48	0.16	2168
All	NYCBSA	0.56	0.21	240291	0.54	0.24	14901	0.52	0.26	4701

Notes : Dissimilarity is computed for Black and exposure is computed for Black and White. The formulas are those in Massey et Denton (1988) : Diss $= \sum_{i=1}^{n} [t_i \mid p_i - P \mid /2TP(1-P)]$ and Expo $= \sum_{i=1}^{n} [x_i/X]\,[y_i/t_i]$, where t_i and p_i are total population and Black proportion of the area i, T and P are the total population and Black proportion for each study area (county, NYC, NYCBSA), x_i and y_i are the total Black and White populations of area i, and X is the total of Black population.

TABLE 1.6: CDFs and excess **E–C** at various distances (all).

	Black			Hispanic			Asian			All population
Distance (km)	CDF	EC	ED	CDF	EC	ED	CDF	EC	ED	CDF
0.2	.0010	.0007	0	.0008	.0005	0	.0011	.0008	0	.0002
0.5	.0028	.0020	0	.0022	.0015	0	.0027	.0020	0	.0007
1.0	.0075	.0056	0	.0056	.0037	0	.0065	.0045	0	.0018
2.5	.0311	.0231	0	.0213	.0132	0	.0205	.0124	0	.0080
5.0	.0814	.0559	0	.0570	.0314	0	.0533	.0277	0	.0254
10	.1559	.0772	0	.1493	.0704	0	.1422	.0632	0	.0786

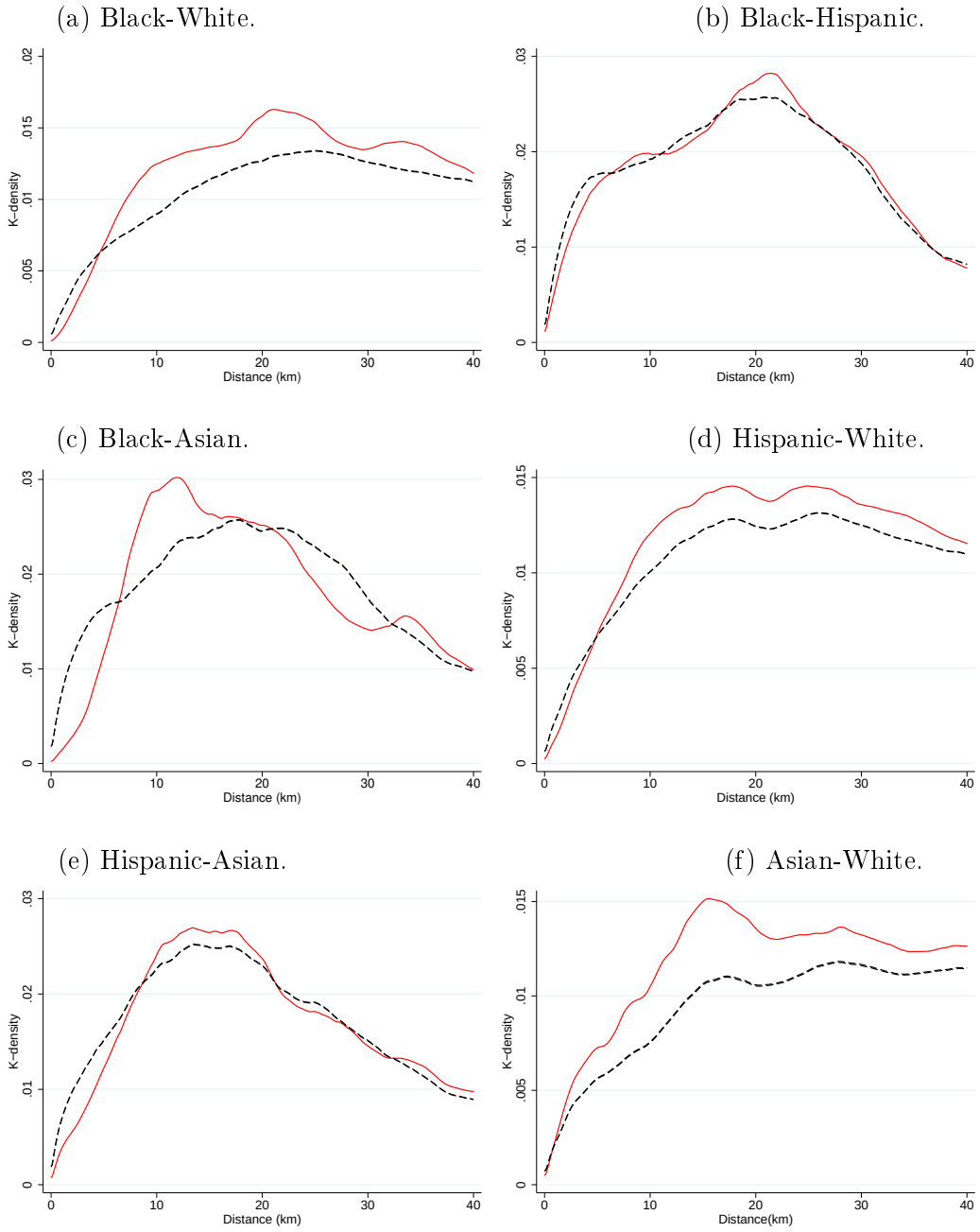
Notes : EC and ED denote excess clustering and excess dispersion, respectively. We report the measures for each racial group, and the benchmark is 1,000 random permutations of each racial group across the whole population distribution of the NYCBSA.

TABLE 1.7: CDFs and excess **E–C** at various distances (poor).

	Black			Hispanic			Asian			All poor
Distance (km)	CDF	EC	ED	CDF	EC	ED	CDF	EC	ED	CDF
0.2	.0019	.0014	0	.0015	.0010	0	.0035	.0031	0	.0004
0.5	.0048	.0034	0	.0041	.0028	0	.0104	.0091	0	.0014
1.0	.0124	.0086	0	.0109	.0070	0	.0249	.0211	0	.0038
2.5	.0507	.0342	0	.0431	.0267	0	.0651	.0487	0	.0164
5.0	.1257	.0795	0	.1078	.0617	0	.1324	.0862	0	.0461
10	.2171	.1017	-.0031	.2199	.1014	0	.2944	.1759	0	.1185

Notes : EC and ED denote excess clustering and excess dispersion, respectively. We report the measures for the poor (bottom quartile of the race-specific income distribution) of each racial group, and the benchmark the observed distribution of poor (bottom quartile of the overall income distribution) across the NYCBSA.

FIGURE 1.10: Exposure-isolation by race, conditional on the geographic distribution of poverty.



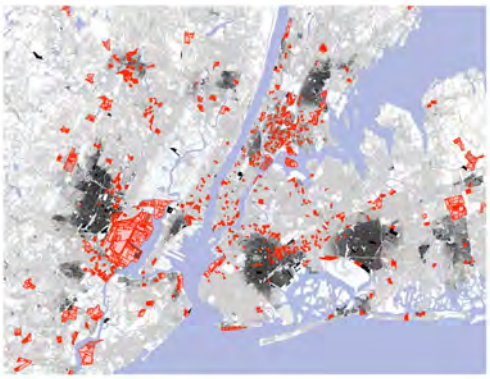
Notes : The figure depicts the K-density PDFs (red lines) and the confidence bands (dashed lines) for a counterfactual distribution based on the benchmark of the population distribution of poor. All confidence bands are the 90% confidence interval computed from 1,000 replications using random permutations of the racial population of the two groups across the benchmark.

TABLE 1.8: CDFs and excess exposure-isolation at various distances for poor.

	Black-White		Black-Hispanic		Black-Asian	
Distance (km)	EI	EE	EI	EE	EI	EE
.2	-.0005	0	-.0004	0	-.0015	0
.5	-.0012	0	-.0008	0	-.0036	0
1	-.0027	0	-.0017	0	-.0084	0
2.5	-.0079	0	-.0045	0	-.0263	0
5	-.0135	0	-.0075	0	-.0522	0
10	-.0135	.0155	-.0083	.0008	-.0567	.0349
	Hispanic-White		Hispanic-Asian		Asian-White	
Distance (km)	EI	EE	EI	EE	EI	EE
.2	-.0005	0	-.0011	0	-.0004	0
.5	-.0011	0	-.0027	0	-.0008	0
1	-.0026	0	-.0061	0	-.0012	0
2.5	-.0080	0	-.0199	0	-.0012	.0023
5	-.0162	0	-.0420	0	-.0012	.0093
10	-.0198	.0002	-.0494	.0066	-.0012	.0229

Notes : EE and EI denote excess exposure and excess isolation, respectively. We report the measures for each pair of racial groups, and the benchmark is 1,000 random permutations of both racial groups across all the blocks of the NYCBSA with positive population in either group.

(a) African American ('Black')

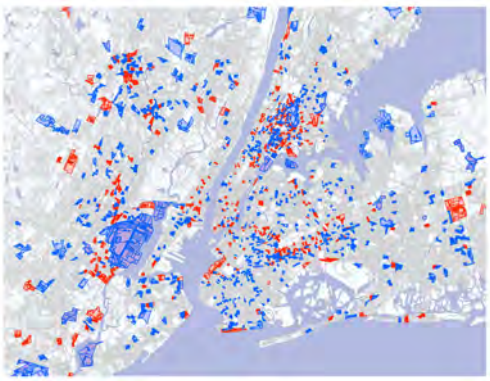


(b) African American ('Black', poor)

(c) Hispanic

(d) Hispanic (poor)

(e) Asian



(f) African American ('Black') and Hispanic, poor

FIGURE 1.11: Spatial distribution of race and poverty.

B. Constructing variables

Employment opportunities. For the purpose of our predicted benchmark, as can be seen in table 1.3, we use employment and firm opportunities as explanatory variables. To construct the latter, we need data on firms and individuals so then we know which firms are likely to hire an individual of a particular racial group. To this end, our strategy is to use U.S. data, except NYCBSA, hiring opportunities for each race and use that information to infer about distance and job opportunities in NYCBSA. To do so, for individuals, we make us of the Current Population Survey (CPS) [33]. It gives us monthly information about earnings, education, race, and other labor market characteristics, for 60,000 households stratified to be representative of the U.S. population. We construct U.S. wide average characteristics by skills, occupation, race and industry. Precisely, we use the CPS to determine which industries hire the most individuals by race, and by educational attainment as a proxy for poor households, in all the U.S., except the 25 counties that form NYCBSA. To keep stable industry classification and avoid the 2008 financial choc, we aggregate the Merged Outgoing Rotation Groups (MORG) of the CPS from 2000-2007 and 2011-2015. We end up analyzing, for U.S. except NYCBSA, around four million households job opportunities, for the four major races (see table 1.9 for all blocks). We will then use this information in NYCBSA to construct variables that capture local employment opportunities and estimate the potential access to jobs by race.

To do so, we now need data on firms and their location in NYCBSA. Hence, our third data source is the National Establishment Time Series (NETS), which includes about 1.45 million establishments in NYCBSA. Using their geographic location, we match each establishment to its census block, and then construct

33. Date obtained from : https ://www.nber.org/data/morg.html

–using population count by race from the Census and the industries identified from the CPS data previously– a measure of local "potential hiring" for each race. Table 1.9 shows the top five industries that are likely to hire each group. These measures include the number of employments and firms that are likely to hire a specific group within different radius ($1km$, $2-5km$ and $5km$). They capture the closeness to opportunity jobs which we will again use among other variables in the predicted benchmark.

Geographic controls. We follow Behrens *et al.* (2019) to derive our geographic controls. We construct 'Number of subway entrances' and 'Distance to nearest subway' using the locations of subway stations, provided by the Metropolitan Transportation Authority (MTA) obtained from the NYC OpenData website. For stations located along the Metro-North and Long Island Railroads, we use the publicly available *NYC Mass Transit Spatial Layers* produced by the GIS Lab at the Newman Library of Baruch College. Finally, the *New Jersey Geographic Information Network* provides us with similar information for lines operated by NJ TRANSIT as well as PATH (operated by Port Authority Trans Hudson) and PATCO (Port Authority Transit Corporation) lines. We then use GIS software to create a variable that gives the minimum distance of each block from a public transit stop, as well as the station count in the block. To construct 'Distance to nearest park', 'Distance to nearest waterfront', and 'Distance to CBD (Wall Street)', we first use the shape-based landmark dataset from the U.S. Census Bureau.The minimum distance of each block to parks is used to create the first variable (we keep only landmarks where the string 'Park' features in the name and drop all others, including those lacking a description). For the second variable, we compute the distance of each block to the closest block that is composed exclusively of water (which we call 'waterfront'). The last variable is the straight-line distance from

TABLE 1.9: The top five industries that are likely to hire each race : All

Race	Top 5 industries	share
Black	Tobacco manufacturing	25 %
	Taxi and limousine service	25%
	Barber shops	24%
	Bus service and urban transit	24%
	Fiber, yarn, and thread mills	23%
White	Coal mining	94%
	Farm product raw materials, merchant wholesalers	92%
	Lawn and garden equipment and supplies stores	90%
	Other motor vehicle dealers	90%
	Fuel dealers	90%
Hispanic	Animal slaughtering and processing	35%
	Cut and sew apparel manufacturing	32%
	Fruit and vegetable preserving and food manufacturing	29%
	Not specified metal industries	29%
	Landscaping services	28%
Asian	Nail salons and other personal care services	28%
	Electronic component and product manufacturing	18%
	Computer systems design and related services	14%
	Computer and peripheral equipment manufacturing	13%
	Software publishing	13%

Notes : In our regression we took the 25 top share industries, to have a broad view we present the top 5 in this table.

the block's centroid to the CBD, taken to be Wall Street.

Other controls. For the zero inflated Poisson model that we estimate for the predicted benchmark, we use other variables such as total population, housing units, owners, and renters count which are provided by the US census. We construct a variable of owner (renter) share by simply dividing the count of owners (renters) by the total count of owners and renters at census block level

Appendix C : Simulation of "toy city"

Why do we permute ? Why do we need to construct a benchmark using random permutations ? Could we not just follow the literature on segregation and take the overall distribution of population as the benchmark against which to compare the concentration of specific groups ? The short answer to that question is "no, we cannot". The reason is that there is a both "lumpiness problem" and potentially a "small sample" problem.

The lumpiness problem stems from the fact that our units of observations are blocks of different sizes. Consider a simple illustrative case. There are nine blocks with size $n_1 = 920$, $n_2 = 25$, $n_3 = 25$ and $n_4 = 25$, and $n_i = 1$ for $i = 5, 6, 7, 8, 9$. Hence, block 1 has 92% of the population, blocks 2–4 have 2.5%, and blocks 5–9 have 0.1%, respectively. Note that all blocks have positive population, i.e., the geography over which the empirical distribution of the population is computed consists of all blocks. Assume now that we have a subsample of 50 people (say black) and that we ask how their distribution would compare to an appropriate benchmark. That benchmark could be the overall population. In that case, we compute the K-density over the nine blocks and compare the empirical distribution of the 50 blacks to the empirical distribution of the total population. How

does that distribution compare to the benchmark in which we consider that the 50 blacks are randomly allocated across the nine blocks. To answer that question, we reshuffle the 50 blacks 1,000 times across the nine blocks, where each location is equally probable. We then aggregate up the number of blacks to the block level as required to compute the K-densities. The following table summarizes the results :

TABLE 1.10: Simulation results, 1,000 random permutations.

	Observed	Simulations		
# block	Pop share	Mean	90% interval	% zeros
1	0.925	0.9226	[0.86 ; 0.98]	0.000
2	0.025	0.0245	[0.00 ; 0.06]	0.279
3	0.025	0.0240	[0.00 ; 0.06]	0.273
4	0.025	0.0236	[0.00 ; 0.06]	0.292
5	0.001	0.0011	[0.00 ; 0.02]	0.947
6	0.001	0.0011	[0.00 ; 0.02]	0.954
7	0.001	0.0011	[0.00 ; 0.02]	0.946
8	0.001	0.0011	[0.00 ; 0.02]	0.944
9	0.001	0.0009	[0.00 ; 0.00]	0.954

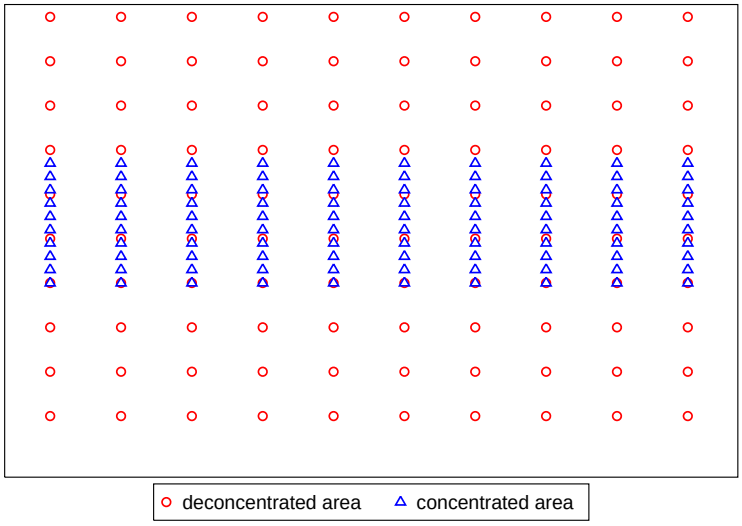
Notes : 1,000 random permutations.

As can be seen, there are a large share of zeros, i.e., many blocks do not enter the computations when we consider random permutations. Hence, the K-densities are estimated over different geographies than the one implied by the observed population distribution. The latter may hence not be the most appropriate benchmark.

To see this further, we run a set of simulation to capture different virtual scenarios. The aim of this exercise is to understand the behavior of our *case-control* strategy in a different scenarios. We want to know how the random permutation test react to different sampling, sizing and shaping. To do so, with regards to the control, we first suppose two different spatial environments : concentrated and not concentrated (see figure 1.12). Second, for each control, whether it is concentra-

ted or not, we also suppose that areal units[34] can have the same or different size. Hence, we analyze four different controls : (A) not concentrated with same size, (B) not concentrated with different size, (C) concentrated with same size and (D) concentrated with different size. Last, with regards to the case, we take 3 different subsample size : 200, 500 and 600 individuals (20 %, 50%, and 60% of the 1000 total population respectively). All in all, as shown in table 1.11, we end up with 12 case-control scenarios, and each one is simulated 1000 times. It is important to note here that the spatial location of the subsample, i.e. the case, has to be within the location universe of the control. This means that for every permutation, there is always a no zero probability that the case equals to the control.

FIGURE 1.12: Control area : concentrated and not concentrated



What do we learn from the simulation ? Figure 1.13 and figure 1.14 show both that when we increase the size of the *case*, from 20%, 50% to 60% , whether the units have same size (subfigures 1, 2 to 3 for not concentrated, and subfigures 7, 8

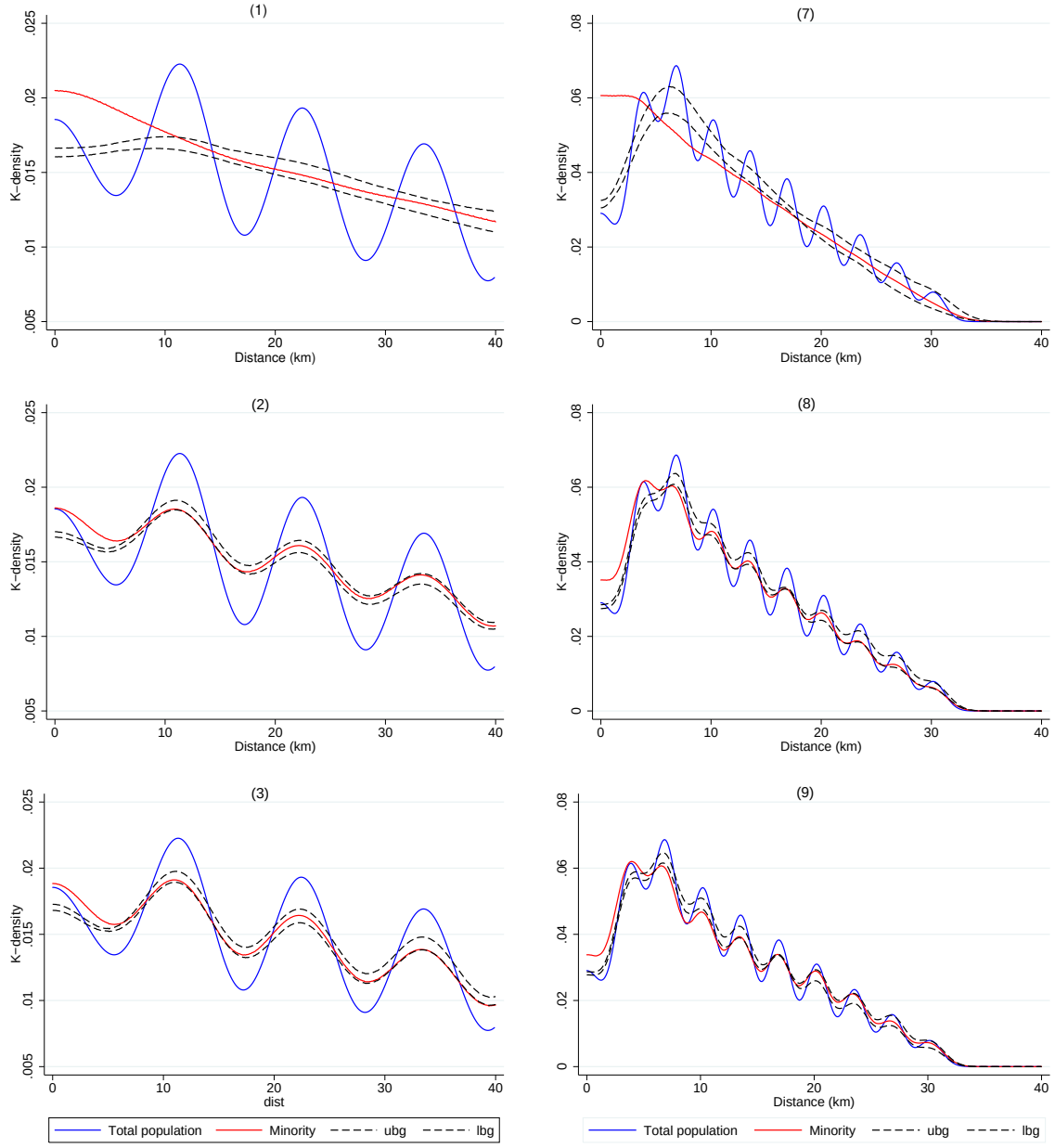
34. We run the simulations for 100 units.

TABLE 1.11: Case-control simulation

Spatial distribution	Block size	Sample size	Simulation
Not concentrated	Same size	20 %	1
		50%	2
		60%	3
	Different size	20 %	4
		50%	5
		60%	6
Concentrated	Same size	20 %	7
		50%	8
		60%	9
	Different size	20 %	10
		50%	11
		60%	12
Number of units : 100 Total population : 1000			

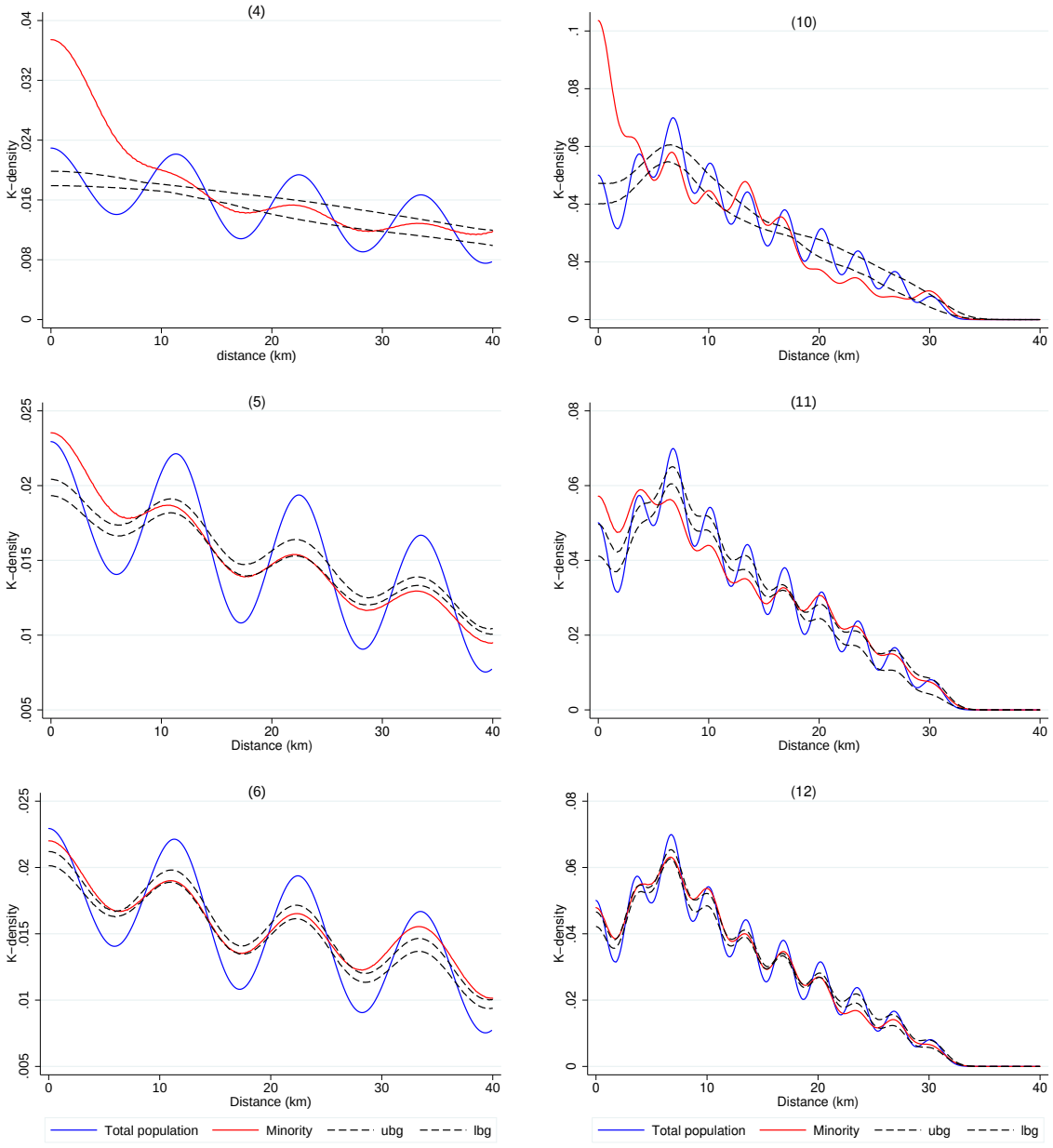
to 9 for concentrated) or different size (subfigures 4, 5 to 6 for not concentrated, and subfigures 10, 11 to 12 for concentrated), tends to be closer to the empirical distribution of the *control*. Hence, asymptotically, when one have large samples, the error measurement of using the empirical distribution as benchmark tend to zero. But, using the overall distribution of population as the benchmark in a "small sample" data could be misleading. In other words, if we have a case with bigger sample, it is less costly to use the overall distribution as reference departure of randomness. Owing to the fact that permutation test is time consuming, and the measurement error due to the comparison of the case distribution to the empirical one, tend to zero.

FIGURE 1.13: The effect of sample size (a)



Notes : in the subfigures (1), (2) and (3) ; the location universe of the control group are spatially equidispersed, whereas in the (7), (8) and (9) ; the location universe is spatially concentrated. In both case, the geographic units have the same size, and lbg and ubg are for lower and upper bound global confidence interval.

FIGURE 1.14: The effect of sample size (b)



Notes : in the subfigures 4, 5 and 6; the location universe of the control group are spatially equidispersed , whereas in the 10, 11 and 12 ; the location universe is spatially concentrated. For all the sixth simulations, the geographic units have different size, and lbg and ubg are for lower and upper bound global confidence interval.

Appendix D : Decomposing K-densities

Recall that one of the desirable property of segregation indices is the ability to disaggregate them to the individual level (**(P7)** : *disaggregation*). This allows for an analysis of segregation at the individual level. In our measures, using equation 1.7 and 1.8, we can decompose the K-densities to the level of the individual block to see which blocks contribute the most to the K-densities at certain distances. To illustrate the idea, figure 1.15 shows for instance that there is blocks that contribute more to the Black concentration than others. Black living in Brooklyn, Suffolk, The Bronks, Westchester and Newark counties seem to contribute the most to magnitude of segregation.

1.7 Conclusion

We have developed new tools that allow us to measure the extent of racial segregation in cities. Contrary to previous measures—which are based on the distribution of racial shares across areal units—our measures build on the distribution of bilateral distances across individuals. They encompass various existing approaches (evenness, exposure, isolation, clustering), satisfy many desirable properties, and allow us to think about evenness-clustering and exposure-isolation within a unified framework. Crucially, our measures allow us to compare the observed distributions against diffferent benchmarks—using a case-control design—thereby allowing us to partly disentangle segregation by race from segregation by poverty. They further allow for statistical testing of the observed patterns of segregation by simulating random distributions within appropriately chosen benchmarks.

To illustrate our measures, we then apply them to NYCBSA census data and find clear patterns of segregation along both income and race. Our results reveal that Black, Hispanic, and Asian are segregated, with income amplifying even

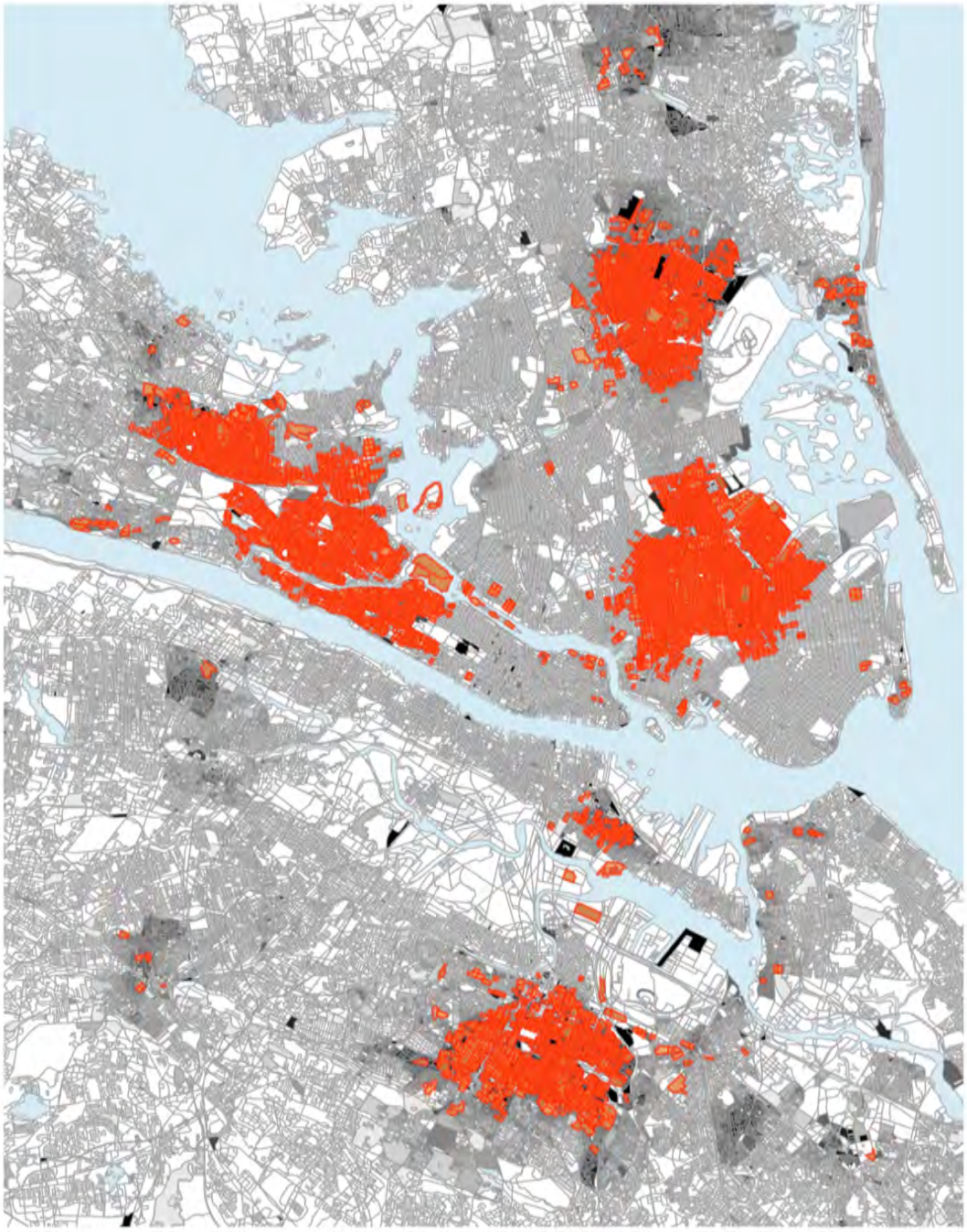


FIGURE 1.15: Top 10 percent of blocks that contribute to the Black CDF, 2010.

further the magnitude of their segregation. Results are similar when we look at how race amplifies income stratification as exemplified by the sorting of lower income groups. Hispanic are the most strongly segregated, whereas poverty adds the most to the segregation of Asian. Turning to measures of the pairwise exposure between groups, our results show that groups tend to isolate themselves from other groups. This effects is again especially strong for poor Asian with respect to the other groups, whereas Black and Hispanic are only little isolated from each other.

Our findings are robust to various distance thresholds and different sets of benchmark distributions. They are consistent with the vast literature about the presence of racial segregation in US cities. We add to the literature by showing the importance of disentangling segregation by income from segregation by race. From the viewpoint of policy makers who aim to adress segregation, understanding the race and income factors—and which one contributes how much—is important to design better policies. We hope that our flexible measures will prove useful to explore the location patterns in other cities and to provide new insights on the factors behind segregation and its dreary consequences.

CHAPITRE II

WHAT MATTERS FOR CHOOSING YOUR NEIGHBORS ? EVIDENCE FROM CANADIAN METROPOLITAN AREAS

Abstract

A corollary of the First Law of Geography and the Principle of Homophily is that "near things are more similar than distant things." We test that proposition using spatially fine-grained data on thousands of colocation patterns of ethnic groups in the six largest Canadian metropolitan areas. The geographic patterns reveal that groups that are more similar along various non-spatial dimensions—language, culture, religion, genetics, and historico-political relationships—colocate more. These results are robust to numerous controls and provide a quantitative glimpse of the 'deep roots' of homophily.

Keywords : colocation patterns ; ethnic segregation ; homophily ; culture and language ; historico-political relationships.

JEL Classification : R23 ; Z13.

2.1 Introduction

The *First Law of Geography* (Tobler, 1970) states that "everything is related to everything else, but near things are more related than distant things." The *Principle of Homophily* (McPherson *et al.*, 2001) in sociology and social psychology posits that "similarity breeds connection."[1] Being related requires to be connected and similar enough to interact. Thus, a corollary of the First Law and of the Principle is that "near things are more similar than distant things."

We test this corollary using spatially fine-grained data on thousands of colocation patterns in the six largest Canadian metropolitan areas. We exploit a unique feature of the census, namely to provide a detailed portrait of the population's ethnic and cultural origins. The census gathers information about ancestry, thus allowing us to measure how groups from diverse backgrounds relate to each other within cities. The colocation patterns reveal that populations that are more similar along various non-spatial dimensions—language, culture, religion, genetics, and historico-political relationships—colocate more. These results are robust to the inclusion of geographic and economic controls and survive an extensive battery of

1. McPherson *et al.* (2001, p.415) summarize the Principle as follows : "Similarity breeds connection. This principle—the homophily principle—structures network ties of every type, including marriage, friendship, work, advice, support, information transfer, exchange, comembership, and other types of relationship. The result is that people's personal networks are homogeneous with regard to many sociodemographic, behavioral, and intrapersonal characteristics. Homophily limits people's social worlds in a way that has powerful implications for the information they receive, the attitudes they form, and the interactions they experience. Homophily in race and ethnicity creates the strongest divides in our personal environments, with age, religion, education, occupation, and gender following in roughly that order. Geographic propinquity, families, organizations, and isomorphic positions in social systems all create contexts in which homophilous relations form."

checks.

Models of segregation date back to at least Schelling (1969, 1971). They show that even weak preferences for own type—homophily—generate spatial clusters of individuals belonging to the same group. While this is well understood theoretically, much of the empirical literature has focused essentially on the outcomes—e.g., peer effects in poverty, crime, and education—rather than on the causes of stratification. What are the 'deep roots' of preferences for own type ? What exactly is 'own type' ? Which 'own type'-characteristics are associated with more or less stratification in cities ? And are the relationships causal ? Providing answers to these questions is important for urban policy that aims at diversity in neighborhoods. If homophily is deeply rooted in language, religion, culture, or long-bygone historical events—such as past conflict or dominance relationships—achieving more diversity in residential patterns may be difficult. Affecting slow-changing fundamentals is hard compared to causes of stratification that originate from discrimination in the housing market, income inequality, red-lining, or other institutional aspects of the economy.

Identifying and disentangling the deep roots of homophily that underlie geographic stratification is difficult for at least three reasons. First, to paint a broad quantitative picture, we need measures of the location patterns of many groups as well as proxies for the different dimensions of 'preference for own type'. There is an extensive literature that has looked at how ethnic and historic characteristics—which shape 'preference for own type'—translate into important outcomes such as the quality of institutions, growth, or armed conflict (e.g., Alesina *et al.*, 2003; Fearon et Laitin, 2003). We draw on the measures developed in that literature. Second, we have to deal with the problem that homophily leads to observationally equivalent outcomes : "near things are more similar than distant things," irrespective of the mechanisms at work. This makes disentangling the mechanisms very

difficult. Last, there are a number of econometric identification concerns we have to deal with. In particular, omitted variable bias and reverse causality loom large. Different ethnic groups may colocate because of unobserved spatial characteristics that are independent of homophily. Furthermore, location patterns usually feed back on homophily as individuals become more similar to the individuals with which they interact (McPherson *et al.*, 2001). We thus need measures of similarity between groups that are exogenous to observed location patterns.

We deal with these problems by exploiting spatially fine-grained census data. We use self-reported data on ethnic origin to compute thousands of *colocation patterns of ethnic groups* in the six major Canadian metropolitan areas. Pairs of ethnic groups display substantial variations in linguistic, religious, cultural, and genetic proximity, as well as in their historico-political past as captured by, e.g., hegemony and colonial relationships. Given that variation, the colocation patterns should reveal—at least partly—if measures of similarity between ethnic groups translate into more geographic proximity. They should also substantiate information on the key dimensions of 'preference for own type'. Using colocation patterns is important and, to our knowledge, novel in this context. The bulk of the literature on segregation has looked at the geographic clustering of own type only—mostly broad ethnic aggregates such as African-Americans or Hispanic. This poses problems because individuals of the same ethnic groups are always similar to each other along almost all dimensions. Instead, we want to analyze location patterns of individuals who are *similar along some dimensions yet dissimilar along others*. Doing so will allow us to alleviate the observational equivalence problem and to better disentangle the contribution of different characteristics of homophily to observed colocation patterns. [2]

2. Ellison *et al.* (2010), Behrens (2016), and Faggio *et al.* (2017) make the same point concerning the location patterns of industries. The geographic concentration of one industry is not very

Our measures of similarity—linguistic, religious, cultural, and genetic—and of historico-political relationships are derived and adapted from existing country-level databases. Using country-level data on ethnic similarity to look at colocation patterns has the obvious advantage to alleviate problems of reverse causality. This is especially important when working at a fine spatial scale as we do, where unobserved spatial features or reverse causality—from colocation patterns to similarity—may be more acute. It will also make it more challenging to uncover significant effects since there is more measurement error using the country-level proxies and much more idiosyncrasy at a fine geographic scale. [3] Despite the sometimes coarse nature of our proxies, the presence of substantial idiosyncrasy, and conservative standard errors, we find statistically strong effects of our covariates on ethnic colocation patterns.

Our key results are summarized as follows. First, religious, linguistic, cultural, and genetic proximity all have positive and significant effects on observed colocation patterns, even when controlling for a wide range of geographic and economic covariates and when including them all simultaneously. We also find that past political relationships have a legacy that extends across time and space to today's location patterns. These results are highly robust to how we measure similarity between groups. We view this as evidence for the corollary that "near things are more

---

informative to understand the underlying agglomeration mechanisms. Colocation patterns of industry pairs are more informative because industry pairs may be similar, and interact, in some dimensions—e.g., patent citations, buyer-supplier relationships, labor market pooling—but not in others.

3. The Second Law of Geography (Arbia *et al.*, 1996) states that "[e]verything is related to everything else, but things observed at a coarse spatial resolution are more related than things observed at a finer resolution." While this is more a technical consideration related to the 'modifiable areal unit problem' (MAUP) than a law properly speaking, we will show that we find strongly signicifant results even at a fine spatial resolution.

similar than distant things." Second, the effects we uncover hold broadly across cities, but with city-level heterogeneity. Some variables—language, religion, and past colonial relationships—even display a fairly pronounced east-west gradient. Linguistic similarity has, e.g., the largest effect in Ottawa and Montréal, but less so in Toronto or the western metropolitan areas. Last, we provide results using sample splits along dimensions that we believe are informative to better understand the observed patterns and that allow us to partly control for unobserved locational characteristics that may confound our results. Using only residents living in poor areas and in rental-dominated areas, we find that our results are basically unchanged. The same holds true when focusing on pairs from Africa that may face more discrimination in the housing market. This suggests that our results are not entirely driven by locational constraints that force some groups—e.g., poorer ethnic groups—to colocate solely because they have no other choice. Results using other splits—e.g., rich residents and owners—are qualitatively very similar.

Our paper is related to several strands of literature. First, it is closely related to the large and diverse literature on the effects of similarity on economic exchange such as migration, trade, and investment between countries (see, e.g., Guiso *et al.* 2009). In particular, it is related to papers that focus on the location decisions of migrants (see, e.g., Lazear 1999, for a model of immigrant sorting). While most of that literature has used large geographic areas—countries or counties—we focus on smaller geographic scales. Most closely related is a recent paper by Falck *et al.* (2012, p.226), who show that historic dialect-similarity between regions still shapes contemporanous interregional migration patterns in Germany. They find that "cultural factors are thus likely to influence [interregional migration] even more strongly than, say, the decision to trade goods with someone from a different region." We show that the results continue to hold at even smaller

geographic scales, namely within cities.

Second, our paper is related to the extensive literature on the causes of segregation in cities (see, e.g., Cutler *et al.* 1999 ; Bayer *et al.* 2004 ; and Boustan 2013 for a recent survey). We contribute to that literature by showing how information on exposure—i.e., contacts between groups—can be used to better identify the deep roots of preference for own type that seem key to understand, at least partly, observed segregation patterns.

Last, our paper is also related to the recent literature that exploits industrial colocation patterns to better identify the sources of agglomeration economies. (see, e.g., Ellison *et al.* 2010 ; Faggio *et al.* 2017). We extend this approach to residential location patterns and show its usefulness to better disentangle the drivers of geographic sorting and the sources of homophily.

The remainder of the paper is organized as follows. Section 2.2 lays out the methodology, and explains our data and measurements. Section 2.3 explains the empirical strategy and discusses identification concerns. Section 2.4 presents our results. It also contains many extensions and a battery of robustness checks. Last, Section 2.6 concludes. We relegate some details on our data and additional results to a set of appendices. Additional material is available in a separate online appendix.

2.2 Measurement and data

We require both measures of geographic proximity of ethnic groups and of non-geographic proximity—similarity—of these groups. We now explain what data we use and how we construct our measures.

2.2.1 Geographic proximity between groups

Census data on ethnic origin

To measure the geographic proximity between different ethnic groups, we firstly require numerous and sufficiently large groups. It is well documented that new immigrants disproportionately arrive and settle in the large metropolitan areas where the ethnic composition is especially diverse : "More than 60% of immigrants and 70% of recent immigrants live in Canada's three largest cities—Toronto, Montréal and Vancouver. Nearly 80% of immigrants live in the thirteen urban areas." [4] We hence restrict our analysis to the six largest Canadian metropolitan areas in 2016 : Toronto, Montréal, Vancouver, Calgary, Ottawa, and Edmonton. These six metropolitan areas all had population above 1 million and together they concentrate 16.37 million people, or 46.6% of the Canadian population.

We secondly require the spatial distribution of the groups. We use geographically fine-grained data from two census waves : 2006 and 2016. We discuss the differences between 2006 and 2016, and why we exclude 2011, in Appendix A.1. Ideally, we would like to know the exact geo-referenced distribution of population by ethnic origin, but this is not publicly available due to confidentiality reasons. We hence use the smallest spatial unit for which publicly available data are reported in Canada : dissemination areas (DA). [5] There are 54,624 DA in the 2006 census

4. See `https://www.canada.ca/en/immigration-refugees-citizenship/corporate/reports-statistics/ research/recent-immigrants-metropolitan-areas-canada-comparative-profile-based-on-2001-census /partg.html`, last accessed on February 1, 2019.

5. The smallest units at which population and dwelling counts are provided are dissemination blocks, but no other data—e.g., ethnic origin—are reported at such small geographic scale. DA are delineated using a population criterion, so that they can be relatively large in rural areas. Yet, they are small geographic units in the densely populated urban areas we focus on : in 2016,

and 56,589 DA in the 2016 census. Of these, 21,155 and 22,261 are located in the six largest metropolitan areas that we focus on.[6] Each dissemination area is geo-referenced by its population-weighted latitude and longitude centroid which we use as our geographic locations in what follows. Figure 2.3 in Appendix A.1 illustrates the granularity of our data.

The Canadian census provides a detailed portrait of ethnicities at the DA level. Ethnic origin is different from citizenship, which is important for our analysis. Indeed, in countries such as Canada—where immigrants constitute a large share of the population and where citizenship can be obtained relatively quickly— using citizenship as a proxy for ethnic origin is often not meaningful. As stated by Statistics Canada : "Ethnic origin refers to a person's 'roots' and should not be confused with citizenship, nationality, language or place of birth. For example, a person who has Canadian citizenship, speaks Punjabi (Panjabi) and was born in the United States may report Guyanese ethnic origin."[7] The Canadian census hence asks explicitly about ethnic origin using the following question : "What were the ethnic or cultural origins of this person's ancestors ?" The question is accompanied by two notes stating : "(1) This question collects information on the ancestral origins of the population and provides information about the composition of Canada's diverse population" ; and "(2) An ancestor is usually more distant than a grandparent."

---

the median surface is 0.3 square kilometers, the average surface is 3.7 square kilometers, and the surface at the 90th percentile is 2.01 square kilometers.

6. For some DA we do not have relevant census data (e.g., on income), so we drop them from our analysis.

7. See also `https://www12.statcan.gc.ca/census-recensement/2016/ref/guides/008/98-500-x2016008-eng.cfm` for additional details, last accessed on February 1, 2019.

The objective of these questions is to understand the roots of the respondent's origins, or his perceptions of his roots. For instance, a person who has Canadian citizenship, speaks Berber and was born in France may report Algerian ethnic origin, and another person with the same background could report French as his ethnicity. Thus, the measure is highly subjective but more likely to capture how people view themselves in terms of their cultural-ethnic background. We choose this measure because of data availability, but also because there is no consensus in the literature about how to measure 'ethnicity' (see, e.g., Burton *et al.* 2010 for a recent discussion). Ethnicity is a multidimensional concept and cannot be readily reduced to a single dimension. Yet, if we only have access to a single dimension—which is usually the case in large datasets such as the census—self-reported perception of ethnic origin seems the most appropriate measure of ethnic background.

Each respondent can report one or more ethnicities. We use the total counts of unique *and* multiple responses, meaning that a person may have a single ethnic origin, or may have multiple ethnicities and thus may be counted twice or more. As a result, when these data are summed across all ethnicities, the total count exceeds that of the total population living in Canada. We view the possibility to report multiple ethnicities as a strong feature of the data because it allows people to finely express how they perceive themselves. This would be more difficult using citizenship data.

While the census data on ethnic origin has many advantages for our analysis, it also has a number of shortcomings. First, like any self-reported data, our data are likely to suffer from reporting bias. For example, people's responses may be—in part—conditioned by their environment : a Chinese person living in China town may report 'Chinese' as ethnic origin, whereas a Chinese person living somewhere else may report 'Canadian' as ethnic origin. In other words, location may shape

self-perception. While we cannot rule out this possibility, we do not think that this is generally a major problem, especially since the census asks explicitly about the ethnic origins of the ancestors and allows for multiple responses. Second, because of confidentiality reasons, ethnic groups are only reported if their national count exceeds 800 individuals. We do not think that this is a problem for us since the samples become so small with less than 800 individuals that a city-by-city estimation of colocation patterns makes hardly sense anymore. Third, we only observe the aggregate population numbers by ethnic group at the DA level, but not the within-DA allocation. Since we do not observed the within-DA allocation, we implicitly assume that all people live at the centroid, which creates measurement error. We explain below how our colocation measure deals with that problem using kernel smoothing. Last, and potentially more worrisome, the public-release ethnic counts at the DA level—as well as all other count variables at that geographic scale—come from 25% samples of the universe and are randomly rounded up or down to the closest multiple of 5. Put differently, when there are 5 Irish reported in a DA—according to the estimates based on the 25% samples—this could represent any number between 1 and 9. Hence, there is additional random measurement error that will affect our colocation measures. We argue below that this should not matter substantially for our analysis given the random nature of the rounding.

Mapping ethnic groups to countries

While there exist many variables that measure relationships and similarity between country pairs, such variables are not readily available for ethnic pairs. The latter are usually not associated with administrative units and thus no data are collected for them. Hence, to construct our explanatory variables that measure the non-geographic proximiy—similarity—between groups, we need to work at the level of countries. This then requires us also to measure the colocation of groups by

country. To this end, we map ethnic groups to countries using the Geo Referencing of Ethnic Groups (GREG) database (Weidmann *et al.*, 2010). We proceed as follows. First, when a respondent reports an ethnic origin using a country name (say Ukrainian, Russian, or Italian), we directly associate this respondent with the corresponding ISO3 country code. Second, when a respondent reports an ethnic origin that is not associated with a precise country (say Basque, Catalan, or Berber) we associate him with the countries that contain the ethnic group, using weights that represent the share of population of that ethnic group living in the different countries where this ethnic group can be found. We provide additional details on the procedure in Appendix B. Let us emphasize that this procedure is applied to less than a third of our ethnic groups.

Measuring geographic colocation of groups

We finally need to measure the geographic colocation of the different groups. Consider two ethnic groups, superscripted by i and j. We only look at geographic concentration patterns for groups $i \neq j$ in the same city c.[8] Assume that there are $n_l^i \geq 0$ and $n_l^j \geq 0$ people of groups i and j located in DA l, and $n_m^i \geq 0$ and $n_m^j \geq 0$ people of groups i and j located in another DA m. Following Duranton et Overman (2005, 2008), we estimate the K-density of all bilateral distances between individuals belonging to i and j at distance d for city c, having L_c locations in total, as follows :

$$\widehat{k}_c^{ij}(d) = \frac{1}{h \sum_{l=1}^{L_c} \sum_{m=1}^{L_c} n_l^i n_m^j} \sum_{l=1}^{L_c} \sum_{m=1}^{L_c} n_l^i n_m^j f\left(\frac{d - d_{lm}}{h}\right), \tag{2.1}$$

8. The main reason for doing so is that a group is always 'similar' to itself along all dimensions, which makes disentangling the drivers of geographic concentration difficult (see Ellison *et al.* 2010 for a discussion).

where $f(\cdot)$ is a Gaussian kernel and h is the bandwidth parameter set using Silverman's rule-of-thumb. The estimator in (2.1) gives us, for each distance d, the kernel-smoothed share of bilateral distances between people of groups i and j in city c. To obtain an aggregate measure of geographic proximity between groups i and j in city c, we then compute the cumulative distribution as follows :

$$\widehat{K}_c^{ij}(d) = \int_0^d \widehat{k}_c^{ij}(\zeta)\mathrm{d}\zeta. \tag{2.2}$$

The measure (2.2) states what share of bilateral distances between people of the two groups is smaller than d in city c. If, for example, $\widehat{K}_c^{ij}(1km) = 0.3$ for $i =$ Nepal and $j =$ Buthan, this means that 30% of bilateral distances between pairs of Nepalese and Buthanese in city c are less than 1 kilometer. Alternatively, we may interpret this as the probability that a random draw of one Nepalese and one Buthanese in city c yields a pair that lives less than 1 kilometer from one another. The larger $\widehat{K}_c^{ij}(d)$, the more colocated are the groups i and j in city c.

Note that the kernel smoothing in (2.1) is important. This is firstly because we assign populations to centroids of the DA, as explained before, since we do not know the exact within-DA distribution. Even if the centroids provided in the data are already population weighted, kernel smoothing is useful to deal with that type of measurement error. Secondly, we compute distances using the great-circle formula (which, at the level of a city, basically is the straight-line distance). Kernel smoothing deals with the fact that the straight-line distance may be a bad proxy for travel distances in the city (see Duranton et Overman 2005 for additional discussion). [9] Last, as explained before, there is random rounding of the population weights n_l^i and n_m^j to the neareast multiples of five in the census data. Since the K-densities are smoothed and computed over the whole metropolitan area, we do

9. However, dense road networks in cities certainly make the straight-line distance a better proxy for travel distance than in less dense rural areas.

not think that this makes a big difference : the rounding is random, so there should be no systematic bias in results. Since the random rounding affects, however, more strongly the smaller groups, we will control for group size in the regressions to partially capture effects that may be due to the differential impact of random rounding across groups of different sizes. [10]

We compute our measures of geographic concentration for all pairs of ethnic groups in each city, both for the 2016 and the 2006 censuses. This yields our dataset with 68,055 kernel densities for 2016, and 56,160 kernel densities for 2006. Each density is estimated on the range from 100 meters to 5 kilometers, with 100 meter steps (hence a total of 50 estimates for each city-ethnic pair combination). We will provide robustness checks using an alternative measure of colocation—the Ellison-Glaeser index (Ellison et Glaeser 1997)—later in the paper.

2.2.2 Similarity between groups

Our second key ingredient are measures of non-geographic proximity—similarity—between groups, which constitute our explanatory variables. We here provide a quick overview of the linguistic, religious, genetic, economic, historico-political, and geographic data that we use in our analysis. A more detailed description is relegated to Appendix A.2, and Table 2.12 there provides the full list of our variables.

---

10. It is also important to point out that the random rounding of the weights makes the use of more 'local' and unsmoothed measure of colocation of ethnic groups more problematic. For example, looking just at some specific locations in the city may provide fairly inaccurate measures of colocation. Our measures are aggregated over the whole metropolitan area and smoothed, so they should be more robust to random rounding of the weights, as well as to potential mismeasurement of distance and within-DA location patterns.

Cultural variables

Culture may be viewed as a symbolic and behavioral marker of ethnic groups. People who share cultural traits and norms may be more inclined to locate near each other for reasons of homophily. We draw on existing sources for language, religion, and cultural distance as our explanatory variables to proxy for 'culture' in a broad sense. We conjecture that speaking the same (or a similar) language, having a common (or a similar) religion, and being generally 'culturally close' will ceteris paribus lead to more coagglomeration between ethnic groups. Our two main data sources are Melitz et Toubal (2014) and Spolaore et Wacziarg (2009). The former provide measures of common language, linguistic proximity, and common religion. The latter provide another set of linguistic distance measures, as well as measures of religious and cultural distances (the latter being constructed from the *World Values Survey*, WVS).

Measures of linguistic proximity. Melitz et Toubal (2014) provide measures of linguistic proximity : Common official language (COL) ; Common spoken language (CSL) ; Common native language (CNL) ; and two measures of linguistic proximity (LP1 and LP2). COL_{ij} is a binary variable that takes value 1 if the pair ij 'shares the same official language', and 0 otherwise. CSL_{ij} takes values from 0 to 1 and reflects the probability that a randomly drawn pair of people from countries ij understand each other. CNL_{ij} is defined analoguously, but restricted to native speakers among all speakers. CSL_{ij} and CNL_{ij} require the languages to be spoken by at least by 4% of the population of each country in the pair ij. Note that CSL_{ij} is necessarily greater or equal than CNL_{ij}, as it includes non-native speakers in addition to native speakers. Linguistic proximity refers to the closeness of two different native languages. Two measures—LP1 and LP2—are used, which both range from 0 to 1. LP1_{ij} compares languages of different trees, branches,

and sub-branches; it takes lower values if two languages spoken in i and j belong to different trees and higher values if they belong to the same sub-branch. LP2_{ij} creates a similarity measure by comparing and analyzing lexical similarities between 100 to 200 words of the languages spoken in i and j.

Spolaore et Wacziarg (2009, 2016, 2018) provide additional measures of linguistic distance. The first measure (TLD_{ij}), is obtained by grouping languages into families and looking at their similarities. It resembles LP1 since it is based on comparisons of trees. It is standardized to range from 0 to 1, with higher values indicating more similarity. A weighted version (TLD_{ij}^{W}), that weights by linguistic group sizes in each country, is also provided. A second type of measure is based on Lexicostatistics that classifies languages based on whether the words used convey some common meaning (i.e., are cognate). Proximity between languages is measured by the percentage of cognate words.

In what follows, we use Common official language (COL) as our baseline measure, but we will show that the results are robust to how we measure linguistic proximity.

Measures of religious proximity. Our first measure from Melitz et Toubal (2014) is referred to as 'common religion'. It is constructed as the probability that two people drawn at random from two countries i and j share the same religion. We further use two measures provided by Spolaore et Wacziarg (2009, 2016, 2018). They compute religious distance in a similar manner than linguistic distance, based on religion trees. Both a weighted and an unweighted measure are provided, and we will show that our results are robust to the measure that is used.

Measures of cultural proximity. Last, Spolaore et Wacziarg (2009, 2016, 2018) also provide different measures of cultural distance, constructed from the WVS.

The latter provides answers to 740 questions about values, norms, and attitudes across countries in the world. They compute eight different Euclidian cultural distance (ECD) indices, based on different subsets of questions asked in the WVS—ranging from questions about "Perception of Life" to "Politics and Society" or "National Identity". More details are provided in Appendix A.2.

Genetic variables

Genetic data is widely used to meausre the relatedness of populations. Genetically closer populations tended to interact more in the past and are more likely to share common traits today. We are interested in whether individuals that report belonging to two genetically close ancestors—or where one is the ancestor of the other—are spatially more colocated. We provide details on how we measure genetic distance in Appendix A.2. We follow Spolaore et Wacziarg (2016), who build on the landmark study by Cavalli-Sforza *et al.* (1994) which measures genetic distance using the distribution of gene variants—e.g., alleles—across populations. The latter provide a worldwide dataset on genetic distance at the population level, which we can match to country-level data using ethnic composition by country from Alesina *et al.* (2003). We also use a second class of measures based on early data on microsatellite variation by Pemberton *et al.* (2013), which has wider coverage of populations (267 populations from Europe, Asia, and Africa). We again match these measures to countries using the ethnic composition by country from Alesina *et al.* (2003).

Our baseline measure of genetic distance is based on 'alelle and plurality groups', but our results are robust to different types of genetic distance, e.g., when using micromarker-based measures. Note also that it is hard to separate genetic distance from cultural distance. Indeed, some authors argue that genetic traits and cultural

traits are intertwined, so that the genetic variables should be viewed as a part of the cultural variables. We take no stand on that issue and report the genetic variables separately. We could equally well include them in the cultural variables and this would not change anything in our subsequent analysis.

Economic variables

Economic interactions between populations and countries help to shape social interactions between groups. For instance, Martin *et al.* (2008) find that trade openness between countries i and j has a negative effect on the likelihood of having a war between those countries. Generally, the literature on the 'gravity equation' in international trade has substantiated that many geographic and historico-political variables are correlated with bilateral trade and investment flows (see, e.g., Head et Mayer 2014 for a recent survey). We are thus interested in how more economic exposure to each other—via more trade, economic agreements, or migration and tourism—is possibly reflected in within-city location patterns of ethnic groups. To this end, we focus on the following economic variables : the value of bilateral trade flows between countries i and j ; the existence of bilateral agreements (e.g., free trade agreements or currency unions) ; and the number of tourists from country i that visited country j. We also take into account the per capita GDP gap between countries i and j, since this gap is related to both trade patterns and foreign direct investment. We add these economic variables as controls to purge effects that may be correlated with our key variables of interest, namely linguistic, religious, and genetic proximity, as well as historico-political factors.

Historical and political variables

We use data provided by Head *et al.* (2011) and made available by the 'Centre d'études prospectives et d'informations internationales' (CEPII) to control for a wide range of historico-political factors affecting the present and past relationships between country pairs ij. In our baseline regressions, we include 'common colonizer'—i.e., a dummy indicating whether the two countries had the same colonizers—and 'colonial relationship' status—if one country was a colonizer of the other. We also include a dummy indicating whether the two countries were part of the same country in the past (e.g., former USSR or Yugoslavia). Furthermore, we use a number of dummy variables as robustness checks : if the pair ij has been in armed conflict ; whether there is a hegemony relationship ; if they have common legal origins ; or if they both belong to the OECD. Because the effect of either conflicts or past colonial relationships are likely to dissipate over time, we also construct time-varying variables. More precisely, we choose post-1945 dates of either conflict or independence and construct variables as the current year minus the conflict year or the current year minus the independence year (conditional on the pair having been in a colonial relationship or in armed conflict).

Geographic variables

Finally, we complement our set of variables with basic geographic controls. The inclusion of these controls is important since it is well known that linguistic, genetic, and cultural distance are all—at least partly—correlated with geographic distance (see, e.g., Ramachandran *et al.* 2005 for a discussion on genetic distance). Hence, purging the effect of geographic distance is necessary to capture the non-geographic part of these measures. We control for common border and continent in our regressions using CEPII data. These measures are highly correlated with

different distance measures between countries, such as the distance between their capitals or their major cities (either unweighted or population weighted). We focus on common border and continent as these measures make more sense to us than the distances between the capitals or major cities. Intuitively, what matters are neighbors and a common history, and those are fairly well captured by common borders and belonging to the same continent. Distances between capitals or major cities also display substantial variation across continents and are a noisier measure than our dummies for common borders or same continent.

2.3 Empirical strategy

We now explain in detail our empirical strategy and discuss the identification concerns we need to deal with.

2.3.1 Estimating equation

Our basic specification is the following linear model :

$$\widehat{K}_c^{ij}(\overline{d}) = \alpha + X^{ij}\beta + \delta_c^i + \delta_c^j + \varepsilon_c^{ij}, \tag{2.3}$$

where $\widehat{K}_c^{ij}(\overline{d})$ is our measure of colocation of groups i and j in city c at distance $\overline{d}$; X^{ij} are country pair-specific covariates that measure linguistic, religious, cultural, genetic, and geographic proximity, as well as historico-political and economic relationships ; and δ_c^i and δ_c^j are city-country fixed effects.[11] They capture, among other things, differences in the sizes of ethnic groups, differences in the spatial extent and the density of cities, and differential tendencies of a group to cluster with itself (i.e., the differential tendency of within-group geographic concentra-

11. Following Ellison *et al.* (2010), the city-country fixed effects are constructed such that $\delta_c^i = 1$ if country i figures in the pair ij (in any order) in city c, and zero otherwise.

tion). We do not think that results without these fixed effects make sense and therefore only report results including them.[12] Note that since the K-densities $\widehat{K}_c^{ij}(\overline{d})$ are by construction symmetric in i and j—since distances are symmetric—we include for each pair ij only one of the ordered pairs (ij or ji). We also exclude all pairs ii, i.e., the geographic concentration of a single group, since we have no measures of similarity of the group with itself. Thus, given N groups we have $N(N-1)/2$ unique pairs.

The K-density on the left-hand side of (2.3) can be evaluated at any distance to capture the geographic concentration of the pair ij up to that distance. Since the effects that we are looking for are likely to operate at small spatial scales—e.g., in the neighborhood of individuals—we look in what follows at distances of $\overline{d} =$ 100 meters, 500 meters, and 1 kilometer. We take 500 meters as our benchmark distance, which corresponds to a 5 minutes walk at reasonable walking speeds. It also corresponds to the distance beyond which numerous neighborhood amenities tend to not be significant anymore in terms of defining the neighborhood (Hidalgo et Castañer, 2016).

We standardize all variables—so that our coefficients measure effect sizes—and we cluster the standard errors by country pairs ij. Recall that we have no variation in ij across cities and this is the dimension of our key variables of interested.[13]

---

12. Larger and less compact cities tend to mechanically have lower K-density CDFs at each given distance than smaller or more compact cities, just because they are geographically more spread out. This is an undesirable effect we need to purge from the estimations. Also, ethnic group sizes vary strongly across cities, and smaller groups tend to be more geographically concentrated. Again, this is not desirable for our estimations. We have experimented with separate country and city fixed effects, as well as with controls for the city-specific sizes of ethnic groups. The results are in line with those we report here.

13. We have a large number of clusters, as required for reliable inference (see Angrist et Pischke

Although effect sizes are useful to assess the relative importance of the explanatory variables, measures of language, culture, religion, genetics, and historico-political relationships might be fairly collinear. Hence, if some measures are better proxies than others, it will be difficult to assess their relative importance. Table 2.13 in Appendix A shows that our explanatory variables are not too strongly correlated. Still, we should not read too much out of the relative magnitudes of the coefficients as they may partially capture the same underlying characteristics.

2.3.2 Identification concerns

Our explanatory variables X^{ij}, described in Section 2.2.2, are arguably exogenous to location patterns in Canadian cities. It is indeed unlikely that the colocation patterns of say Indians and Pakistanis in Toronto have any bearing on trade between Pakistan and India or linguistic or religious proximity between those countries. There is not a single of our variables at the ij level between countries that could be fundamentally determined by how ethnic groups colocate in Canada. Hence, there are no problems of reverse causality that we would need to address using instrumental variables. In what follows, we report OLS estimations.

There may, however, be omitted variables specific to the country pairs ij that are correlated with both our X^{ij} and $\widehat{K}_c^{ij}(\overline{d})$.[14] We have no cross-city variation in the X^{ij}, and little to no time variation (since colocation patterns change slowly and the similarity measures X^{ij} are time invariant), so we cannot include ij fixed effects. We mitigate the problem of omitted variables the best we can by controlling for an exhaustive set of ij-specific covariates related to geographic proximity

---

2009).

14. We discuss the scope for selection bias in the supplemental online appendix. Given that we do not think this is a problem, we do not provide more details here.

and economic relationships. Of special importance is the inclusion of geographic controls to purge the potential correlations of our similarity measures with geographic distance, thus making sure that we are not picking up purely geographic effects in terms of proximity between country pairs and the ethnic groups that populate them. Furthermore, we include country-city fixed effects δ_c^i and δ_c^j in all specifications. These control, in a fairly exhaustive way, for all country-city-specific factors such as the sizes of ethnic groups, the spatial extent and density of the cities, and differences in province-level immigration requirements and city-level policies. Last, we will also report results where we first-difference the geographic patterns between 2006 and 2016 and regress them on the the initial levels of X^{ij}.

Given our set of controls and the variables that we include related to geography, economics, culture, language, religion, historico-political relationships, and genetics, it is hard to think of other omitted factors that would be both correlated with the X^{ij} and that would have a direct effect on the colocation patterns of groups i and j. One notable exception is linked to factors that arise *within* Canadian cities and that are related to both the locations of groups i and j and correlated with X^{ij}. To understand that problem, let $\widetilde{K}_c^{ij}(\overline{d})$ denote the *counterfactual* colocation measure between groups i and j in a world where the two groups make independent random choices *within their feasible location sets* (i.e., the sets of locations they could potentially choose in the city). To fix ideas, assume that groups i and j share a common religion, yet do not seek to be close to each other based on that criterion. Assume further that there is religious discrimination in the city, which targets systematiclly people with that religious affiliation ('religious redlining'). Then, groups i and j may be *constrained* to pick from the same spatial choice sets and, therefore, may end up being close together in the city. This would create a spurious correlation between religious similarity and geographic proximity that is unrelated to homophily but originates from discrimination in the

housing market.[15] Formally, if $\mathbb{E}(\widetilde{K}_c^{ij}(\overline{d})X^{ij}) \neq 0$, and since $\mathbb{E}(\widetilde{K}_c^{ij}(\overline{d})\widehat{K}^{ij}(\overline{d})) > 0$ by construction, our coefficients will be biased if we do not control for the counterfactual distribution. The true model would be

$$\widehat{K}_c^{ij}(\overline{d}) = \alpha + X^{ij}\beta + \delta_c^i + \delta_c^j + \left[\widetilde{K}_c^{ij}(\overline{d}) + \varepsilon_c^{ij}\right], \tag{2.4}$$

a classic case of omitted variables.[16]

---

15. This fundamental problem is related to the classical question in spatial economics of what the observed colocation patterns of groups i and j would be in a world where the two groups make independent random choices conditional on their set of feasible choices (see, e.g., Ellison et Glaeser, 1997; Ellison *et al.*, 2010). This problem has been emphasized in the literature measuring the coagglomeration of industries, and various strategies have been put forth to construct counterfactual distributions that only depend on 'locational fundamentals' of the industries (e.g., resource endowments, or access to waterways or the sea; see Ellison et Glaeser 1999, Klier et McMillen 2008, Carillo et Rothbaum 2016, and Behrens et Moussouni 2018 for different ways of constructing counterfactual spatial distributions). To fix ideas—and to illustrate the concept of spurious coagglomeration patterns—consider the colocation of the 'shipbuilding' and 'seafood processing' industries in Canada. These industries are highly colocated, yet they have little interactions with each other in terms of buyer-supplier links, the hiring or exchange of similar workers, or the transmission of knowledge and ideas. These two industries just happen to be in the same place since the set of feasible locations they can choose from overlaps substantially : both need access to the sea, but conditional on that they want to be neither close to each other nor far from each other. Hence, finding them together does not carry much information on interactions between them.

16. Observe that if all groups had a priori the same choice set—namely, all DA in the city—then $\widetilde{K}_c^{ij}(\overline{d}) = \widetilde{K}_c(\overline{d})$ would not vary significantly across groups if they made the same independent random choices and it would be absorbed by the constant term. This is, however, unlikely to be the case. We also have to assume that groups i and j have 'sufficiently large choice sets'. Assume, on the contrary, that the choice sets of groups i and j are just the ones they have actually choosen (i.e., the observed distribution is the only possible one given their choice set). Then, $\widetilde{K}_c^{ij}(\overline{d}) = \widehat{K}_c^{ij}(\overline{d})$ coincide, and the coefficents for our variables of interest would not be identified (of course, we can still estimate something since we do not observe $\widetilde{K}_c^{ij}(\overline{d})$, but it is

Ideally, we need a good proxy for the feasible location sets $\widetilde{K}_c^{ij}(\overline{d})$. Yet, such proxies are very hard to construct at the DA level. Indeed, the relevant characteristics that we have access to are themselves likely to be endogenous to location choices (e.g., if an ethnic group is poorer, it may not maintain the housing stock as well as richer groups, but then using the quality of housing as a determinant would be unwarranted ; also other important determinants of the choice sets—e.g., social networks and discrimination in the housing market—are clearly highly endogenous). We hence have no good benchmark distribution of the coagglomeration we should expect if groups picked random locations among their feasible location sets.

We will use three characteristics of our data to partly deal with that problem : income and tenure status for housing, and restrictions to subgroups that we know are likely to face substantial discrimination in the housing market (namely, groups from Africa). In both cases, the underlying idea is to focus on groups that have more restricted location choices in the city. Hence, if conditional on those more restricted location choices we observe the same relations between colocation patterns and measures of similarity, this means that the former are not driven exclusively by restrictions in spatial choice sets.

Concerning income and tenure status, we split our DA into poor DA and rich DA, based on the DA per capita income *across all groups in the* DA. We take the bottom quartile of the per capita income distribution by DA in each city c and refer to

---

hard to interpret the results in that case). In a nutshell, the identifying assumptions we have to make are the following : (i) groups i and j have sufficiently large choice sets, so that observing their actual pattern represents just one possible outcome compared to a random location within their choice sets ; and (ii) the unobserved counterfactual benchmark $\widetilde{K}_c^{ij}(\overline{d})$ that would prevail in the presence of a random allocation within the set of feasible choices is not systematically correlated with our explanatory variables. These two conditions are hard to verify empirically.

it as the poor DA. Conversely, we take the top quartile of the per capita income distribution in each city c and refer to it as the rich DA.[17] The logic of splitting along those lines is that if some ethnic groups must predominantly pick from 'poor DA'—but are otherwise not likely to colocate—then looking at their pattern for the whole city might be dominated by the colocation driven by that in the poor DA (which are spatially concentrated) ; whereas looking only at the poor DA might reveal a pattern that is closer to randomness (since the poor can pick a priori any location among poor DA). In a nutshell, the assumption underlying this reasoning is that looking at the patterns of colocation among poor areas controls for the fact that the choice set of poor people is mostly restricted to poor places. If we see a lot of sorting based on non-geographic characteristics conditional on being in poor locations, this implies that the patterns pick up real effects that are not solely driven by geographic patterns in choice sets. We can apply a similar logic to split samples along another line : renters vs owners. The majority of renters are constrained to locations where rentals are available, whereas owners are a priori less constrained. Again, if the rental market is highly concentrated (e.g., the inner city), whereas the owner market is more dispersed (e.g., the suburbs), this could imply spurious patterns. Analogously to the distinction between rich and poor, we split the DA in the city into 'renter' DA (the top quartile in the distribution of DA rental property shares in the city), and 'owner' DA (the bottom quartile in that distribution). The effects we estimate on the more restricted choice set (renters)

17. We do not observe income by ethnic group. Yet, since there is a lot of sorting by income in cities, ethnic groups in rich DA are also likely to be rich ; whereas ethnic groups in poor DA are also likely to be poor.

are again more likely to be informative of the true effects we are looking for. [18]

The logic underlying the analysis of groups that a priori are more likely to face discrimination in the housing market is similar. Assume that people from Africa face either more discrimination because of the color of their skin or because of their religion. Then, looking only at the colocation patterns of those groups, we should not see any effect of similarity on geography anymore if there is no homophily. To summarize, focusing on poor people, renters, and minorities is likely to tell as more as to the importance of homophily. Indeed, what we basically observe in the data is a spatial configuration at a given point in time. Hence, we cannot assess how this configuration has been established in the first place. As discussed in the literature, there are three broad reasons behind segregation along racial or ethnic lines. First, immigrants may prefer to live among people of their own ethnic group, thereby creating ethnic enclaves. This is the mechanism we are interested in. Second, natives may want to avoid immigrants—e.g., White flight or collective action racism—thereby also creating enclaves (see, e.g., Cutler *et al.* 1999 for a test on discrimination vs self-segregation). Last, income sorting (Bayer *et al.*, 2004) may also lead to segregation. Focusing on colocation patterns of groups that face more discrimination or controlling (at least partly) for sorting along income helps us in being confident that we capture mostly the first mechanism. If observed patterns were due exclusively to White flight or sorting along income—without any consideration of 'preference for own type'—then we should not observe colocation patterns that reflect similarity among either poor groups or groups that face potentially more discrimination.

---

18. Alternative potentially informative sample splits would be in terms of housing consumption (apartments vs detached or semi-detached units), or in terms of occupations and jobs. Unfortunately, we do not have those data for our small geographic units.

2.4 Results

Tables 2.1 and 2.2 summarize descriptive results for the geographic colocation patterns of the groups within our cities. As shown in panel (a), groups from the same continent tend—as expected—to colocate more. This effect seems especially strong for groups from Africa and weaker for groups from Asia and Europe, as shown by panels (b) and (c). Note also that groups from Europe are the least coagglomerated with groups from other countries, but this effect is likely to be partly mechanical since larger groups tend to appear less coagglomerated with other groups. As explained before, we will control for these effects by including group-city fixed effects in all our subsequent regressions. Panel (d) of Table 2.1 finally shows that groups that immigrated more to Canada after 2002 tend to be slightly more colocated in the cities. This may be due to the dynamics of the housing market, which has become tighter in the 2000s. If groups of immigrants that arrive massively at the same time are constrained to locate together in areas where housing is available at that time, this may also lead to higher degrees of colocation if there are strong patterns in where housing is available. We will control for that aspect of simultaneity in arrival later.

Which pairs are the most coagglomerated in Canadian cities ? Table 2.2 list the top-10 most coagglomerated groups on average across our six metropolitan areas. As shown, and consistent with the descriptives summarized in Table 2.1, it is mostly couples of African countries that top the list. The only other couple is Bhutan and Nepal, two Asian countries that are geographically and culturally close. These results already suggests that geographic proximity needs to be controlled for in our analysis, and that 'culturally similar' countries also tend to have more colocated populations. Observe also that it is hard to know at this stage why pairs of groups from Africa tend to be usually more strongly colocated than other pairs.

TABLE 2.1: Coagglomeration measures by continents and timing of arrival, 2016 census.

	# of pairs	Mean CDF	Stdev. CDF	Min	Max
		(a) Aggregate results			
All	83,365	0.0091	0.0041	0.0005	0.0549
All same continent	19,410	0.0096	0.0045	0.0008	0.0461
All different continent	63,955	0.0090	0.0039	0.0005	0.0549
		(b) Same continent			
Africa-Africa	5,522	0.0126	0.0049	0.0020	0.0409
Pacific-Pacific	60	0.0110	0.0038	0.0039	0.0196
America-America	3,964	0.0096	0.0038	0.0009	0.0251
Asia-Asia	5,418	0.0087	0.0039	0.0008	0.0461
Europe-Europe	4,446	0.0069	0.0029	0.0017	0.0266
		(c) Different continents			
Africa-America	9,586	0.0107	0.0041	0.0015	0.0349
America-Pacific	1,105	0.0102	0.0033	0.0024	0.0246
Asia-Africa	11,180	0.0100	0.0041	0.0012	0.0360
Asia-Pacific	1,290	0.0096	0.0037	0.0017	0.0277
Africa-Pacific	1,290	0.0096	0.0037	0.0017	0.0277
Europe-Pacific	1,170	0.0089	0.0033	0.0025	0.0253
Asia-America	9,503	0.0088	0.0038	0.0005	0.0288
Europe-Africa	10,140	0.0086	0.0037	0.0015	0.0549
Europe-America	8,619	0.0078	0.0033	0.0009	0.0294
Europe-Asia	10,062	0.0072	0.0033	0.0010	0.0304
		(d) Timing of arrival			
Both mainly pre-2002	16,950	0.0083	0.0037	0.0008	0.0279
Both mainly post-2002	24,868	0.0099	0.0044	0.0005	0.0461

Notes : We report simple (unweighted) averages across groups. The variable is the cumulative distribution function (CDF) of the Duranton-Overman K-densities computed city-by-city at a distance of 500 meters. Panel (b) reports all pairs where both countries belong to the same continent, while panel (c) does the same for pairs belonging to different continents. Panel (d) reports results by timing of arrival. Groups are split by couples where both arrive 'early' (i.e., pre-2002 in our sample, which is the median population-weighted arrival year) and couples where both arrive 'late' (i.e., post-2002 in our sample).

This could be due to homophily, but also to a variety of other causes—such as discrimination in the housing market—as explained before.

TABLE 2.2: Top-10 colocated groups represented in more than 20 DA on average across cities.

Country i	Country j	Avg. K-density CDF	Avg. #DAi	Average #DA j
Mauritania	Niger	0.0271	28.31	28.31
Bhutan	Nepal	0.0239	175.22	250.26
Guinea-Bissau	Mauritania	0.0238	39.64	28.18
Guinea-Bissau	Niger	0.0238	39.64	28.31
Gambia	Guinea-Bissau	0.0205	27.84	39.82
Mauritania	Chad	0.0204	28.31	44.00
Niger	Chad	0.0203	28.08	44.00
Gambia	Mauritania	0.0203	27.84	28.18
Gambia	Chad	0.0200	27.84	44.00
Guinea-Bissau	Chad	0.0199	39.64	44.00

Notes : Avg. #DA i and Avg. #DA j are the average number of DA with positive population in that group across the six metropolitan areas. The variable is the cumulative distribution function (CDF) of the Duranton-Overman K-densities computed city-by-city at a distance of 500 meters.

2.4.1 Baseline results

We now present our baseline empirical findings. We provide results for the 2016 Census and for a distance of 500 meters. Results for distances of 100 meters or 1 kilometer, as well as for the 2006 Census, are fairly similar and mostly relegated to Appendix C and to the supplemental online appendix. To get a first idea of how the different variables affect the tendency of groups to colocate, we start by running univariate regressions of each variable separately on our K-densities, including a full set of country-city fixed effects and clustering the standard errors by ij pairs. The results are summarized in Table 2.3.

Table 2.3 shows that all coefficients are precisely estimated and have the expected

sign. Starting with geography, both contiguity and being on the same continent have a positive and significant effect on colocation patterns in Canadian cities. While this is expected, it does not tell us much about why geographic proximity of the countries leads to more colocation in Canada. Next, the economic variables (Common currency, Free trade agreement, Both OECD, GDP per capita gap, Bilateral trade flows, and Bilateral tourism flows) also have the expected effects. Sharing a common currency, being both OECD members, having free trade agreegments, and having larger bilateral exchanges of goods and people all are associated with more colocation. This suggests that people who are from countries that are economically close also tend to colocate more. Again, it is not clear why this should be the case. We thus turn next to what we think are the 'deep roots' of homophily : language, religion, culture, genetics, and historico-political relationships. As shown, people from countries that were in past colonial relationships colocate more. So do people from countries that share a common official language or that share religions. All these aspects of language, culture, and religion can be broadly subsumed by genetic distance which, as shown by the last line of Table 2.3, has a strong negative effect on colocation patterns : ethnic groups that are genetically more distant tend to colocate less. [19]

We next include all variables jointly into our baseline specification. Table 2.4 summarizes our results, where we progressively add the economic, historico-political, cultural, and genetic variables to our basic geographic variables. As Table 2.4 shows, the coefficients on the geographic variables progressively decrease as we add our economic, linguistic, historic, religious, and genetic variables. As expec-

---

19. The large number of fixed effects explains the bulk of the R^2, i.e., there is a lot of idiosyncrasy in the data. Nevertheless, we can identify statistically strong effects of our main variables on colocation patterns, even with that large number of fixed effects and conservative standard errors.

TABLE 2.3: Univariate baseline results, 2016 Census.

Dependent variable : $\widehat{K}_c^{ij}(500m)$	Coeff.		R^2	N
Contiguity	0.05^a	(0.00)	0.86	68,055
Same continent	0.07^a	(0.00)	0.86	68,055
Common currency	0.05^a	(0.00)	0.86	68,055
Free trade aggrement	0.07^a	(0.00)	0.86	68,055
Both OECD	0.09^a	(0.00)	0.86	68,055
Bilateral trade flows	0.03^a	(0.01)	0.86	64,509
Bilateral tourist flows	0.03^a	(0.01)	0.86	66,400
GDP per capita gap	-0.12^a	(0.00)	0.86	67,153
Were same country	0.04^a	(0.00)	0.86	68,055
Common colonizer	0.05^a	(0.00)	0.86	68,055
Colonial relationship	0.01^a	(0.00)	0.85	68,055
Common official language	0.05^a	(0.00)	0.86	68,055
Common religion	0.04^a	(0.00)	0.86	68,055
Genetic distance (alelle, plurality groups)	-0.07^a	(0.00)	0.86	68,055

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. ^{a}p$<$0.01, ^{b}p$<$0.05, ^{c}p$<$0.1.

TABLE 2.4: Multivariate baseline results, 2016 Census.

Dependent variable :	$\widehat{K}_c^{ij}(500m)$							EG_c^{ij}		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Contiguity	0.03^a	0.02^a	0.01^a	0.01^a	0.01^a	0.01^b	0.01^a	0.01	0.00^b	0.00^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Same continent	0.07^a	0.04^a	0.04^a	0.04^a	0.03^a	0.03^a	0.03^a	0.01^a	0.01^a	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common currency		0.02^a	0.01^a	0.01^a	0.01^a	-0.01^b	-0.01^a	0.01	-0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Free trade agreement		0.03^a	0.03^a	0.03^a	0.02^a	0.00	0.01^a	0.01^c	0.00^a	0.01^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Both OECD		0.03^a	0.03^a	0.03^a	0.03^a	0.03^a	0.03^a	-0.01^b	0.00^a	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Bilateral trade flows		0.01^a	0.01^a	0.01^a	0.01^a	0.01^a	0.01^a	0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Bilateral tourism flows		-0.01^a	-0.01^a	-0.01^a	-0.01^a	-0.01^a	-0.01^a	-0.00	-0.00 ^{a}	-0.00^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GDP per capita gap		-0.07^a	-0.07^a	-0.07^a	-0.07^a	-0.05^a	-0.05^a	-0.01^a	-0.01^a	-0.01^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Were same country			0.01^a	0.01^a	0.01^a	0.03^b	0.01^b	0.02	0.01^b	0.01^b
			(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.02)	(0.00)	(0.00)
Common colonizer			0.04^a	0.03^a	0.03^a	0.04^a	0.03^a	0.01^c	0.01^a	0.01^a
			(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Colonial relationship			0.01^a	0.00^a	0.00^a	0.00^c	0.00^b	-0.00	-0.00	-0.00
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common official language				0.01^a	0.01^a	0.01^a	0.02^a	-0.00	0.00^b	0.00^a
				(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Common religion				0.01^a	0.01^a	0.01^a	0.02^a	0.01^b	0.00^a	0.01^a
				(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Genetic Distance (alelle, plurality)					-0.04^a	-0.04^a	-0.04^a	-0.01^a	-0.01^a	-0.01^a
					(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Weighted	no	no	no	no	no	yes[1]	yes[2]	no	yes[1]	yes[2]
Fixed effects	ic and jc (country-city) fixed effects									
Country pairs	All pairs included									
Sample size	68,055	62,145	62,145	62,145	62,145	62,145	62,145	62,145	62,145	62,145
R^2	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.06	0.15	0.08

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. ^{a}p<0.01, ^{b}p<0.05, ^{c}p<0.1, [1]population weights, [2]geographic weights.

ted, the coefficients drop from about 0.03 and 0.07 to 0.01 and 0.03. Yet, they remain significant. As can be seen from columns (2)–(5) in Table 2.4, adding all variables reduces their individual effects—because of the correlations among them—yet we still find significant effects for all of them in our full specification in column (5). Same continent, the GDP per capita gap, a past common colonizer, and genetic distance have the largest effect sizes at 0.03. Yet, all other variables—in particular common official language and common religion—remain highly significant too. Tables 2.14 and 2.15 in Appendix C show the same results as Tables 2.3 and 2.4 for the 2006 Census. Our qualitative results are stable across censuses.

2.4.2 Robustness checks

We next run a battery of robustness checks for : (i) the way we measure cultural, linguistic, and genetic distance, as well as the type of historico-political variables that we use ; (ii) the distance at which we evaluate our measure of geographic concentration ; and (iii) where we control for the 'quality' of our K-density estimates, i.e., where we retain the left-hand side variable only for ethnic groups that are present in a sufficiently large number of DA in our cities.

Alternative measures of similarity

How robust are our results to how we measure cultural, linguistic, religious and genetic distance, as well as historico-political factors such as colonial relationships and other ties ? Table 2.5 summarizes results for our baseline specification (5) from Table 2.4 where we use different variables related to cultural, linguistic, genetic and historico-political factors. As shown, our results are very robust across the different specifications. All linguistic distance measures—except the two that are built on language trees—indicate that speaking the same or a close language

increases the geographic colocation measures. All genetic distance measures have a negative sign and are precisely estimated : genetically more distant groups tend to colocate less. Furthermore, we provide results where we replace both language and religion with broader measures of 'cultural proximity' (Euclidian cultural distance measures constructed from the *World Values Survey* ; see Appendix A.2.2 for details). All cultural distance measures are negatively related to colocation patterns : ethnic groups that report being culturally more different tend to colocate less. Finally, as shown, the historico-political variables have a sizeable and lasting effect on colocation patterns. In particular, ethnic groups that are 'siblings' (i.e., that belonged to the same empire or had a common colonizer) tend to colocate more. Yet, the ties dissipate with time, as shown by the highly negative coefficient on the variable 'Number of years since no longer siblings'. This result mimics the one uncovered for trade patterns between countries (see **?**) : the long shadow of history extends to contemporary location patterns.

To summarize, our results are highly robust to how we measure linguistic, religious, cultural, and genetic proximity. They are also robust to different ways of measuring past historico-political relationships between countries.

Distance and alternative colocation measure

We can evaluate our K-density measures at any distance d between 100 meters and 5 kilometers. Our baseline results use 500 meters. How do the results change with smaller or larger distances, respectively ? Table 2.19 in the supplemental online appendix shows results for distances of 100 meters and 1 kilometer. The results are very stable across distances. In a nutshell, the distance threshold does not really matter for our analysis. The reason is that the K-densities are cumulative measures and thus are strongly correlated across distances. The relative K-density

TABLE 2.5: Alternative measures of our key variables, 2016 Census.

Description	Stata variable name	Coeff.		Sample size	R^2
Common spoken language	lang_csl	0.014[a]	(0.003)	62,145	0.872
Common native language	lang_cnl	0.004[a]	(0.002)	62,145	0.872
Linguistic proximity (Tree, unadjusted)	lang_prox1	0.009[a]	(0.002)	62,145	0.872
Linguistic proximity (Tree, adjusted)	lang_lp1	0.009[a]	(0.002)	57,635	0.875
Linguistic proximity (ASJP, unadjusted)	lang_prox2	0.007[a]	(0.002)	62,145	0.872
Linguistic proximity (ASJP, adjusted)	lang_lp2	0.006[b]	(0.002)	57,635	0.875
Common Language Index (log specification)	lang_cl	0.014[a]	(0.003)	57,635	0.875
Common Language Index (level specification)	lang_cle	0.012[a]	(0.003)	62,145	0.872
Common official or primary language	lang_comlang_off	0.012[a]	(0.003)	62,145	0.872
Language is spoken by at least 9 % of the population	lang_comlang_ethno	0.006[b]	(0.003)	62,145	0.872
Linguistic distance (words, plurality languages)	lang_cognate_dominant	-0.008[a]	(0.004)	14,748	0.904
Linguistic distance (words, weighted)	lang_cognate_weighted	-0.012[b]	(0.005)	7,760	0.931
Linguistic distance (trees, plurality languages)	lang_lingdist_dom_formula	0.004[c]	(0.002)	52,073	0.866
Linguistic distance (trees, weighted)	lang_lingdist_weighted_formula	0.003	(0.002)	52,073	0.866
Genetic distance (microsatellite variation, weighted)	gent_new_gendist_weighted	-0.058[a]	(0.004)	57,805	0.871
Genetic distance (microsatellite variation, plurality groups)	gent_new_gendist_plurality	-0.053[a]	(0.004)	57,805	0.871
Genetic distance (alelle, weighted)	gent_fst_distance_weighted	-0.043[a]	(0.003)	59,462	0.871
Euclidian cultural distance, all categories	cult_total	-0.032[a]	(0.006)	13,674	0.922
Euclidian cultural distance, category A only	cult_total_a	-0.020[a]	(0.005)	13,674	0.922
Euclidian cultural distance, category C only	cult_total_c	-0.014[a]	(0.005)	13,674	0.921
Euclidian cultural distance, category D only	cult_total_d	-0.014[a]	(0.005)	13,674	0.921
Euclidian cultural distance, category E only	cult_total_e	-0.019[a]	(0.006)	13,674	0.922
Euclidian cultural distance, category F only	cult_total_f	-0.007[a]	(0.004)	13,674	0.921
Euclidian cultural distance, binary choice questions only	cult_total_binary	-0.019[a]	(0.005)	13,674	0.922
Euclidian cultural distance, non-binary choice questions only	cult_total_non_binary	-0.027[a]	(0.006)	13,674	0.922
Country was post-45 colonizer of the other	poli_col45	-0.000	(0.001)	62,145	0.871
Countries in the same 'empire' or had common colonizer	poli_sibling	0.017[a]	(0.003)	62,145	0.871
Hegemony relationship	poli_heg	0.003[a]	(0.002)	62,145	0.871
Number of years since no longer siblings (cond. on sibling = 1)	poli_nb_years_sev	-0.035[a]	(0.011)	10,871	0.896
Common legal origins pre-independence	poli_comleg_pre	0.023[a]	(0.002)	62,145	0.872
Common legal origins post-independence	poli_comleg_post	0.014[a]	(0.002)	62,145	0.871
Common legal origins across countries changed	poli_comleg_change	-0.004[a]	(0.003)	62,145	0.871
Religious distance (plurality Fearon et al.)	cult_reldist_dominant_formula	-0.007[b]	(0.003)	51,594	0.866
Religious distance (weighted, Fearon et al.)	cult_reldist_weighted_formula	-0.011[a]	(0.003)	51,594	0.866
Religious distance (plurality, WCD)	cult_reldist_dominant_WCD_form	-0.013[a]	(0.003)	59,532	0.872
Religious distance (weighted, WCD)	cult_reldist_weighted_WCD_form	-0.017[a]	(0.004)	59,532	0.872

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. The specification that we use is (6) in all regressions, with only the language, religion, culture, politics or genetics variable changed. We replace variables as follows in the different regressions : (i) Language : We drop 'common official language' and we replace with the new language variable; (ii) Genetics : We replace the genetics variable with the new genetics variable; (iii) Culture : We replace both language and religion with the cultural variables; (iv) Historico-political : We replace 'common colonizer' and 'colonial relationship' with the new variables; and (v) Religion : We replace 'common religion' with the new religion variable. [a]p<0.01, [b]p<0.05, [c]p<0.1.

CDFs across groups (which pick up most of our identifying variation, recall that we have city-country fixed effects) are fairly stable across distances. We hence stick with a 500 meters distance measure in what follows.

We next check the robustness of our baseline results using an alternative measure for the colocation of ethnic groups. More precisely, we use the measure proposed by Ellison *et al.* (2010), given by :

$$EG_c^{ij} = \frac{\sum_m (s_{m,c}^i - x_{m,c})(s_{m,c}^j - x_{m,c})}{1 - \sum_m (x_{m,c})^2}, \tag{2.5}$$

where $s_{m,c}^i$ is the share of group i in city c located in the DA m; and where $x_{m,c}$ is the share of city c population (all groups) in DA m. Observe that the measure (3.4) can be viewed as a 'spatial covariance' that corrects for the granularity in the distribution of population across dissemination areas. However, as is well known, this measure is aspatial in the sense that any random permutation of the spatial units across the cities will not change its value. Put differently, the relative position of the dissemination areas does not matter.

Columns (8)–(10) of Table 2.4 summarize our results. As shown, the coefficients are smaller using the EG index, but the qualitative patterns are fairly similar. In particular, common religion, genetic distance, and common colonizer have the same impact and are precisely estimated. As can be further seen from Table 2.4, the R^2 drops substantially when using the EG index as the dependent variable. The main reason for this is that the EG index loses the spatial patterns across DA in the data, whereas the DO index captures these. In any case, irrespective of whether we measure the colocation of ethnic groups using the EG or the DO index, we uncover evidence for homophily from the colocation patterns.

Sample size for K-density estimation

Until now, we have included all pairs ij for all cities c into our regressions, even those for which we have only few DA in each city to estimate the K-densities. Since the K-density estimation is less precise for smaller samples (i.e., for ethnic groups present in fewer DA in the city), we replicate our main results by excluding 'small ethnic groups' as follows.[20] We compute the distribution of the number of DA with non-zero presence of each ethnic group i. Then, we drop the bottom quartile of that distribution, i.e., we only keep the K-density estimates for the pairs ij where both groups i and j are not in the bottom quartile of the distribution.[21] In doing so, we exclude the small groups for which the K-densities are estimated on a small number of DA and, therefore, are arguably less precisely measured. Table 2.6 summarizes our results. As shown, they change little compared to the baseline results in Table 2.4. Actually, the results in column (5) of Table 2.6 are almost identical to the correponding results in column (5) of Tables 2.4. We further show in Table 2.16 in Appendix C that our results are robust to the use of our alternative measures for linguistic, religious, genetic, and cultural proximity, as well as the

---

20. Figure 2.5 in the supplemental online appendix shows that there are many relatively small ethnic groups in the cities, and that the distribution of groups across DA is skewed : there are many groups that are small in the sense that they are only present in a small number of DA in each city. This may pose problems for the reliability of our measures of geographic concentration (2.2).

21. We take the distribution across *all* cities and drop the bottom quartile. This has the downside of introducing selective trimming across cities—smaller cities will also be disproportionately represented in the bottom of the distribution. However, using a city-specific threshold—e.g., the bottom quartile in each city—would imply that we still have many less precisely measured K-densities in the smaller cities, whereas we trim away more precise estimates in the larger cities. There is no optimal solution, and results change little with the choice that we make.

other historico-political variables. While there are some minor changes for the historico-political variables, the effects of language, religion, culture, and genetics remain very stable. Last, columns (6) and (7) of Table 2.4 provide estimates for all pairs, where we weight pairs by either their population size in the city or by the number of DA in which they are present. The results from the weighted regressions are close to the unweighted ones. The same holds for columns (6) and (7) of Table 2.6.

Timing of arrival

There are immigration 'waves' and the broad geographic origins of immigrants change over time (e.g., shifting from Europe to Asia). Hence, the simultaneous arrival of different groups may lead to their colocation in specific parts of the city depending on the available housing supply at their time of arrival. To control for this, we use immigration data by country of origin between 1980 and 2018.[22] As shown in panel (d) of Table 2.1, there is some evidence that groups that arrive both 'early' (i.e., pre 2002 in our sample, which is the median population-weighted arrival year) are less colocated than groups that arrive both 'recently' (i.e., post 2002 in our sample). In the former case, the average K-density at 500 meters is 0.008, whereas in the latter case it is 0.010.

To control for potential 'timing of arrival'-effects, we compute, for each pair i and j, the time correlation of the arrival of populations in those two groups, and we

22. Unfortunately, we do not have detailed immigration data for all countries going back in time more than 1980. Technically, we could go back to 1967, but this would imply to digitize archived paper files or extract data from old (scanned) pdf documents. Furthermore, the coverage in terms of countries of origin is substantially sparser. We do not think that this adds much to the analysis and thus have not done it.

TABLE 2.6: Multivariate results, 'high quality' K-densities only, 2016 Census.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Contiguity	0.03^{a}	0.02^{a}	0.01^{a}	0.01^{a}	0.01^{a}	0.01^{b}	0.01^{a}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Continent	0.07^{a}	0.04^{a}	0.04^{a}	0.04^{a}	0.03^{a}	0.04^{a}	0.04^{a}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common currency		0.00	-0.00	-0.00	-0.00	-0.01^{a}	-0.01^{a}
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Free trade aggrement		0.04^{a}	0.03^{a}	0.03^{a}	0.03^{a}	0.01^{b}	0.01^{a}
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Both OECD		0.03^{a}	0.03^{a}	0.03^{a}	0.03^{a}	0.02^{a}	0.02^{a}
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Trade flows		0.01^{a}	0.00^{a}	0.00^{a}	0.00^{a}	0.00^{a}	0.00^{a}
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Tourism flows		-0.01^{a}	-0.01^{a}	-0.01^{a}	-0.01^{a}	-0.01^{a}	-0.01^{a}
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GDP per capita gap		-0.06^{a}	-0.06^{a}	-0.06^{a}	-0.06^{a}	-0.06^{a}	-0.05^{a}
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Were same country			0.02^{a}	0.01^{a}	0.01^{a}	0.03^{b}	0.01^{b}
			(0.00)	(0.00)	(0.00)	(0.01)	(0.01)
Common colonizer			0.04^{a}	0.04^{a}	0.04^{a}	0.05^{a}	0.04^{a}
			(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Colonial relationship			0.01^{a}	0.00^{b}	0.00^{b}	0.00	0.00^{c}
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common official language				0.02^{a}	0.02^{a}	0.02^{a}	0.02^{a}
				(0.00)	(0.00)	(0.00)	(0.00)
Common religion				0.01^{a}	0.02^{a}	0.01^{a}	0.02^{a}
				(0.00)	(0.00)	(0.00)	(0.00)
Genetic Distance (allele, plurality groups)					-0.03^{a}	-0.03^{a}	-0.03^{a}
					(0.00)	(0.01)	(0.00)
Weighted	no	no	no	no	no	yes[1]	yes[2]
Fixed effects	ic and jc (country-city) fixed effects						
Country pairs	Only pairs ij in the top-75%.						
Sample size	38,715	35,883	35,883	35,883	35,883	35,883	35,883
R^2	0.81	0.82	0.83	0.83	0.83	0.86	0.85

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. [a]p<0.01, [b]p<0.05, [c]p<0.1, [1]population weights, [2]geographic weights.

include that variable as an additional ij control in our regressions. Our results barely change and this correlation is insignificant in all but one specification. Hence, we do not report those results (they are available upon request). In a nutshell, the simultaneity of the arrival of groups does not significantly affect our results.

2.4.3 Estimates on restricted samples

As explained before, one key concern of our analysis is that we do not observe the counterfactual colocation patterns that would prevail if ethnic groups made location choices independent from considerations of homophily within their feasible location sets. Constructing such counterfactual patterns would require both strong assumptions and data that we do not have. We thus proceed differently to indirectly control for that problem. More precisely, we now limit our analysis to subgroups that are likely to be more constrained in their choice sets, either because of financial reasons (poor residents and renters) or because of discrimination in the housing market (people from Africa).

Poor residents and renters

It seems reasonable to assume that the poor are constrained in their location choices : they can only pick locations where housing prices or rents are cheap. The same holds—though less stringently—for renters : it is difficult for broad segments of the population to move from renting to buying, which constrains many renters to pick areas where enough rentals are available. Looking at the colocation patterns generated only within poor or renter dominated areas is thus more informative as to whether or not homophily really matters. The reason is that the poor and renters are relatively unconstrained within poor or renter dominated areas, so that

those zones constitute a better proxy for their feasible choice set.

We do not observe individuals in our data, only dissemination areas. Hence, we have to make assumptions as to what we mean by 'poor' and by 'renter'. We classify DA into 'rich' and 'poor' based on average per capita income in the DA across all inhabitants of the DA. Ideally, we would like to know income by ethnic group and by DA, but this is not available. We thus make the assumption that all groups in a poor DA are poor, which seems reasonable since there is a lot of stratification by income in space and since poor and rich usually do not mix much within small spatial locations. We consider that the bottom quartile of the DA in the city-specific per capita income distribution by DA is 'poor', whereas the top quartile in that distribution is 'rich'. We classify, in the same way, the DA by their shares of tenure status : renter DA are those in the top renter-share quartile of the city-wide distribution, whereas owner DA are in the bottom quartile of that distribution.

Note that the ethnic groups present in poor and rich areas—or in renter vs owner-dominated areas—may vary substantially. To purge potential composition effects, we also present results where we compare estimates for the poor DA and for renter DA with city-wide estimates *restricted to the same sets of ethnic pairs*. In words, we compute results for location patterns in the whole city but only for the ethnic pairs that are also represented in the poor DA. This allows for a cleaner comparison and better isolates the pure effect of the choice set.

Table 2.7 summarizes our results. First, column (1) replicates our baseline results. Second, columns (2) and (3) show results where the colocation K-densities are estimated using only the poor DA in the city. Column (3) is 'restricted to poor', i.e., presents results for all DA in the city but only for the groups that are present in the poor DA. In other words, columns (2) and (3) are computed over different locations

TABLE 2.7: Results for poor and renter DA, 2016 Census.

	(1) All	(2) Poor DAs only	(3) Restricted to poor	(4) Renter DAs only	(5) Restricted to renters
Contiguity	0.01^a	0.01^b	0.01^a	0.01^c	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Continent	0.03^a	0.02^a	0.03^a	0.02^a	0.03^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common currency	0.01^a	0.00	0.01^a	0.01^b	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Free trade aggrement	0.02^a	0.02^a	0.03^a	0.02^a	0.03^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Both OECD	0.03^a	0.02^a	0.03^a	0.01^a	0.03^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Trade flows	0.01^a	0.01^a	0.01^a	0.01^a	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Tourism flows	-0.01^a	-0.00^c	-0.01^a	-0.01^a	-0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GDP per capita gap	-0.07^a	-0.04^a	-0.07^a	-0.07^a	-0.07^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Were same country	0.01^a	0.01^b	0.01^b	0.01^a	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common colonizer	0.03^a	0.02^a	0.03^a	0.02^a	0.03^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Colonial relationship	0.00^a	0.00^a	0.00^a	0.01^a	0.00^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common official language	0.01^a	0.02^a	0.01^a	0.01^a	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common religion	0.01^a	0.02^a	0.01^a	0.02^a	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Genetic Distance (allele, plurality groups)	-0.04^a	-0.02^a	-0.03^a	-0.03^a	-0.03^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Fixed effect	ic and jc (country-city) fixed effects				
Country pairs	All, computed on poor or renter DAs only.				
Sample size	62,145	58,174	58,174	58,939	58,939
R^2	0.87	0.84	0.87	0.80	0.87

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. ^{a}p<0.01, ^{b}p<0.05, ^{c}p<0.1.

but for the same pairs of groups. Comparing columns (2) and (3) provides an idea of how the set of feasible location choices affects the coefficients on our variables of interest. As shown, our results are fairly similar between the two columns, with generally slightly smaller and less precisely estimated effects in column (2) than in column (3). Yet, the coefficients on language and religion are slightly larger and precisely estimated in column (2), thus suggesting that colocation by language and religion is not spuriously driven by location choice sets and may be especially valued by lower-income residents. Results for renters (see columns (4) and (5)) are fairly similar. Again, our main effects are robust to estimates on a restricted sample.

Rich residents and owners

Along the same lines as for poor and renters, we can provide estimates for the rich and owners. These categories of residents may have different preferences.[23] Owners, for example, make longer term decisions than renters. Thus, they could be more 'picky' when choosing their neighbors and thus more sensitive to the 'deep roots' of homophily. Also, rich residents face different constraints with respect to location choices. One of them is housing quality, and high-quality housing is

23. Differences in coefficients may reflect heterogeneity in 'tastes', i.e., some attributes may be valued differently by rich and poor or by renters and owners. We know from previous research that owners put more weight on neighbors' characteristics than renters since they stay longer in the same location and are thus more likely to sort. The same may hold for the rich, who sort on income, educational attainment, school quality or other neighborhood characteristics that may be important for peer effects (see, e.g., Nechyba, 2006). It is thus not clear that if we find, e.g., a larger effect of 'common official language' on the colocation patterns of the poor, that this reflects the desire of poor to be closer to groups with a similar linguistic background or that the location sets of the poor are more restricted. We cannot separate the two effects, so some caution is in order.

unevenly distributed across cities. Furthermore, they are known to be sensitive to school quality and the potential for peer effects (either for themselves or for their children). In a nutshell, the rich and owners may value differently the ethnic composition of their neighborhood. Table 2.17 in Appendix C shows our results, along the same lines as Table 2.7. The results for the rich in columns (2) and (3) are fairly similar, thus suggesting again that the effects are unlikely to be driven by strong geographic patterns in choice sets. The results for owners in columns (4) and (5) are interesting. The coefficients on geographic contiguity and genetic distance increase, thus suggesting that there is slightly more stratification along those lines for owners than for the population in general. Since the effect sizes are, however, fairly similar across all specifications, we do not want to read too much out of this.

Potential discrimination in the housing market

As a third exercise, we replicate our analysis to see if the measured effects of linguistic, religious, and genetic similarity vanish once we look at the colocation patterns of groups that are likely to face substantial discrimination in the housing market. To this end, we estimate separate effects for pairs ij that originate both from Africa. These populations are likely to face discrimination based on either the color of their skin or their religion.[24] Table 2.8 and Figure 2.1 show that there are indeed Africa-specific effects.

As shown, especially common religion and the variables related to the colonial past have strong effects for pairs from Africa. As to common official language

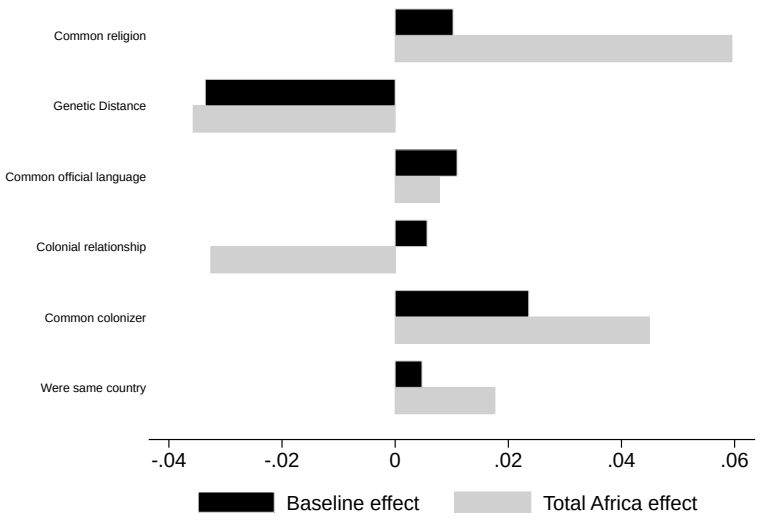
24. We also estimated the specification for pairs ij that originate both from Asia. The results are similar, except that religion matters less whereas language matters more for these couples. In any case, our results suggest that the effects do not vanish when looking at these groups.

TABLE 2.8: Are there Africa-specific effects ?

	Coefficient	Std dev.	Total Africa effect
Common official language	0.0109^a	(0.0027)	0.0078
Common religion	0.0102^a	(0.0024)	0.0595
Genetic distance (alelle, plurality groups)	-0.0336^a	(0.0029)	-0.0357
Were same country	0.0047	(0.0036)	0.0176
Common colonizer	0.0236^a	(0.0029)	0.0449
Colonial relationship	0.0056^a	(0.0016)	-0.0326
Both Africa	-0.0032	(0.0229)	
Common official language × Both Africa	-0.0031	(0.0108)	
Common religion × Both Africa	0.0493^a	(0.0143)	
Genetic distance (alelle, plurality groups) × Both Africa	-0.0021	(0.0119)	
Were same country × Both Africa	0.0129^c	(0.0077)	
Common colonizer × Both Africa	0.0213^a	(0.0095)	
Colonial relationship × Both Africa	-0.0382^a	(0.0058)	

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. We impose common coefficients for all variables except the ones that we interact with a 'Both Africa'-dummy. The latter takes value 1 if i and j are African countries, and zero otherwise. ^{a}p<0.01. ^{b}p<0.05. ^{c}p<0.1.

FIGURE 2.1: Pairs from Africa display at least as much homophily than the other pairs.



Notes : The black bars are the baseline effects, whereas the grey bars are the 'Total Africa effect' (sum of the baseline plus the interaction).

and genetic distance, while there is no specific effect for Africa, the effect also does not disappear : pairs from Africa have a positive coefficient for common official language and genetic distance, and that effect is not significantly different from that of the other ethnic pairs. In a nutshell, even if groups from Africa face discrimination in the housing market and are constrained as to where they can locate, conditional on their choice sets they still sort in a way such that religious, linguistic, and genetic similarity—as well as common history—matter. These results strengthen our view that we pick up real effects and not just spurious colocation patterns driven by income sorting or discrimination.

2.4.4 Extensions : Heterogeneity by city and mean reversion

Heterogeneity across cities

Until now, we have considered common coefficients across all cities. Yet, there may be heterogeneous effects across cities. First, historic immigration patterns differ across cities in Canada. Thus, language may be more important in some cities whereas religion may be more important in others. Second, institutional settings related to housing and immigration differ somewhat across Canada, which may have a direct effect on differential colocation patterns in cities. For example, language is traditionally a thornier issue in the east than in the west. Thus, eastern cities may see more stratification along linguistic divides than western cities.

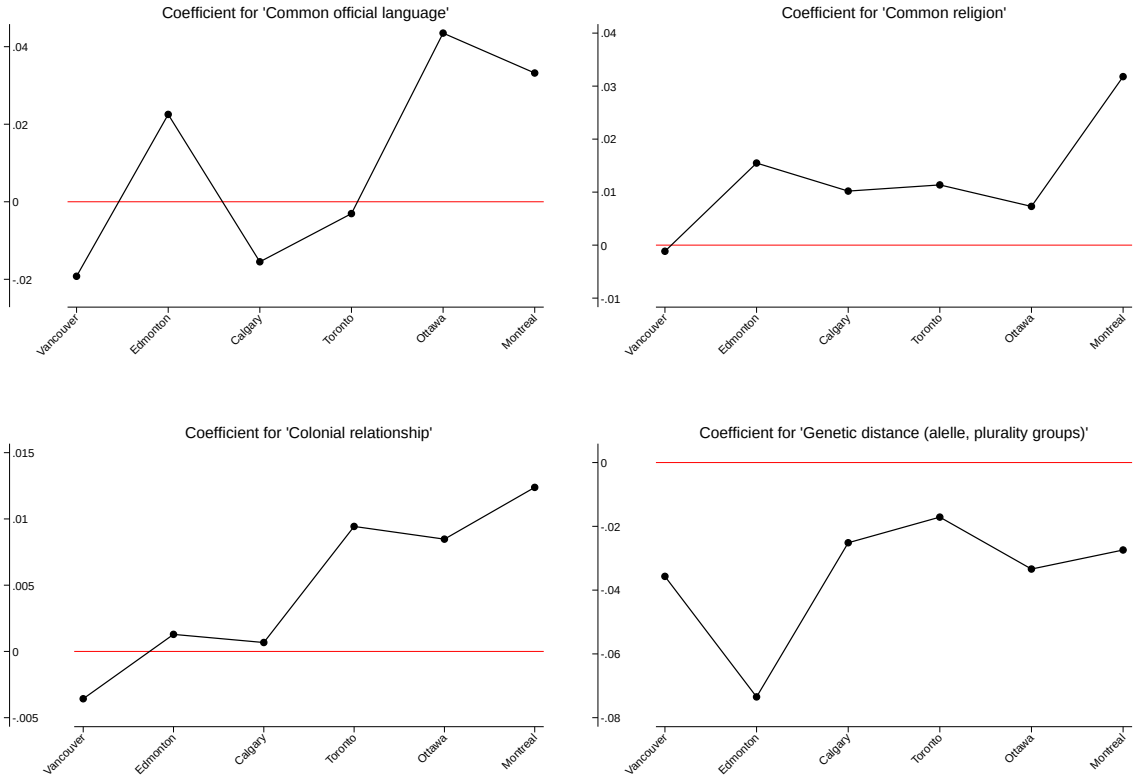
To look for heterogeneous effects, we estimate (2.3) by allowing some of our key coefficients of interest to vary between cities. This allows us to see if there are substantive differences in the role of language, religion, history, or genetics between cities when it comes to the choice of neighbors. We interact our variables of interest—one-by-one—with a city dummy, while keeping common coefficients for the other variables. [25]

Table 2.9 and Figure 2.2 show that, as expected, language is more important in Montréal and Ottawa. The latter is due to the fact that the Ottawa-Gatineau metropolitan area straddles two provinces with different official languages (French in Québec, and English in Ontario), which leads to more opportunity for sorting along linguistic lines. This effect is, however, not only due to the two-province location. It can also be seen in Montréal, where colocation patterns reflect linguistic

---

25. We also ran the models city-by-city, i.e., letting all coefficients vary by city. Results are available upon request. In that case, we cannot cluster by ij as we only have one observation per pair.

FIGURE 2.2: Heterogeneous effects of language, religion, colonial relationships, and genetics by city.



Notes : See Table 2.9 for detailed results. Standard errors for the city-specific coefficients are also reported in that table. We depict the coefficientsusing all variables and city-interaction effects for our variable of interest.

TABLE 2.9: Heterogeneous effects of language, religion, colonial relationships, and genetics by city.

	Montréal	Ottawa	Toronto	Calgary	Edmonton	Vancouver
	2016 Census					
Common official language	0.033^a	0.044^a	-0.003	-0.015^a	0.023^a	-0.019^a
	(0.004)	(0.005)	(0.003)	(0.005)	(0.006)	(0.005)
Common religion	0.032^a	0.007	0.011^a	0.010^a	0.015^a	-0.001
	(0.004)	(0.005)	(0.003)	(0.005)	(0.005)	(0.004)
Colonial relationship	0.012^a	0.008^a	0.009^a	0.001	0.001	-0.004^c
	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)
Genetic distance (alelle, plurality groups)	-0.027^a	-0.033^a	-0.017^a	-0.025^a	-0.073^a	-0.036^a
	(0.004)	(0.005)	(0.003)	(0.005)	(0.006)	(0.006)

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. We impose common coefficients for all variables except the one that we interact with a city-dummy. ^{a}p<0.01. ^{b}p<0.05. ^{c}p<0.1.

similarity. Generally, the effect of sharing a common official language on colocation patterns is weaker in the west, with the exception of Edmonton where it seems to play a sizable role. Similar as for language, past colonial relationships also display a substantial east-west gradient, being more important for ethnic groups in the east than in the west. Common religion appears the most important in Montréal—home to the largest share of the Jewish community in Canada—displays a fairly flat pattern across the country, and appears the least important for colocation patterns in Vancouver. Last, genetic distance has through the board a negative effect across Canadian cities. The results using 2006 Census data (available upon request) are broadly in line with those using 2016 data though the coefficients are smaller and less precisely estimated since we have fewer ethnic origins reported (see Appendix A.1).

Mean reversion

Finally, we run a first-differenced specification, where we regress the decadal 2006–2016 changes in the colocation measure on the initial values of our explanatory variables, including the 2006 colocation measure. This first-differenced specification is akin to a convergence regression and provides an answer to the question whether groups that are more similar along different dimensions tend to increase or decrease their degree of colocation over the decade, conditional on their initial colocation patterns. Since colocation patterns tend to be relatively stable over time, this is a demanding exercise.

Table 2.10 shows that there is strong mean reversion—the coefficient on the initial level of colocation is negative and large—but that the other coefficients do not change substantially compared to our cross-sectional baselines (reported in columns (1) and (2) of Table 2.10). This suggests that, although the extent of colocation tends to decrease over time for pairs that were initially strongly colocated, it does less so for pairs that are similar in terms of language, culture, religion, genetics, or that share a common history. While these findings suggest that ethnic stratification in Canadian cities has not increased in the last decade—and that there may even be slightly more mixing along some dimensions than ten years ago (see Glaeser et Vigdor 2012 who find that segregation has decreased in U.S. cities after 2000)—they need to be interpreted with caution. Indeed, less coagglomeration between groups i and j could simply mean that there is more concentration within groups i and j.

2.5 Appendix

This set of appendices is structured as follows. Appendix **A** presents additional details and information on our data. Appendix **B** explains in more detail our

TABLE 2.10: Mean reversion regressions, difference 2006–2016 Census.

	(1)	(2)	(3)	(4)	(5)
	Baseline 2016	Baseline 2006	Difference CDF all DAs	Difference CDF poor DAs	Difference CDF renter DAs
Dependent var.	$\widehat{K}_c^{ij}(500m)$	$\widehat{K}_c^{ij}(500m)$	$\Delta\widehat{K}_c^{ij}$, 2016-06	$\Delta\widehat{K}_c^{ij}$, 2016-06	$\Delta\widehat{K}_c^{ij}$, 2016-06
$\widehat{K}_c^{ij}(500m)$, 2006			-1.27^a	-1.24^a	-1.14^a
			(0.02)	(0.01)	(0.01)
Contiguity	0.01^a	0.01	0.01^a	0.01^c	0.00
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Same continent	0.03^a	0.02^a	0.02^a	0.02^a	0.02^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common currency	0.01^a	0.02^c	0.01^b	0.00	0.01^a
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Free trade area	0.02^a	0.02^a	0.01^a	0.01^a	0.02^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Both OECD	0.03^a	0.02^a	0.02^a	0.01^a	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Trade flows	0.01^a	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Tourism flows	-0.01^a	-0.00	-0.01^b	-0.00^c	-0.00^b
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GDP per capita gap	-0.07^a	-0.03^a	-0.06^a	-0.05^a	-0.05^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Were same country	0.01^a	0.02^c	0.01^b	0.01^b	0.01^b
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Common colonizer	0.03^a	0.02^a	0.03^a	0.03^a	0.02^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Colonial relationship	0.00^a	0.00^c	0.01^a	0.01^a	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common official language	0.01^a	0.00	0.01^a	0.02^a	0.01^b
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common religion	0.01^a	0.01^a	0.01^b	0.02^a	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Genetic distance (alelle, plurality groups)	-0.04^a	-0.02^a	-0.03^a	-0.02^a	-0.03^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Fixed effects			ic and jc (country-city) fixed effects		
Country pairs			All, computed on poor or renter DAs only.		
Sample size	62,145	51,820	51,582	42,881	49,502
R^2	0.87	0.78	0.90	0.84	0.90

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. ^{a}p<0.01, ^{b}p<0.05, ^{c}p<0.1.

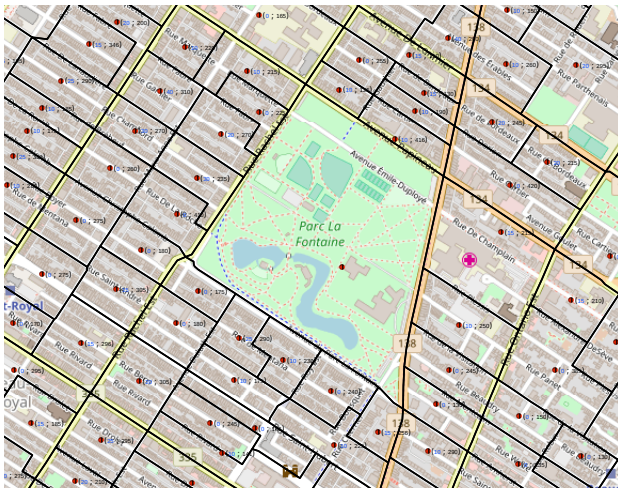
procedure for mapping ethnic groups to countries. Last, Appendix **C** contains additional tables and results.

Appendix A. Addititional information on the data

A.1. Census data

Figure 2.3 illustrates the granularity of our data by depicting the dissemination areas in the area known as 'le plateau' in Montréal. The red dots are the (population-weighted) centroids—as provided by Statistics Canada—and the blue figures next to them report the count of ethnic groups (Belgian and French in our example) living in each DA. These are the data we use to compute our measures of ethnic colocation. Table 2.11 reports summary statistics by city, including population figures and the number of dissemination areas.

FIGURE 2.3: Dissemination areas and centroids in 'le plateau' in Montréal in 2016.



The raw census data encompass a wide range of ethnic groups. In 2016, for example, there were more than 250 ethnic groups in the census, and 50% of the population reported more than one ethnic origin. Although we aggregate the data to the country level, as explained before, thereby losing ethnic diversity, this is still a

fine division along ethnic lines. As expected—besides Canadian—British, French, and other European origins were the most reported. Figure 2.5 and Table 2.21 in the Online Appendix provide summary statistics on the representation of different ethnic groups and their distribution across the DAs in our six metropolitan areas. One thing to notice immediately is that there are many small groups. For these, measures of colocation may be more noisy and we provide robustness checks (either using weights or excluding the small groups) that show that our results are not driven by the small groups.

We use the 2006 and 2016 census waves. Although they are largely comparable, there are some minor differences between the two censuses. First, changes in immigration source countries, the political context, and the increasing diversity of Canada's population have made recent censuses richer in ethnic origins. There are groups in the 2016 census that are not reported in the 2006 census (for example, Arawak, Bavarian, Bhutanese, Catalan, Corsican, Djiboutian, Edo, Ewe, Guadeloupean, Hazara, Karen, Kyrgyz, Malinké, Turkmen and Wolof). Second, the geographical units changed between 2006 and 2016, with slightly more DAs in 2016 than in 2006. While a finer geography makes for more precise estimates of our geographic concentration measures, the changes are marginal at best, especially in the central parts of the cities where there is very little change in the census geography over time.

Note that while the 2006 and 2016 census long-form questionnaires were obtained from a mandatory survey that had a high response rate (94% and 97% for 2006 and 2016, respectively), the 2011 ethnic information was collected from the 2011 National Household Survey (NHS), which is a voluntary survey that replaced the former mandatory 2006 census long-form questionnaire. The NHS sample frame was approximately one-third of all Canadian household, with a lower response rate (68.6%, or around 7 million individual responses). The estimated data, if any,

TABLE 2.11: Summary statistics by city

	Montréal	Ottawa	Toronto	Calgary	Edmonton	Vancouver
	2016					
Population (millions)	4.07	1.31	5.87	1.38	1.3	2.44
# ethnicities (in sample)	153	153	153	152	151	146
# of DAs in our analysis	6,355	1,904	7,293	1,706	1,622	3,381
Average income	85,115	105,530	120,064	144,135	120,920	104,333
# of DAs (poor) in our analysis	1,588	476	1,823	425	405	845
Average income (poor)	47,886	56,940	64,167	76,812	69,109	62,418
	2006					
Population (millions)	3.6	1.11	5.08	1.07	1.02	2.09
# ethnicities (in sample)	142	141	143	133	132	133
# of DAs in our analysis	6026	1,769	6,960	1,572	1,522	3,306
Average income	64,180	83,680	89,755	91,779	79,367	75,750
# of DAs (poor) in our analysis	1,506	442	1,740	393	380	826
Average income (poor)	35,357	42,930	45,279	44,610	43,097	42,456

Notes : This table report the statistics (e.g., # of DAs) only for those units for which we have all the data (e.g., income data from the census). Hence, we drop some DAs from the table.

from the 2011 NHS would be more affected by the response rate than those from the 2006 and 2016 long-form questionnaires. They are also subject to potentially higher non-response error than in the census due to the survey's voluntary nature. Unlike the census, Canadian citizens and landed immigrants living outside the country were excluded from the NHS (collectives, such as hotels, hospitals or work camps, were also excluded). In what follows, we disregard the 2011 NHS and work with the 2006 and 2016 census waves only. Also, location patterns change slowly, so decennial changes seem more appropriate than five year changes to check the robustness of our results and their dynamics over time.

A.2. Other data

This appendix provides additional details on our main data sources and on our key explanatory variables. We spend more time explaining the linguistic and genetic variables as those are conceptually more complex and less widely used. We spend comparatively less time explaining the standard variables of the gravity equations (e.g., distance, trade flows, colonial relationships etc.) since those have been abundantly documented elsewhere (see Head *et al.* 2011 ; Head et Mayer 2014). Table 2.12 provides a full list of the variables that we use, as well as information on where to find additional details. We also provide the name of the Stata variable for the ease of reading the appendices. Red-colored ones are used in the baseline model. Table 2.13 provides the correlations between these variables (which are in red in the table).

A.2.1 Measures of linguistic distance.

Common official (`lang_col`), common native (`lang_cnl`), common spoken language (`lang_csl`), and language index (`lang_cl`, `lang_cle`). Our data come

TABLE 2.12: Summary of the key variables and data sources.

Category	Stata variable names	Appendix
Language	lang_col, lang_cnl, lang_csl	A.1.
Language	lang_prox1, lang_prox2, lang_lp1, lang_lp2	A.1.
Language	lang_lingdist_weighted_formula, lang_lingdist_dom_formula	A.1.
Language	lang_cognate_dominant, lang_cognate_weighted	A.1.
Language	lang_cl, lang_cle, lang_comlang_off, lang_comlang_ethno	A.1.
Religion, culture	cult_comrelig	A.2.
Religion, culture	cult_reldist_dominant_formula, cult_reldist_weighted_formula	A.2.
Religion, culture	cult_reldist_dominant_WCD_form, cult_reldist_weighted_WCD_form	A.2.
Religion, culture	cult_total, cult_total_a, cult_total_c, cult_total_d	A.2.
Religion, culture	cult_total_e, cult_total_f, cult_total_binary, cult_total_non_binary	A.2.
Genetics	gent_new_gendist_weighted, gent_new_gendist_plurality	A.3.
Genetics	gent_fst_distance_dominant, gent_fst_distance_weighted	A.3.
Politico-historic	poli_smctry, poli_comcol, poli_colony	A.4.
Politico-historic	poli_sibling, poli_heg, poli_comleg_pre, poli_comleg_post	A.4.
Politico-historic	poli_col45, poli_nb_years_sev, poli_comleg_change	A.4.
Geographic (controls)	geog_contig, geog_continent	A.5.
Economic (controls)	econ_com_cur, econ_fta, econ_gap_gdpcap_mean, econ_flow_mean	A.5.
Economic (controls)	econ_oecd, econ_tour_mean	A.5.

Notes : Variables included in our baseline specification are highlighted in red. The other variables are used in robustness checks. Details on data sources and construction are provided in Appendix A.

TABLE 2.13: Correlation matrix, controls and key variables.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Geography : contiguity														
2. Geography : same continent	0.2481													
3. Economics : common curreny	0.1312	0.2456												
4. Economics : FTA	0.1877	0.2871	0.2315											
5. Economics : both OECD	0.0657	0.1661	0.0485	-0.0792										
6. Economics : trade flows	0.1772	0.0994	0.1008	0.1026	-0.0101									
7. Economics : tourism	0.3234	0.1314	0.1322	0.1559	0.0272	0.6822								
8. Economics : p.c. GDP gap	-0.0805	-0.1311	-0.0478	0.0377	-0.5233	0.0192	-0.0024							
9. Was same country	0.3598	0.1780	0.1945	0.1841	0.0562	0.0201	0.0491	-0.0660						
10. Common colonizer	0.0601	0.1122	0.1143	0.0193	0.2193	-0.0298	-0.0164	-0.0700	0.1444					
11. Colonial relationship	0.0965	-0.0164	0.0005	0.0675	-0.1132	0.0475	0.1048	0.0433	0.0365	-0.0367				
12. Common official language	0.124	0.1997	0.104	0.0976	0.0784	0.0115	0.0312	-0.016	0.1562	0.3700	0.1602			
13. Common religion	0.1401	0.2177	0.0677	0.1388	0.0499	0.0021	0.0433	-0.0589	0.1038	-0.0041	0.0558	0.2151		
14. Genetic distance	-0.1425	-0.2985	-0.1387	-0.2167	0.035	-0.0782	-0.0981	-0.0673	-0.0867	-0.0225	-0.0300	-0.0654	-0.0278	

from Melitz et Toubal (2014). `lang_col` is a binary variable that takes value 1 if the country pair ij shares the same official language and 0 otherwise. It measures the likelihood that residents from i and j will understand each other. A restrictive definition is that two countries share a common official language when this language is official and formally used in different administrations, schools, and public organizations. In this paper, we use a slightly broader and more liberal definition. `lang_col` can take a value of 1 even when the pair does not share 'officially' same the same language, and it can take value 0 even if it does. For instance, even if country i = Sudan adopted English as an official language since 2005, another country j that has English as an official language will yield $\texttt{lang\_col}_{ij} = 0$ because the decision of Sudan to adopt this language is purely trade-related. It is still unlikely that someone from an officially English-speaking country will understand someone from Sudan. Consequently, `lang_col` can take value 0 even if the two countries share the same official language. Also, countries that had colonial relationships tended to often adopt the language of the colonizer as an official language. After independence, one of the first symbolic decisions was often to reverse this, even though the language remains widely used in official documents and daily life (e.g., French in Morocco, Algeria, and Tunisia). For such pairs, `lang_col` will take a value of 1 since an Algerian, Moroccan, or Tunisian person is likely to easily communicate with other French-speaking persons. Furthermore, some languages can be official in some specific parts of a country only (e.g., German is official in some parts of Denmark and French in some parts of Lebanon). In both case, `lang_col` will equal 1. As a result of this special definition of `lang_col`, there are 19 official languages that are shared by at least one country pair : Arabic, Bulgarian, Chinese, Danish, Dutch, English, French, German, Greek, Italian, Malay, Persians, Portuguese, Romanian, Russian, Spanish, Swahili, Swedish, and Turkish.

Common native language (`lang_cnl`) and common spoken language (`lang_csl`) require that the languages be spoken by at least by 4% of the population of each country in the pair ij, irrespective of the official status of the language. This yields 42 different languages that are shared by country pairs (including the 19 official languages listed above).[26] $\texttt{lang\_cnl}_{ij}$ and $\texttt{lang\_csl}_{ij}$ are then calculated as the probability that two randomly drawn individuals from countries i and j have the same native language or speak the same language.[27]

Finally, we also took an aggregated measure of common language (`lang_cl` and `lang_cle`) that summarize some of the measures cited above and that is used to look at the relation between trade and language. It is a 0–1 common language index that is resting strictly on exogenous linguistic factors (think about potential reverse causality between trade and language), and summarize COL, CNL and LP alone. (see Melitz et Toubal 2014 for more details on these measure and their context to bilateral trade).

Linguistic proximity (`lang_prox1`, `lang_prox2`, `lang_lp1`, and `lang_lp2`). Linguistic proximity measures to the 'closeness' of two different native languages. Two

---

26. The 23 shared languages that are not official in both countries ij are : Albanian, Armenian, Bengali, Bosnian, Croatian, Czech, Fang, Finnish, Fulfulde, Hausa, Hindi, Hungarian, Javanese, Lingala, Nepali, Pashto, Polish, Quechua, Serbian, Tamil, Ukrainian, Urdu, and Uzbek.

27. Formally, for each pair ij we compute $\alpha_{ij} = \sum_{n=1}^{N} L_{ni}L_{nj}$, where L_{ni} and L_{nj} are the shares of people in countries i and j that speak (native or not) language $n = 1, 2, \ldots, N$. As people can speak more than one language, α_{ij} may exceed one. To correct for this problem, an adjusted version of $\texttt{lang\_csl}_{ij}$ (or of $\texttt{lang\_cnl}_{ij}$) is computed for all data using the following formula $\texttt{lang\_csl}_{ij} = \max(\alpha_{ij}) + (\alpha_{ij} - \max(\alpha_{ij}))(1 - \max(\alpha_{ij}))$, where $\max(\alpha_{ij})$ denotes the largest contribution of a given language n to the pair ij. When α_{ij} is greater than 1, $\alpha_{ij} - \max(\alpha_{ij})$ is always smaller than 1, so that $\texttt{lang\_csl}_{ij}$ is adjusted to be smaller than 1.

measures, `lang_prox1` and `lang_prox1`, are used, which range from 0 to 1.[28] They are constructed using the proximity of at most two native languages common to each pair ij. A country that has too high a linguistic diversity—or where the native language is not spoken by the majority—will have a measure equal to 0 in the couple ij. If the pair shares the exact same native language then `lang_lp1` or `lang_lp2` equal 1.[29] Based on the *Ethnologue* data (Lewis, 2009), the measure `lang_lp1` compares languages of different trees, branches, and sub-branches. `lang_lp1` takes lower values if two languages belong to different trees and higher values if they belong to the same sub-branch (see, e.g., Fearon, 2003). There are four possibilities : 0 if the two languages belongs to different trees ; 0.25 if they belong to different branches within a tree ; 0.5 if they belong to the same branch ; and 0.75 if they belong to the same sub-branch. To overcome problematic comparisons between trees, `lang_lp2` uses the *Automated Similarity Judgment Program* (ASJP ; see Brown *et al.* 2008 for more details). ASJP attributes score by comparing and analyzing lexicographic similarities between 100 to 200 words of the two languages. Finally, once billateral proximity measures ranging from 0 to 1 are obtained for all pairs of language, the final step is to convert them to country-pair scores.

Linguistic distances (`lang_lingdist_dom_formula`, `lang_lingdist_weighted_formula`, `lang_ cognate_dominant` and `lang_cognate_weighted`). Our source of data is Spolaore et Wacziarg (2009, 2016). The first measure of linguistic dis-

---

28. To make the two measures coefficient comparable between them and along with `lang_col`, `lang_prox1` and `lang_prox2` are again normalized and noted `lang_lp1` and `lang_lp2`. By doing so, their values now range from 0 to more than 1.

29. In Melitz et Toubal (2014), perfect correspondence is coded as 0, but this is controlled for in the regressions via the inclusion of another variable.

tance is obtained by grouping languages into families, and by looking at their similarities, a concept borrowed from cladistics. It is similar to `lang_lp1` since it is based on tree comparisons, but the measures are structurally different and have a lower correlation (Table 2.20).

Languages which split into other languages over time and variations in common nodes reflect linguistic distances. [30] Once measures for language pairs are obtained, the data has to be mapped to the level of countries. To do so, Fearon (2003) provides information on the prevalence of different languages for a large set of countries. Using this information, two country-level measures are computed. First, an unweighted measure, `lang_lingdist_dom_formula` that takes simply the number of common nodes for two major languages of each country in a pair. Second, a weighted measure where the weights are given by the country's linguistic groups. [31]

The second set of linguistic distance measures that we use, `lang_cognate_dominant` and `lang_cognate_weighted`, is based on Lexicostatistics that classifies languages based on whether the words used do convey some common meaning. Two words can derive from the same ancestor, i.e., they are cognate. Thus, two languages with many cognates are closer. For instance, the words "tavola" in Italian and

30. For instance, Spolaore et Wacziarg (2016, p.11) explain that French and Italian share four nodes since French is classified as Indo-European, Italic, Romance, Italo-Western, Western, Gallo-Iberian, Gallo-Romance, Gallo-Rhaetian, Oil, and Français ; whereas Italian is classified as Indo-European, Italic, Romance, Italo-Western, and Italo-Dalmatian. This makes these languages 'close'.

31. Formally, we compute `lang_lingdist_weighted_formula` $= \sum_{i=1}^{I} \sum_{j=1}^{J} (S_{1i} \times S_{2j} \times c_{ij})$, where S_{1i} and S_{2j} are the shares of linguistic groups i and j in countries 1 and 2 respectively, and where c_{ij} is the number of common nodes between language i and j. Both `lang_lingdist_weighted_formula` and `lang_lingdist_dom_formula` range between 0 to 15, and these measures are then standardized to range from 0 to 1.

"table" in French both stem from the Latin word "tabula" and are, therefore, cognate. Linguistic proximity is measured by the percentage of cognate words between the two languages. In the same way as for `lang_lingdist_dom_formula` and `lang_lingdist_weighted_formula`, a weighted and an unweighted measure are computed. The advantage of the measures based on cognate words is that they are more continuous than those using a cladistic approach. We also add two other variables : a dummy variable equal to one of the language is at least spoken by 9% of the population (`lang_comlang_ethno`) ; and a dummy variable equal to one if the pair shares a common official or primary language (`lang_comlang_off`).

Table 2.20 in the supplemental online appendix provides more detailed correlations within the language measures.

A.2.2. Measures of religious and cultural distance.

Common religion (`cult_comrelig`). This measure comes from Melitz et Toubal (2014). It measures the probability that two people drawn at random from two countries i and j will have the same religion. The measure is constructed using mainly the *CIA World Factbook* that reports population shares for major religions (Buddhist, Christian, Hindu, Jewish, and Muslim) for the different countries of the world. Then, the information is aggregated to the country-pair level, using the same methodology as for the `lang_cnl` measure (i.e., the sum of the products of the population shares, plus the standardization).

Religious distance measures (`cult_reldist_dominant_formula`, `cult_reldist_weighted_formula`, `cult_reldist_dominant_WCD_form` and `cult_reldist_weighted_WCD_form`). These measures are drawn from Spolaore et Wacziarg (2009, 2016). They are computed using a tree-based approach, i.e., religious distance is

reflected by distances between nodes in a tree. One tree comes from Mecham *et al.* (2006) and another tree, less disaggregated, comes from WCD (2007). Both also provide frequency distributions of each religion by country. The religious distance, weighted and unweighted, can be computed in the same way as for `lang_cnl`.

Euclidian cultural distance measures (`cult_total`, `cult_total_`x, `cult_total_binary`, and `cult_total_non_binary`). A second source of cultural data in Spolaore et Wacziarg (2009, 2016) is based on information from the *World Values Survey* (WVS). This survey reports answers to 740 questions about values, norms, and attitudes. The answers are divided into 7 categories, of which 5 are used to construct distance measures ($x = a, c, d, e, f$ in our variable `cult_total_`x) : A : Perception of Life, C : Work, D : Familly, E : Politics and Society, F : Religion and Moral. The Euclidian cultural distance is computed as follows. Consider countries 1 and 2, and some question i that allows for answers $j = 1, 2, \ldots, J$, where J may differ between questions. Let s_{ij}^c denote the share of respondents in country c giving answer j to the question i. If the question has a binary answer then the cultural distance is measures as $C_i^{12} = |S_{i1}^1 - S_{i1}^2|$. If the question has multiple responses, then the distance is $C_i^{12} = \sqrt{\sum_{j=1}^{J}(S_{ij}^1 - S_{ij}^2)^2}$.

One problem with the WVS is that not every question was asked in every country. When calculating the Euclidian cultural distance between pairs of countries, it is important to have the same number of question for each pair. Hence, if we want to cover a large number of questions, the cost is to have less countries. If we want to have a large number of countries, the cost is to have less questions. We choose to have the broadest coverage of countries, using 98 questions that were asked to all countries. This gives us 2,701 country pairs. Oberserve that this coverage of country pairs is low compared to all the country pairs for which we can compute coagglomeration patterns. Hence, we will use these Euclidian cultural distance

measures with caution and as robustness checks only.

Last, different versions of the Euclidian cultural distance can be computed by either summing across all the 98 questions—to have an overall index `cult_total`—or for each of the categories separately (`cult_total_`x, with $x = a, c, d, e, f$). We can also create create an index for binary questions only (`cult_total_binary`), and for non binary questions only (`cult_total_non_binary`).

A.2.3. Measures of genetic distance.

Genetic distances, allele-based (`gent_fst_distance_dominant` and `gent_fst_distance_weighted`). The first measure uses alleles—variants of a given gene—as genetic markers to compute genetic distances. Spolaore et Wacziarg (2016), following the landmark study by Cavalli-Sforza *et al.* (1994), provide a data set containing genetic distances computed for 42 representative populations worldwide using 120 alleles. The underlying idea is that two people are genetically related if one is the ancestor of other or they share common ancestors. This requires the people to having similar genetic markers.[32] The allele-based distance measure is based on the following formula : $F_{ST} = V_p/[\bar{p}(1-\bar{p})]$, where V_p is the variance between genes across populations and $\bar{p}$ is the average. Consider two alleles, if F_{ST} equals to 0, this means that the variance of frequency genes is null, thus the alleles are identical. If F_{ST} =1, this means that one population has only one allele and the other has only the other allele ($V_p = \bar{p}$). Thus, the higher the

32. For instance, all homo sapiens share four main blood groups, A, B, AB, and O, which are the outcomes of three different alleles, A, B, and O, of the same gene. Early studies in genetics used blood groups to look at the genetic differences between populations. Yet, the information on A, B, and O groups only is too coarse to provide measures of distance. Recent microbiology advancements in DNA sequencing and genotyping allow us to make use of new measures that provide much more precise information.

variation across the two populations, the higher the F_{ST}.

Cavalli-Sforza *et al.* (1994) provide a worldwide dataset on genetic distance at the population level. However, we require data at the country-pair level to run our regressions. Therefore, we match the genetic data to the country level using ethnic composition by country from Alesina *et al.* (2003) and the population labels from Cavalli-Sforza *et al.* (1994). For each pair, we compute the distance taking the largest population group represented in each country of the pair. The issue in doing so is that some countries contain equal-sized sub-populations. To overcome this problem, we use a second measure that weights each subgroup accordingly. Formally, suppose that two countries 1 and 2 have population subgroups $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$ respectively. The weighted formula is : $F_{ST}^{W} = \sum_{i=1}^{I} \sum_{j=1}^{J} (S_{1i} \times S_{2j} \times d_{ij})$, where S_{1i}, S_{2j} are the shares of subgroups i and j in country 1 and 2, respectively, and where d_{ij} is the genetic distance between the pairs. F_{ST}^{W} thus may be interpreted as the expected genetic distance between two randomly selected people in the two countries.

Genetic distances, microsatellite-based (`gent_new_gendist_weighted` and `gent_new_gendist_ plurality`). Our foregoing measures belong to a class of measures that uses the distribution of gene variants across populations. It thus captures the general genetic relatedness of two countries. We will also use a second class of measures based on early microsatellite-variation data by Pemberton *et al.* (2013). Microsatellites are DNA sequences that contain motifs which are repeated across thousands of locations within a genome. Their micro definition is precise and widely used for DNA profiling of some diseases, e.g., cancer diagnosis. Thus, because of their diversity and the pertinent information they carry, we use them to have another measure of genetic distance. Pemberton *et al.* (2013) cover 267—more than Cavalli-Sforza *et al.* (1994)—populations from Europe, Asia and Africa,

with 645 common microsatellite loci. As for the first class of measures, the data are at the population level and are matched to the country level using the same matching rules as before. We again compute the distance as before, using the same formulas and weighting schemes.

A.2.4. Politico-historic variables.

Colonial and politico-historical linkage variables can be used to proxy for similarities in cultural, political or legal institutions. We use three main variables in the baseline model and several variables for alternative measures as follow :

Baseline variables (poli_smctry, poli_comcol, poli_colony). Same country (poli_smctry) variable complement common colonizer (poli_comcol) variable setting to one if the pair was or is in the same state or administration entity for a long period. It covers countries that belong to the same empire, countries that have been divided (e.g., Czechoslovakia, Yugoslavia), and countries that have been belong to the same administrative colonial area. For example, Spanish colonies are distinguished following their administrative divisions on the colonial period (viceroyalities), therefore Argentina, Bolivia, Paraguay and Uruguay were a single country in the colonial period. Similarly, the Philippines were subordinated to the New Spain viceroyalty and thus same country variable equals to one with Mexico. We also provide a dummy variable of colony (poli_colony) that equals to one if one was a colony of the other at some point in time.

Alternative measures (poli_sibling, poli_heg, poli_col45, poli_comleg_pre, poli_comleg_post, poli_comleg_change, poli_nb_years_sev). As regards political alternative measures, we use sibling relationship (poli_sibling) dummy variable for origin and destination ever in sibling relationship, i.e. two colonies of

the same empire. If sibling=1, we constructed a variable (`poli_nb_years_sev`) of how many years since no longer sibling of i and j. Additionally, we make us of hegemony dummy variable (`poli_heg`) if country i (or j) is current or former hegemon of j (or i), a dummy equals to 1 for pairs in colonial relationship post 1945 (`poli_col45`). Finally, on such reasoning, we use dummy variables that equals to one if i and j share common legal system (e.g., civil law or common law) before transition (`poli_comleg_pre`), after transition, and if common legal origin changed since transition (`poli_comleg_change`) .

A.2.5. Geographic and economic controls.

Finally, we use a battery of geographic and economic variables to control for possible interactions between country pairs. The geographic controls are especially important since the linguistic, cultural, genetic, and historico-political variables are all spatially correlated. Thus, we want to see if there remains any effect on within-city location patterns once geographic proximity has been purged.

Geographic controls (`geog_contig` and `geog_continent`). To control for geographic features, we use variables from the CEPII bilateral distance database.[33] Contiguity is a dummy variable that takes value one if the pair shares of common borders. Continent is also a dummy variable that takes value one if the two countries are on the same continent.

Economic controls (`econ_flow_mean`, `econ_tour_mean`, `econ_gap_gdpcap_mean`, `econ_com_cur`, `econ_fta`, and `econ_oecd`). For trade (`econ_flow_mean`), we take the observed nominal trade flow provided by the Historical Bilateral Trade

33. See `www.cepii.fr/anglaisgraph/bdd/distances.htm`

and Gravity Data set (TradHist). The original CEPII trade data comes from different sources. It is mostly reported by the exporter and importer, but often the importer sources are more used since they have more incentive to properly assess the value of trade flows. Data concern merchandise trade and excludes services, bullion, and species. Data are at the ISO3 standard country coding and pertain to national territories, excluding colonies. For our 2016 regressions, we take the 2009–2013 average of trade. In the same manner, for our 2006 regressions, we take the 1999–2006 average of trade. We also use data on tourism flows (`econ_tour_mean`), which may be viewed as a particular type of trade in services, we obtained from the United Nations World Tourism Organization (UNWTO). It covers both origin and destination of tourists for each country of the pair, and we take the mean of influx and outflux between i and j as our measure. As for trade, we take the average, in the same manner for 2016 and 2006. In addition, we construct a GDP per capita gap variable (`econ_gap_gdpcap_mean`) between two countries i and j and take again the average across years as for trade and tourism. Finally, we also have dummy variables that equal one if a pair has a free trade agreement, as well as belongs both to the OECD or shares a common currency. With regards to dummy variables, we make them equal to 1 if at any year of the regression the dummy equals to 1 (e.g., in our 2016 regressions, we make common currency equals to 1 if it equals to 1 for any year between 2009 and 2013).

Appendix B. Mapping ethnic groups to countries

We map ethnic groups to countries using the Geo Referencing of Ethnic Groups (GREG) database (Weidmann *et al.*, 2010). This database provides a digital representation of the Soviet Atlas "Narodov Mira" (Bruk et Apenchenko, 1964). It delineates the territories of ethnic groups associated with more than 8,900 poly-

gons worldwide.[34] To understand how the procedure works, consider Figure 2.4, which depicts the border between France (in green), Spain (in pink), and Andorra (in yellow). The shaded polygons are ethnic zones from the GREG data with Basque populations (to the west) and Catalan populations (to the south-east). The grey points in Figure 2.4 depict population centroids that we use to compute population weights. We use the administrative unit center points population estimates from the Gridded Population of the World (GPW) dataset in 2016.[35] We map these population points to the ethnic polygons from the GREG database.[36] Then, we sum populations within polygons-countries where the ethnicity is present and use the resulting population totals of ethnic groups by country to compute the share of each ethnic group within each country (see Table 2.18 in the

---

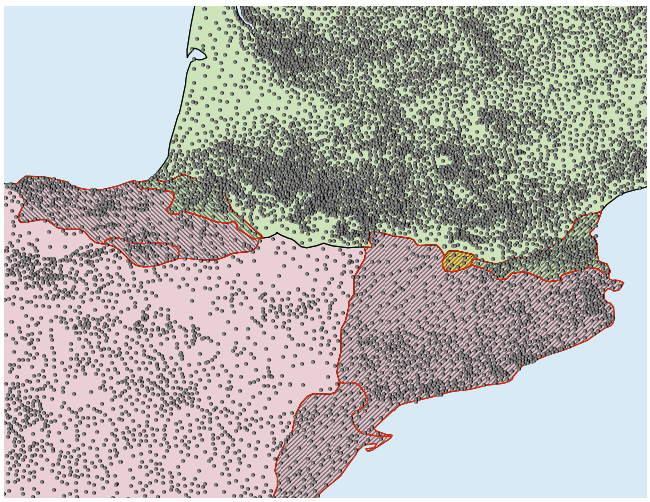
34. See Weidmann *et al.* (2010) and Bridgman (2008) for a discussion of that data and their limitations.

35. Gridded Population of the World, Version 4 (GPWv4) : Administrative Unit Center Points with Population Estimates, Revision 10. Center for International Earth Science Information Network – CIESIN – Columbia University. Palisades, NY : NASA Socioeconomic Data and Applications Center (SEDAC). URL : `https://dx.doi.org/10.7927/H4F47M2C`, accessed February 2018. Clearly, the spatial resolution of these estimates varies between countries, with some having a very high resolution, whereas others have a fairly low resolution. The advantage of this database is that it covers the world using the best available country-level data.

36. In some rare occasions, we use Wikipedia for the mapping (e.g., if the ethnic group is not reported in GREG, or if a country reported by a respondent does not exist anymore or has a different name now). Also, ethnic polygons may report up to three different ethnic groups in the same polygon (e.g., Catalans and Spaniards). Since we have no information on how to split between these different groups, we count each person once for each of the ethnic groups when computing the shares. We could also use equal splits (e.g., 1/3, 1/3 and 1/3, but this changes little and is as arbitrary). Finally, there are cases where one or more ethnic groups are present in a single country only (e.g., Bretons in Brittany, which lies in France). In that case the mapping is straightforward.

online appendix for a detailed breakdown of the mapping from ethnic groups to countries).

FIGURE 2.4: Mapping ethnic groups—for example, Basques and Catalans—to countries.



Formally, let θ_i^c denote the share of ethnic group i in country c, with $\sum_c \theta_i^c = 1$. We use these shares to split out ethnic groups in the different DAs among countries. For example, a dissemination area in city c that reports 100 residents of Flemish ethnicity will be split into $100 \times \theta_{Flemish}^{BEL} = 96$ people from Belgium and $100 \times \theta_{Flemish}^{FRA} = 4$ people from France, using the shares summarized in Table 2.18.[37] Observe that by splitting the Flemish into French and Belgian, we 'artificially' create a set short bilateral distances within the couple France-Belgium. However, how this affects our measures of colocation is unclear since in doing so we also create a new set of long bilateral distances between the other French and Belgian populations. In any case, our results are robust to excluding all groups that we

37. We round fractional splits to the closest integers since our weights in the K-density computations need to be integers. We do not think that this makes a substantial difference since, as explained before, the census numbers are already randomly rounded up or down to the nearest multiple of five.

'split'.

Appendix C : Additional results

This appendix contains additional tables and results.

2006 Census. Tables 2.14 and 2.15 show the same results as Tables 2.3 and 2.4 but for the 2006 Census. As can be seen, our results are very robust and change little compared to the 2016 Census. The only exceptions are for bilateral trade and tourism flows, and for common official language, which tend to become insignificant using the 2006 Census data. Actually, all coefficients (including those on geographic proximity) become smaller and are less precisely estimated. As explained in Appendix A.1, the 2006 Census features less disaggregated data of ethnic groups, which explains why we have smaller sample sizes and why the results are generally less precise.

Robustness to large enough ethnic groups. Table 2.16 presents results using our alternative measures of similarity and the K-densities estimated for sufficiently large ethnic groups only. The results are qualitatively similar to those in Table 2.5. The only difference is that some language variables become insignificant, and that some of the historico-political variables are affected. But globally, the results are very similar to those in our baseline regressions.

Results for the rich and owners. Table 2.17 depicts our results where we estimate the K-densities for the rich DAs and the 'owner' DAs as defined in the main text.

TABLE 2.14: Univariate baseline results, 2006 Census.

Dependent variable : $\widehat{K}_c^{ij}(500m)$	Coeff.		R^2	N
Contiguity	0.04^a	(0.00)	0.76	56,160
Same continent	0.06^a	(0.00)	0.76	56,160
Common currency	0.04^a	(0.01)	0.76	56,160
Free trade aggrement	0.07^a	(0.01)	0.76	56,160
Both OECD	0.07^a	(0.00)	0.76	56,160
Bilateral trade flows	0.02^a	(0.01)	0.76	54,470
Bilateral tourist flows	0.02^a	(0.01)	0.76	54,816
GDP per capita gap	-0.08^a	(0.00)	0.76	54,553
Were same country	0.04^a	(0.01)	0.76	56,160
Common colonizer	0.04^a	(0.00)	0.76	56,160
Colonial relationship	0.00^c	(0.00)	0.76	56,160
Common official language	0.03^a	(0.00)	0.76	56,160
Common religion	0.04^a	(0.00)	0.76	56,160
Genetic distance (alelle, plurality groups)	-0.05^a	(0.00)	0.76	56,160

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. ^{a}p$<$0.01, ^{b}p$<$0.05, ^{c}p$<$0.1.

TABLE 2.15: Multivariate baseline results, 2006 Census.

Dependent variable : $\widehat{K}_c^{ij}(500m)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Contiguity	0.02^a	0.01^a	0.01	0.01	0.01	0.00^c	0.01^a
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Same continent	0.05^a	0.03^a	0.03^a	0.03^a	0.02^a	0.02^a	0.03^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common currency		0.02^a	0.02^c	0.02^c	0.02^c	-0.00^a	-0.00
		(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Free trade agreement		0.03^a	0.03^a	0.02^a	0.02^a	-0.00	0.00^b
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Both OECD		0.02^a	0.02^a	0.02^a	0.02^a	0.03^a	0.02^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Bilateral trade flows		0.00	0.00	0.00	0.00	0.00^b	0.00^c
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Bilateral tourism flows		-0.01	-0.00	-0.00	-0.00	-0.00^b	-0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GDP per capita gap		-0.03^a	-0.03^a	-0.03^a	-0.03^a	-0.03^a	-0.03^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Were same country			0.02^c	0.02^c	0.02^c	0.02^b	0.01^b
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Common colonizer			0.02^a	0.02^a	0.02^a	0.05^a	0.04^a
			(0.00)	(0.00)	(0.00)	(0.01)	(0.01)
Colonial relationship			0.00^a	0.00^c	0.00^c	0.00^b	0.00^a
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common official language				0.00	0.00	0.01^b	0.00
				(0.00)	(0.00)	(0.00)	(0.00)
Common religion				0.01^a	0.01^a	0.01^a	0.01^a
				(0.00)	(0.00)	(0.00)	(0.00)
Genetic Distance (alelle, plurality groups)					-0.02^a	-0.04^a	-0.03^a
					(0.00)	(0.00)	(0.00)
Weighted	no	no	no	no	no	yes[1]	yes[2]
Fixed effects	*ic* and *jc* (country-city) fixed effects						
Country pairs	All pairs included						
Sample size	56,160	51,820	51,820	51,820	51,820	51,820	51,820
R^2	0.76	0.78	0.78	0.78	0.78	0.86	0.85

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. ^{a}p<0.01, ^{b}p<0.05, ^{c}p<0.1, [1]population weights , [2]geographic weights.

TABLE 2.16: Robustness of alternative measures of linguistic and genetic proximity, 'high quality' K-densities only, 2016 Census

Description	Stata variable name	Coeff.		Sample size	R^2
Common spoken language	lang_csl	0.031[a]	(0.003)	35,883	0.829
Common native language	lang_cnl	0.015[a]	(0.002)	35,883	0.829
Linguistic proximity (Tree, unadjusted)	lang_prox1	0.005	(0.003)	35,883	0.828
Linguistic proximity (Tree, adjusted)	lang_lp1	0.004	(0.003)	34,244	0.834
Linguistic proximity (ASJP, unadjusted)	lang_prox2	0.004	(0.003)	35,883	0.828
Linguistic proximity (ASJP, adjusted)	lang_lp2	0.004	(0.003)	34,244	0.834
Common Language Index (log specification)	lang_cl	0.027[a]	(0.003)	34,244	0.835
Common Language Index (level specification)	lang_cle	0.025[a]	(0.003)	35,883	0.829
Common official or primary language	lang_comlang_off	0.023[a]	(0.003)	35,883	0.829
Language is spoken by at least 9 % of the population	lang_comlang_ethno	0.015[a]	(0.004)	35,883	0.829
Linguistic distance (words, plurality languages)	lang_cognate_dominant	-0.021[a]	(0.005)	9,623	0.855
Linguistic distance (words, weighted)	lang_cognate_weighted	-0.037[a]	(0.005)	5,149	0.902
Linguistic distance (trees, plurality languages)	lang_lingdist_dom_formula	-0.007[a]	(0.003)	31,619	0.824
Linguistic distance (trees, weighted)	lang_lingdist_weighted_formula	-0.007[a]	(0.003)	31,619	0.824
Genetic distance (microsatellite variation, weighted)	gent_new_gendist_weighted	-0.044[a]	(0.005)	33,776	0.828
Genetic distance (microsatellite variation, plurality groups)	gent_new_gendist_plurality	-0.043[a]	(0.006)	33,776	0.828
Genetic distance (alelle, weighted)	gent_fst_distance_weighted	-0.027[a]	(0.004)	34,380	0.827
Euclidian cultural distance, all categories	cult_total	-0.029[a]	(0.006)	11,354	0.908
Euclidian cultural distance, category A only	cult_total_a	-0.017[a]	(0.005)	11,354	0.908
Euclidian cultural distance, category C only	cult_total_c	-0.008[c]	(0.005)	11,354	0.907
Euclidian cultural distance, category D only	cult_total_d	-0.014[a]	(0.004)	11,354	0.908
Euclidian cultural distance, category E only	cult_total_e	-0.022[a]	(0.006)	11,354	0.908
Euclidian cultural distance, category F only	cult_total_f	-0.008[b]	(0.004)	11,354	0.907
Euclidian cultural distance, binary choice questions only	cult_total_binary	-0.015[a]	(0.005)	11,354	0.908
Euclidian cultural distance, non-binary choice questions only	cult_total_non_binary	-0.027[a]	(0.006)	11,354	0.908
Country was post-45 colonizer of the other	poli_col45	0.001	(0.002)	35,883	0.827
Countries in the same 'empire' or had common colonizer	poli_sibling	0.018[a]	(0.003)	35,883	0.828
Hegemony relationship	poli_heg	0.003	(0.002)	35,883	0.827
Number of years since no longer siblings (cond. on sibling $=1$)	poli_nb_years_sev	-0.011	(0.012)	5,572	0.870
Common legal origins pre-independence	poli_comleg_pre	0.021[a]	(0.003)	35,883	0.828
Common legal origins post-independence	poli_comleg_post	0.014[a]	(0.003)	35,883	0.828
Common legal origins across countries changed	poli_comleg_change	0.001	(0.003)	35,883	0.827
Religious distance (plurality Fearon et al.)	cult_reldist_dominant_formula	-0.009[a]	(0.003)	31,247	0.825
Religious distance (weighted, Fearon et al.)	cult_reldist_weighted_formula	-0.015[a]	(0.004)	31,247	0.825
Religious distance (plurality, WCD)	cult_reldist_dominant_WCD_form	-0.015[a]	(0.003)	34,032	0.830
Religious distance (weighted, WCD)	cult_reldist_weighted_WCD_form	-0.022[a]	(0.004)	34,032	0.830

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for country pairs with large size only (HQ). The specification that we use is (6) in all regressions, with only the language, religion, culture, politics or genetics variable changed. We replace variables as follows in the different regressions : (i) Language : We drop 'common official language' and we replace with the new language variable ; (ii) Genetics : We replace the genetics variable with the new genetics variable ; (iii) Culture : We replace both language and religion with the cultural variables ; (iv) Historico-political : We replace 'common colonizer' and 'colonial relationship' with the new variables ; and (v) Religion : We replace 'common religion' with the new religion variable. [a]p<0.01, [b]p<0.05, [c]p<0.1.

TABLE 2.17: Results for rich and owner DAs, 2016 Census.

	(1)	(2)	(3)	(4)	(5)
	All	Rich DAs	Restricted	Owner DAs	Restricted to
Dependent variable : $\widehat{K}_c^{ij}(500m)$		only	to rich	only	to owners
Contiguity	0.01^a	0.01^a	0.01^a	0.04^a	0.01^a
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Continent	0.03^a	0.02^a	0.02^a	0.02^a	0.02^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common currency	0.01^a	0.00	0.01^a	0.01	0.00
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Free trade aggrement	0.02^a	0.01^a	0.02^a	0.02^a	0.02^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
OECD	0.03^a	0.02^a	0.03^a	0.02^a	0.03^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Trade	0.01^a	0.00	0.01^a	0.01^b	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Tourism	-0.01^a	-0.00	-0.01^a	-0.01^b	-0.01^a
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
GDP per capita gap	-0.07^a	-0.03^a	-0.07^a	-0.05^a	-0.06^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
same country	0.01^a	0.00	0.01^b	-0.01	0.01^b
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Common colonizer	0.03^a	0.02^a	0.03^a	0.03^a	0.03^a
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Colonial relationship	0.00^a	0.00^a	0.00^b	0.00	0.00^c
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
COL	0.01^a	0.00^c	0.02^a	0.02^a	0.02^a
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Common religion	0.01^a	0.02^a	0.02^a	0.00	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Genetic Distance (allele, plurality groups)	-0.04^a	0.00	-0.03^a	-0.04^a	-0.03^a
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Fixed effect	ic and jc (country-city) fixed effects.				
Country pairs	All, computed on rich or owner DAs only.				
Sample size	62,145	51,461	51,461	49,373	49,373
R^2	0.87	0.83	0.87	0.61	0.85

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. ^{a}p<0.01, ^{b}p<0.05, ^{c}p<0.1.

Supplemental online material

This set of supplemental online appendices is structured as follows. In Appendix **S.1**, we briefly discuss why self-selection is not an issue for our analysis. Appendix **S.2** contains additional figures and tables that summarize results concerning the mapping of ethnic groups to countries and the distribution of ethnic groups across the dissemination areas of the cities.

Appendix S.1. Self-selection into migration and across cities

Another possible identification concern in our analysis is that there is likely to be self-selection of ethnic groups *into* migration and *across* cities. Migration is a multi-stage problem. First, people decide on whether or not to migrate ; second, conditional on coming to Canada, they pick provinces and cities ; and third, conditional on picking cities, they choose neighborhoods within cities. Some ethnic groups may have stronger incentives to migrate—because of international wars, internal conflicts, or adverse climatic or economic conditions—and within those groups migrants are unlikely to be a random sample (see, e.g., Borjas 1987). While this is well understood, there is little we can do about it in our study. If immigrants are, e.g., more educated and open-minded than people who do not migrate, we may see that there is more mixing in Canadian cities between ethnic groups than would prevail if immigrants were randomly drawn from their respective populations. Turning to location choices across cities, it is well understood that some groups historically immigrate more to some provinces and cities in Canada (e.g., North Africans and people from Black Africa to Montréal ; Indians and Pakistani to Toronto ; and Asians to Vancouver).[38] Thus, the observed split of groups across

38. This is further complicated by the fact that part of the immigration process takes place at the federal level, but that the provinces have special competences to modulate part of that

cities reflects the between-city location problem, which could—at least partly—depend on the same X^{ij} that we are interested in. The city-specific K-densities may thus encapsulate this upper-tier location problem, i.e., there is a selection problem.

We cannot really address this problem in a satisfying way since we cannot control for the first-stage location choices. Yet, our country-city fixed effects will soak up any variation linked to country-city pairs, which is likely to subsume most effects linked to regional variation in historical immigration patterns and immigration requirements and policy. What we cannot control is that spatial sorting may be across cities and not within cities. Assume, e.g., that ethnic groups i and j dislike each other strongly and hence pick different cities altogether. In that case they will not show up in our data—recall that we compute colocation measures only for pairs within cities—and our coefficients would be biased. A similar problem arises if the ethnic groups tend to predominantly pick different cities so that the joint distribution of the two groups in the same city always has one group of small size. In that case, if we drop these observations because the small size makes the K-density estimations more noisy, we would also introduce a bias into our analysis. Hence, we present estimation results where all pairs ij are kept in the analysis because this is likely to alleviate this type of selection bias. [39]

To summarize, there are two types of potential section biases : into migration, and

---

process (e.g., selection is based on different quantitative criteria in Québec, and Ontario has leeway for pushing specific groups in terms of skills or education.

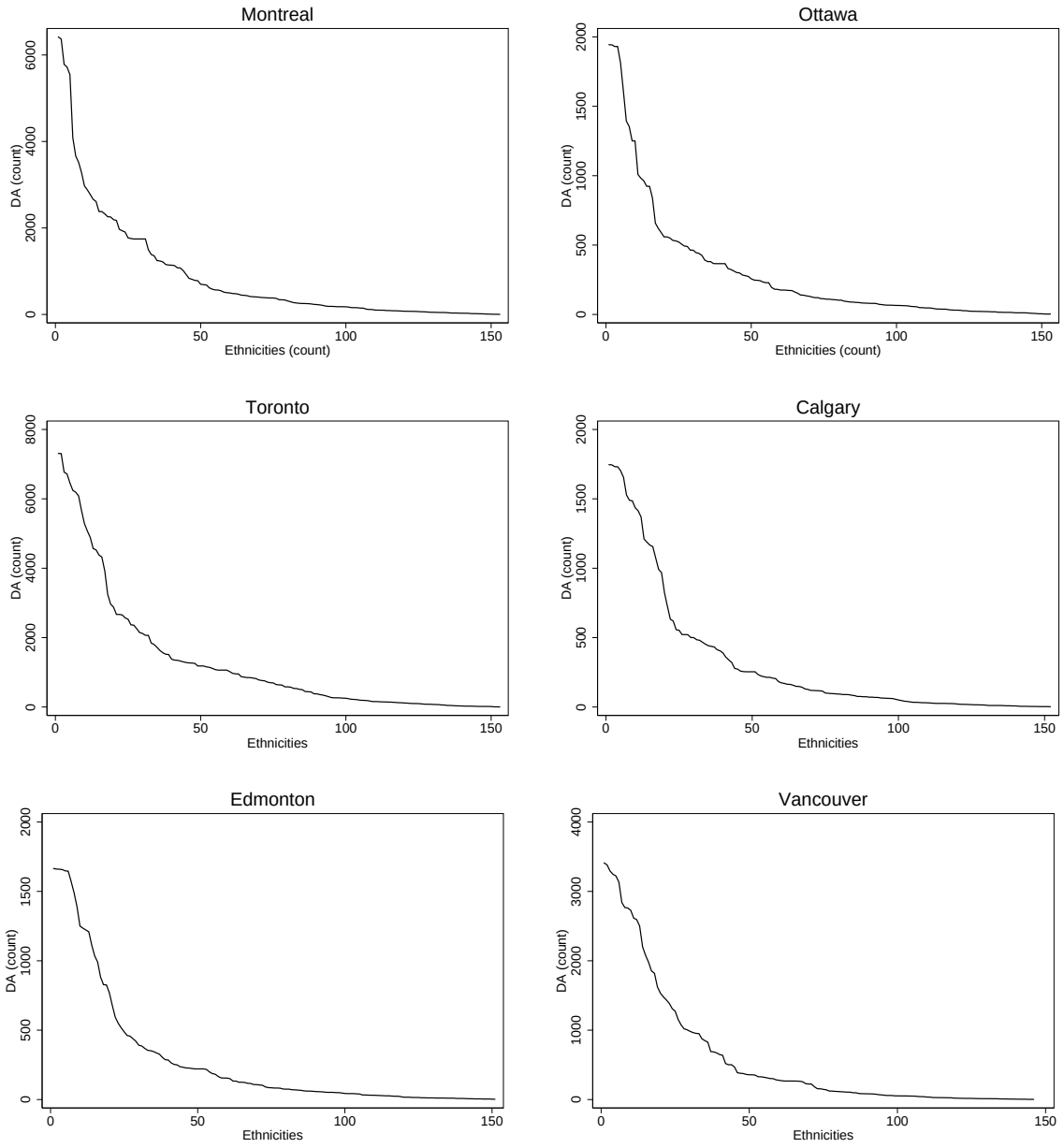
39. Extreme sorting into disjoint cities is not present in our data. For example, we have 169 different countries in our dataset in 2016, which allows potentially for 85,176 pairs (= $(169 \times 168)/2$ unsorted pairs for each of the 6 cities). We have K-densities for 83,365 pairs, implying that we only loose 2.23% of the pairs (which are pairs that are always completely disjoint between cities). These are few pairs and correspond to quite small ethnic groups.

across cities. While we cannot control the former, we think that presenting results that include all pairs of ethnic groups into the analysis will help to mitigate the latter.

Appendix S.2. Additional tables and results

— Table 2.18 shows the mapping of ethnic groups to countries, including the different population shares.
— Figure 2.5 shows that there are many relatively small ethnic groups in the cities, and that the distribution of groups across DAs is skewed : there are many groups that are small in the sense that they are only present in a small number of DAs in each city.
— Table 2.19 summarizes results for the 2016 and 2006 censuses for coagglomeration patterns measured at 100m and 1km distance thresholds, respectively.
— Table 2.20 shows the correlations between the different measures of linguistic distance that we use. While some of these correlations are large, they are not too large on average, meaning that our explanatory variables related to language capture different aspects.
— Finally, Table 2.21, provides a detailed breakdown of the largest and smallest ethnic groups by cities.

FIGURE 2.5: Distribution of ethnic groups across dissemination areas (2016).



Notes : Distribution of number of dissemination areas with non-zero population for ethnic groups across the six metropolitan areas. The long right tails in the figure show that many ethnic groups are represented in a small number of DAs only.

TABLE 2.18: Mapping from ethnic groups to countries.

Ethnicity	Country	Share	Ethnicity	Country	Share	Ethnicity	Country	Share
Afrikaner	South Africa	96%	Corean	North Korea	32%	Peulh	Mali	10%
Afrikaner	Namibia	4%	Corean	South Korea	62%	Peulh	Senegal	18%
Arab	Saudi Arabia	18%	Corean	China	5%	Peulh	Cameron	12%
Arab	Turkey	2%	Corean	Russia	1%	Peulh	Nigeria	25%
Arab	Egypt	52%	Flemish	France	4%	Peulh	Burkina Faso	6%
Arab	Kuwait	2%	Flemish	Belgium	96%	Peulh	Niger	6%
Arab	Oman	3%	Karen	Thailand	38%	Tadjik	Afghanistan	97%
Arab	Bahrain	1%	Karen	Myanmar	62%	Tadjik	Iran	3%
Arab	Qatar	3%	Kurde	Syria	7%	Tamoul	India	88%
Arab	Yemen	14%	Kurde	Iraq	36%	Tamoul	Sri Lanka	8%
Arab	U. A. Emirates	5%	Kurde	Iran	23%	Tamoul	Malaysia	4%
Akan	Togo	1%	Kurde	Turkey	32%	Tatar	Romania	0.7%
Akan	Ghana	70%	Kurde	Azerbaijan	1%	Tatar	Russia	99%
Akan	Cote d'Ivoire	29%	Kurde	Armenia	1%	Tatar	China	0.3%
Bantou	Central African Republic	2%	Malinke	Guinea-Bissau	2%	Tzigane	Hungary	0.1%
Bantou	Congo Democratic	27%	Malinke	Senegal	10%	Tzigane	Romania	0.6%
Bantou	Rwanda	13%	Malinke	Cote d'Ivoire	7%	Tzigane	Serbia	0.3%
Bantou	Congo	2%	Malinke	Gambia	8%	Tzigane	Ukraine	99%
Bantou	Cote d'Ivoire	19%	Malinke	Guinea	18%	Wolof	Gambia	1%
Bantou	Liberia	37%	Malinke	Mali	49%	Wolof	Senegal	99%
Basque	Spain	95%	Malinke	Sierra Leone	1%	Yoruba	Togo	1%
Basque	France	5%	Malinke	Burkina Faso	5%	Yoruba	Nigeria	96%
Bengali	Nepal	0.2%	Maya	Belize	5%	Yoruba	Benin	3%
Bengali	Bhutan	0.1%	Maya	Mexico	95%	Zulu	Mozambique	1%
Bengali	Bangladesh	56.3%	Pendjabi	India	37%	Zulu	South Africa	99%
Bengali	India	43%	Pendjabi	Pakistan	63%			
Bengali	Myanmar	0.4%	Peulh	Guinea-Bissau	0.1%			
Catalan	Spain	95%	Peulh	Guinea	18%			
Catalan	Italy	0.1%	Peulh	Mauritania	4%			
Catalan	France	4%	Peulh	Chad	2%			
Catalan	Andorra	0.9%	Peulh	Togo	0.9%			

Notes : Our computations, based on GREG and GPW data.

TABLE 2.19: Robustness to spatial scale, 2016 Census.

	Dependent var. : $\widehat{K}^c_{ij}(100m)$							Dependent var. : $\widehat{K}^c_{ij}(1km)$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Contiguity	0.03^a	0.02^a	0.01^a	0.01^a	0.01^a	0.01^b	0.01^a	0.03^a	0.02^a	0.01^a	0.01^a	0.01^a	0.01^b	0.01^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Same continent	0.07^a	0.04^a	0.04^a	0.04^a	0.03^a	0.03^a	0.03^a	0.07^a	0.04^a	0.04^a	0.04^a	0.03^a	0.03^a	0.04^a
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common currency		0.02^a	0.01^a	0.01^a	0.01^a	-0.01^b	-0.01^a		0.02^a	0.01^a	0.01^a	0.01^a	-0.01^b	-0.01^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Free trade aggrement		0.03^a	0.03^a	0.03^a	0.02^a	0.00	0.01^a		0.03^a	0.03^a	0.03^a	0.02^a	0.00	0.01^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Both OECD		0.03^a	0.03^a	0.03^a	0.03^a	0.03^a	0.03^a		0.03^a	0.03^a	0.03^a	0.03^a	0.03^a	0.03^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Trade flows		0.01^a	0.01^a	0.01^a	0.01^a	0.01^a	0.01^a		0.01^a	0.01^a	0.01^a	0.01^a	0.01^a	0.01^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Tourism flows		-0.01^a	-0.01^a	-0.01^a	-0.01^a	-0.01^a	-0.01^a		-0.01^a	-0.01^a	-0.01^a	-0.01^a	-0.01^a	-0.01^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
GDP per capita gap		-0.07^a	-0.07^a	-0.07^a	-0.07^a	-0.05^a	-0.05^a		-0.07^a	-0.07^a	-0.07^a	-0.07^a	-0.05^a	-0.05^a
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Were same country			0.01^a	0.01^a	0.01^a	0.03^b	0.01^b			0.01^a	0.01^a	0.01^a	0.03^b	0.01^b
			(0.00)	(0.00)	(0.00)	(0.01)	(0.01)			(0.00)	(0.00)	(0.00)	(0.01)	(0.01)
Common colonizer			0.04^a	0.03^a	0.03^a	0.04^a	0.03^a			0.04^a	0.03^a	0.03^a	0.04^a	0.03^a
			(0.00)	(0.00)	(0.00)	(0.01)	(0.00)			(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Colonial relationship			0.01^a	0.00^a	0.00^a	0.00^c	0.00^b			0.01^a	0.00^a	0.00^a	0.00^c	0.00^b
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Common official language				0.01^a	0.01^a	0.01^a	0.02^a				0.02^a	0.01^a	0.01^a	0.02^a
				(0.00)	(0.00)	(0.00)	(0.00)				(0.00)	(0.00)	(0.00)	(0.00)
Common religion				0.01^a	0.01^a	0.01^a	0.02^a				0.01^a	0.01^a	0.01^a	0.02^a
				(0.00)	(0.00)	(0.00)	(0.00)				(0.00)	(0.00)	(0.00)	(0.00)
Genetic Distance (allele, plurality groups)					-0.03^a	-0.04^a	-0.04^a					-0.04^a	-0.04^a	-0.04^a
					(0.00)	(0.00)	(0.00)					(0.00)	(0.00)	(0.00)
Fixed effect	ic,jc													
Country pairs	All country pairs													
Sample size	68,055	62,145	62,145	62,145	62,145	62,145	62,145	68,055	62,145	62,145	62,145	62,145	62,145	62,145
R^2	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.86	0.87	0.87	0.87	0.87	0.87

Notes : Standardized OLS regression coefficients. All standard errors provided in parentheses are clustered by country pairs ij. All regressions include ic and jc (country-city) fixed effects and are run using the K-densities for all country pairs. ^{a}p<0.01, ^{b}p<0.05, ^{c}p<0.1.

TABLE 2.20: Correlation matrix, measures of linguistic distance.

		Obs.	Mean	SD	1.	2	3	4	5	6	7	8	9	10	11	12	13	14
1.	Common official language	68,055	0.13	0.33														
2.	Common spoken language	68,055	0.11	0.22	0.81													
3.	Common native language	68,055	0.03	0.15	0.94	0.81												
4.	Linguistic proximity (Tree, unadjusted)	68,055	0.07	0.14	-0.60	-0.44	-0.65											
5.	Linguistic proximity (Tree, adjusted)	61,212	0.61	1.13	-0.60	-0.44	-0.65	1.00										
6.	Linguistic proximity (ASJP, unadjusted)	68,055	0.07	0.08	-0.45	-0.27	-0.49	0.91	0.91									
7.	Linguistic proximity (ASJP, adjusted)	61,212	0.64	0.73	-0.45	-0.27	-0.49	0.91	0.91	1.00								
8.	Common Language Index (log specification)	61,212	0.13	0.17	0.87	0.80	0.90	-0.28	-0.28	-0.06	-0.06							
9.	Common Language Index (level specification)	68,055	0.13	0.16	0.86	0.80	0.89	-0.28	-0.28	-0.06	-0.06	1.00						
10.	Common official or primary language	68,055	0.14	0.35	1.00	0.81	0.93	-0.59	-0.59	-0.45	-0.45	0.86	0.86					
11.	Language spoken by at least 9% of the population	68,055	0.15	0.36	0.91	0.77	0.91	-0.61	-0.61	-0.47	-0.47	0.82	0.81	0.92				
12.	Linguistic distance (words, plurality languages)	15,046	0.63	0.29	-0.74	-0.72	-0.75	0.04	0.04	-0.11	-0.11	-0.93	-0.93	-0.74	-0.69			
13.	Linguistic distance (words, weighted)	7,902	0.60	0.28	-0.74	-0.71	-0.76	0.04	0.04	-0.10	-0.10	-0.93	-0.94	-0.74	-0.69	0.98		
14.	Linguistic distance (trees, plurality languages)	56,716	0.96	0.15	-0.86	-0.74	-0.90	0.45	0.45	0.33	0.33	-0.88	-0.87	-0.86	-0.81	0.80	0.80	
15.	Linguistic distance (trees, weighted)	56,716	0.97	0.11	-0.84	-0.72	-0.90	0.41	0.41	0.30	0.30	-0.88	-0.88	-0.83	-0.79	0.79	0.82	0.95

TABLE 2.21: Top- and bottom-20 ethnic groups in each city (2016).

Montréal						Ottawa						Toronto					
All		**Rich**		**Poor**		**All**		**Rich**		**Poor**		**All**		**Rich**		**Poor**	
Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA
Canada	6418	Canada	1587	Canada	1581	France	1944	Canada	476	Canada	476	U.K	7308	Canada	1796	U.K	1795
France	6369	France	1577	France	1560	Canada	1943	France	476	France	476	Canada	7304	U.K	1794	Canada	1786
Ireland	5784	Ireland	1472	Ireland	1380	U.K	1931	U.K	476	U.K	471	Ireland	6770	Ireland	1724	Ireland	1654
Italy	5722	Italy	1458	Italy	1363	Ireland	1931	Ireland	475	Ireland	468	Italy	6719	Italy	1724	Italy	1612
U.K	5546	U.K	1436	U.K	1330	Germany	1813	Germany	469	Germany	413	China	6459	Germany	1660	India	1608
Germany	4087	Germany	1093	Haiti	1045	Italy	1608	Italy	433	Italy	357	Germany	6245	Poland	1612	China	1606
Spain	3663	Spain	871	Spain	979	Poland	1394	Poland	394	Poland	280	India	6191	China	1606	France	1516
Haiti	3517	China	870	Germany	948	Netherlands	1354	Netherlands	379	Netherlands	273	France	6082	France	1572	Germany	1503
China	3280	Poland	858	Morocco	900	Ukraine	1251	Ukraine	370	China	272	Poland	5660	India	1445	Philippines	1486
Portugal	2974	Lebanon	741	China	887	China	1250	China	367	Lebanon	247	Portugal	5285	Ukraine	1400	Jamaica	1457
Poland	2881	Greece	737	Algeria	880	Lebanon	1009	India	298	Ukraine	247	Philippines	5075	Russia	1368	Portugal	1326
Morocco	2774	Portugal	730	Turkey	761	Russia	982	Russia	297	Spain	236	Ukraine	4888	Netherlands	1244	Spain	1318
Algeria	2663	Russia	697	Egypt	757	India	963	U. S. A.	274	Haiti	209	Russia	4567	Portugal	1194	Poland	1300
Egypt	2612	Haiti	681	Portugal	706	U. S. A.	924	Lebanon	264	India	207	Jamaica	4530	Greece	1075	Ukraine	1123
Greece	2381	Romania	667	Poland	655	Spain	923	Spain	229	Portugal	204	Spain	4383	Philippines	959	Russia	1100
Lebanon	2378	Egypt	661	Yemen	643	Portugal	839	Portugal	207	Russia	194	Netherlands	4320	Hungary	943	Netherlands	1025
Russia	2325	Belgium	641	S.A	635	Hungary	658	Hungary	200	U. S. A.	188	Greece	3921	Spain	926	Greece	901
Belgium	2260	Ukraine	624	U.A.E	633	Philippines	620	Sweden	183	Egypt	181	Hungary	3251	U. S. A.	916	Guyana	871
Turkey	2255	Morocco	579	Bahrain	633	Egypt	590	Austria	180	Turkey	178	U. S. A.	2969	Romania	770	Pakistan	842
Romania	2193	U. S. A.	552	Kuwait	633	Haiti	559	Romania	180	Philippines	165	Pakistan	2880	Austria	749	Sri Lanka	837
-		-		-		-		-		-		-		-		-	
Macedonia	45	Angola	9	Iceland	15	Gambia	19	Georgia	5	Bolivia	8	Guinea	56	Fiji	11	Bermuda	30
New Zealand	41	C. A. R.	9	Kenya	14	Bahamas	16	Guinea	5	Grenada	8	Burundi	45	Cameroon	10	Burundi	27
S. K. N.	34	Congo	9	Uzbekistan	14	Panama	16	Sierra Leone	5	A. B.	6	Liberia	45	Honduras	10	Liberia	26
Bahamas	33	Georgia	9	Eritrea	13	Costa Rica	15	Bahamas	4	Bahamas	6	Gambia	39	Angola	6	Gambia	24
Uzbekistan	33	Uganda	9	Estonia	13	Georgia	15	Bolivia	4	Ecuador	6	Turkmenistan	34	Côte d'Ivoire	6	Mali	24
Eritrea	30	Uzbekistan	8	Cyprus	11	Uzbekistan	15	Gambia	4	Gambia	6	Mali	33	Paraguay	5	Tunisia	23
Kenya	29	Bahamas	7	S. K. N.	11	Zambia	14	Grenada	4	Mauritania	6	Zambia	29	Rwanda	5	Singapore	20
A. B.	28	Kenya	7	Bahamas	10	Cyprus	12	Madagascar	4	Niger	6	Paraguay	26	Mozambique	4	C. A. R.	15
Paraguay	28	Gambia	6	Malta	10	Kazakhstan	12	Uganda	4	Uzbekistan	6	C. A. R.	25	Seychelles	4	Congo	15
Uganda	21	S. K. N.	6	New Zealand	10	Mauritania	12	Angola	3	Zambia	6	Congo	25	Burundi	3	Chad	15
Sudan	19	A. B.	5	Paraguay	10	Niger	12	Costa Rica	3	Costa Rica	5	Burkina Faso	22	Madagascar	3	Turkmenistan	15
Zimbabwe	19	Sudan	5	Djibouti	9	Uruguay	11	Djibouti	3	Georgia	5	Chad	19	Chad	3	Djibouti	13
Djibouti	16	Eritrea	4	Zimbabwe	9	Bermuda	9	Honduras	3	New Zealand	5	Madagascar	17	Turkmenistan	3	Burkina Faso	12
Tanzania	15	Sierra Leone	4	A. B.	8	S. K. N.	8	Zambia	3	Mauritius	4	Mozambique	17	Zambia	3	Guinea-Bissau	10
Bermuda	10	Bermuda	3	Macedonia	7	Turkmenistan	7	Fiji	2	Bermuda	3	Djibouti	16	Burkina Faso	2	Paraguay	10
Singapore	9	Singapore	3	Sudan	6	Fiji	5	Gabon	2	Kazakhstan	3	Seychelles	15	C. A. R.	2	Seychelles	8
Mozambique	6	Seychelles	3	Tanzania	6	Paraguay	5	Guinea-Bissau	2	S. K. N	3	Guinea-Bissau	14	Congo	2	Madagascar	7
Fiji	5	Djibouti	2	Uganda	6	Mozambique	3	Kazakhstan	2	Panama	3	Mauritania	5	Guinea	2	Mozambique	6
Turkmenistan	4	Turkmenistan	2	Bermuda	2	Singapore	3	Paraguay	2	Turkmenistan	3	Niger	5	Liberia	2	Mauritania	4
Zambia	2	Zimbabwe	2	Singapore	2	Seychelles	3	Chad	2	Uruguay	2	Gabon	2	Sierra Leone	2	Niger	4

Table 15 (continued).

Calgary						Edmonton						Vancouver					
All		Rich		Poor		All		Rich		Poor		All		Rich		Poor	
Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA	Ethnicity	# DA
Canada	1746	U. K.	426	Canada	424	U. K.	1665	Canada	405	U. K.	404	U. K.	3411	U. K.	843	U. K.	837
U. K.	1745	Ireland	426	U. K.	424	Canada	1661	Germany	405	Germany	402	Canada	3383	Canada	836	Canada	835
Ireland	1732	Canada	425	Germany	419	Germany	1660	U. K.	405	France	402	Ireland	3295	Ireland	834	Ireland	808
Germany	1731	Germany	425	Ireland	419	Ireland	1657	Ireland	405	Canada	401	Germany	3247	Germany	825	Germany	801
France	1703	Ukraine	420	France	413	Ukraine	1648	Ukraine	405	Ireland	401	China	3223	China	810	China	798
Ukraine	1655	France	418	Ukraine	386	France	1646	France	397	Ukraine	399	France	3134	France	797	France	789
Poland	1529	Poland	399	China	362	Poland	1572	Poland	392	Poland	371	Ukraine	2840	Ukraine	726	Russia	723
Netherlands	1490	Norway	384	Philippines	357	Netherlands	1490	Netherlands	368	Netherlands	356	Russia	2768	Russia	721	Ukraine	722
China	1486	Netherlands	382	Poland	344	Norway	1389	Norway	361	China	322	India	2761	Italy	704	Philippines	720
Norway	1436	China	378	Netherlands	335	Russia	1250	Sweden	331	Philippines	321	Italy	2727	Poland	697	India	693
Russia	1414	Russia	366	Russia	315	Italy	1235	Russia	327	Norway	314	Poland	2615	Netherlands	678	Italy	691
Italy	1369	Italy	363	Norway	309	Sweden	1222	Italy	316	Russia	290	Netherlands	2593	India	622	Poland	663
Philippines	1211	U. S. A.	335	Italy	298	China	1209	U. S. A.	302	Italy	284	Philippines	2501	Norway	599	Netherlands	647
India	1188	Sweden	325	India	290	Philippines	1113	China	284	India	275	Norway	2200	Sweden	567	Spain	604
Sweden	1167	India	301	Spain	243	U. S. A.	1035	Denmark	248	Sweden	256	Sweden	2078	U. S. A.	519	Norway	543
U. S. A.	1156	Hungary	295	Sweden	242	India	991	India	243	U. S. A.	213	Spain	1980	Philippines	469	Sweden	541
Hungary	1074	Denmark	283	U. S. A.	239	Denmark	883	Austria	241	Spain	211	U. S. A.	1854	Japan	466	Japan	492
Denmark	992	Austria	245	Hungary	234	Austria	828	Hungary	228	Denmark	185	Japan	1820	Austria	416	Korea	484
Spain	968	Spain	226	Denmark	195	Hungary	826	Philippines	215	Hungary	174	Hungary	1624	Hungary	405	U. S. A.	466
Austria	824	Philippines	225	Viet Nam	191	Spain	768	Romania	190	Austria	166	Korea	1534	Spain	384	Hungary	438
			-						-						-		
C. F. A.	11	Saint Lucia	4	Costa Rica	6	Guinea-Bissau	11	Liberia	4	Uzbekistan	6	A. B.	14	Jordan	6	Congo	7
Congo	11	Mauritius	4	Mauritius	6	S. V. G.	11	Tanzania	4	Gambia	5	Panama	14	Panama	6	Zambia	6
Zambia	11	Bahamas	3	Guinea	5	Zambia	11	Belize	3	Guinea-Bissau	5	C. F. A.	13	Tunisia	6	Guinea	5
Paraguay	10	Libya	3	Burkina Faso	4	A. B.	10	Ecuador	3	Macedonia	5	Congo	13	Belize	5	Saint Lucia	5
Bermuda	9	Paraguay	3	Georgia	4	Georgia	10	Mauritius	3	Mali	5	Zambia	13	Costa Rica	5	Paraguay	5
Gambia	9	Rwanda	3	Panama	4	Gambia	10	Bahamas	2	S. V. G.	5	Burundi	12	Dominican Republic	5	A. B.	4
Chad	8	Senegal	3	Chad	4	Angola	8	Bermuda	2	Zambia	5	Benin	12	Somalia	5	Benin	4
Uruguay	8	Somalia	3	Bermuda	3	Bolivia	8	C. F. A.	2	Angola	4	Saint Lucia	11	Congo	4	Bolivia	4
Turkmenistan	6	Tunisia	3	Bolivia	3	Kazakhstan	8	Congo	2	Georgia	4	Cameroon	9	Tanzania	4	Cameroon	4
Burkina Faso	5	Uzbekistan	3	Guinea-Bissau	3	Bahamas	7	Cyprus	2	Mauritius	4	S. K. N.	9	S. V. G.	4	Panama	4
S. K. N.	5	S. V. G.	3	Mali	3	Panama	6	Georgia	2	A. B.	3	Sierra Leone	8	Bermuda	3	Senegal	4
Mali	5	Burundi	2	Uzbekistan	3	Chad	6	Guinea	2	Mauritania	3	Guinea	7	Grenada	3	Turkmenistan	4
Cyprus	4	Cyprus	2	Zambia	3	Cyprus	5	Saint Lucia	2	Niger	3	Madagascar	7	Kazakhstan	3	Angola	3
Guinea-Bissau	4	Dominican Republic	2	C. F. A.	2	Mauritania	5	Nicaragua	2	Chad	3	Angola	6	Sudan	3	S. K. N.	3
Mozambique	4	Georgia	2	Congo	2	Niger	5	Singapore	2	Uruguay	3	Mozambique	5	Zambia	3	Sierra Leone	3
Madagascar	3	Guinea	2	Grenada	2	Bermuda	4	Sierra Leone	2	Bahamas	2	Senegal	5	Bolivia	2	Burkina Faso	2
Mauritania	3	Grenada	2	Mauritania	2	Madagascar	4	Tunisia	2	Belize	2	Turkmenistan	5	Eritrea	2	Bermuda	2
Niger	3	Liberia	2	Niger	2	Turkmenistan	4	Uzbekistan	2	Ecuador	2	Burkina Faso	3	S. K. N.	2	Gambia	2
Seychelles	3	Panama	2	Paraguay	2	Paraguay	3	S. V. G.	2	Honduras	2	Gambia	3	Libya	2	Madagascar	2
Gabon	2	Uruguay	2	Uruguay	2	S. K. N.	2	Zambia	2	Madagascar	2	Mali	3	Paraguay	2	Mali	2

Notes : This table reports the number of dissemination areas (DAs) in which there is at least one person of the reported ethnic origin. 'Poor' ('rich') DAs are DAs in the bottom ('top') quartile of the metropolitan income distribution. For example, there are 426 DAs with income in the top quartile in Calgary with positive population of ethnic origin 'U.K.'

2.6 Conclusion

We have explored the causal effects of exogenous country-level measures of cultural, religious, linguistic, and genetic proximity between populations, as well as of historico-political relationships, on the colocation patterns of these populations in Canadian cities. We find that, conditional on geographic and economic controls,

these variables have a statistically strongly significant impact on the exposure of different groups to one another : sharing the same language or religion, being genetically closer, and having common past colonizers makes populations colocate more. These results are robust to identification concerns, a large set of alternative measures of our key covariates, and across both the 2016 and 2006 census waves. The effects also vary across cities and display an east-west gradient, with preferences over language, religion, and past colonial relationships playing a larger role in eastern than in western Canada.

Our results confirm that "near things are more similar than distant things." Being similar along non-spatial dimensions, when coupled with homophily, seems to be one explanation for the observed stratification of cities. Our results may also shed light on a preference-based explanation to the existence of cities : cities are places that provide 'ethnic variety', and if people want to interact with similar people they can get better matches for interactions in larger cities—which are more diverse—than in smaller places. This may explain in part the somewhat puzzling importance and persistence of sorting of people, especially immigrant minorities, into urban areas, despite poverty, crime, and congestion. Exploring the causal effect of ethnic diversity on city size and sorting thus seems to be an exciting extension for future research.

CHAPITRE III

RACE AND FIRM LOCATION : WHO MOVES WHERE ?

Abstract

This paper provides a measure for, and empirical application of, the spatial mismatch hypothesis in order to better understand patterns of firms and individuals locations. We first explore how the dynamic of decentralization in the New York area affected different groups in a different magnitude between 1990 and 2010. We then test the spatial physical disconnection between individuals and their potential employers. The article shows a robust empirical regularities of the effect of race and poverty on the spatial mismatch. For instance, we find that White shifted towards jobs while Black, Hispanic and Asian shifted away from total employment. Within each group, the shift between jobs and poor individuals is even more pronounced.

Keywords : Spatial Mismatch, Decentralization, Race, New York.

JEL Classification : R2 ; R3.

3.1 Introduction

The unemployment rate was around 3.9% in the U.S. at the end of 2018. However, this varies across groups and there is disparity in unemployment across race and ethnicity. While White and Asian workers had the lowest rate with 3.1% and 3.2% respectively, Black had more than twice as much with 6.5% and Hispanic with 4.5%.[1] These racial gaps remained unchanged over several decades and the most classical reasons indicate that discrimination and educational attainment of minorities are two of the major reasons. This paper sketches the geography of racial opportunities and the location of minorities as a potential ingredient of racial gaps in unemployment and other labor market outcomes.

Precisely, on the one hand, we look at the decentralization patterns in the New York Metropolitan Statistical Area (NYMSA) and explore the joint distributions of jobs and active populations to analyze how race and poverty can reshape these patterns between 1990 and 2010. On the other hand, we revive Kain (1968)'s Spatial Mismatch Hypothesis (SMH) and test the actual physical disconnection between population and employment, racial groups, and their opportunities as well as the poor of each group with their potential employers. For population, we gather 1990, 2000, and 2010 census data for the NYMSA at the smallest geographic level, with information on race, income, and education. For firms, we use data for over two million plants with information on employment and industry classification. We also gather other individual data from the Current Population Survey (CPS) that provides information on the national distribution of jobs by race and industry, which we use to perform our measure of SMH.

Empirical and theoretical work has shown the importance of this question. Physi-

1. U.S. Bureau of Labor and Statistics (2019)

cal disconnection is potentially harmful for employment, income, and other labor outcomes. First, being far from the potential place of work is inefficient for job search. Job seekers struggle to identify potential employers that are distant and with which they are unfamiliar, they only search efficiently around their residences (Wasmer et Zenou, 2006). Second, housing prices tend to be lower in their residence, providing no incentive to search actively or even to move close to the location of jobs (Patacchini et Zenou, 2006). Finally, the cost of search (if unemployed) or commute (if employed) increases with distance to jobs which gives less incentives to search farther (Coulson *et al.*, 2001). If workers experience a long commute, they are likely to be tired when they are on the job and are less productive. This is a real concern for workers who don't have cars or an efficient public transport system with a complete network, efficient interconnection nodes, and synchronization between transport modes.

This paper makes progress on three major points regarding the spatial mismatch. First, we provide a novel way to test the SMH. We adopt the Duranton et Overman (2005, 2008) continuous measures of firm colocation and extend it to capture the physical distance between people and jobs. It allows us to overcome the lumpiness of existing measures that suffer from the so called Modifiable Areal Unit Problem (MAUP). Moreover, these measures are fairly flexible to test different benchmarks which results in rich interpretations. We take advantage of a finer data on population at the census block level and on firms at a single location to measure most accurately job shifts and spatial disconnection. Second, most of the previous empirical work on SMH focuses on Black and White comparison. In this paper, we will extend to Asian and Hispanic and explore the poverty factor on both decentralization and spatial mismatch. Third, we push the previous literature and go beyond testing only the spatial disconnection between total jobs and population. We refine the notion of opportunities and define potential employers

using a nationwide distribution of employment by race and poverty. This will help us to obtain an accurate picture of the likely employers that hire each group.

There are several notable findings. First, there is a decentralization of jobs in NYMSA while total population seems to be stable between 1990 and 2010. However, poor population shifted away from the city center. The parallel shift of total jobs and the poor is suggestive of a bigger disconnection of poor and jobs in general. Second, race shapes the patterns of the decentralization of people. Indeed, we find that White shifted closer to jobs over the two decades while Black, Hispanic and Asian shifted away from jobs. When we look at poverty within these racial groups, we find that the shift is even more substantial for the poor groups. Third, applying our measures of spatial mismatch, we find that spatial disconnection increased over time : people are less close to jobs between 1990 and 2010. Fourth, minorities tend to be far from total jobs than White. Also, not the least important, looking at opportunities instead of total jobs, our measures show that White and Asian are closer to their opportunities than Black and Hispanic. The similar ranking of spatial mismatch and unemployment across race is probably suggesting that the more you are spatially mismatched from your opportunities, the more likely you suffer from unemployment.

This paper is related to three main literatures. First, the literature of job decentralization. Glaeser *et al.* (2001) show the decentralization of jobs in U. S. MSAs and the focus was more on jobs and industries. In our paper, the richness of data on firms allows us to explore deeper the jobs-population gradient by looking at potential employers and population of different groups, and how poverty affects these patterns. Second, our paper is related to the literature of colocation measurements. More specifically, we are related to two strands : measures that look to the exposure-isolation dimension of segregation (Reardon et O'Sullivan (2004), Behrens et Moussouni (2018)) and those that look at patterns of industry co-

agglomeration (Ellison *et al.* (2010) ; Duranton et Overman (2005)). We extend these two strands and provide new insights on how to measure the spatial colocation of firms and people. Last, the literature of SMH. Our paper is a revival of the Kain (1968)'s landmark work on the effects of race on housing markets and labor market opportunities. In our paper, we push further this literature and emphasize the effect of different racial groups and poverty on the spatial colocation of population and their suitable jobs by using a new continuous measure.

The paper is organized as follows. The next section 3.2 describes the data employed in this paper. Section 3.3 lays out the empirical strategy and the measurement used to test decentralization and spatial mismatch. Section 3.4 presents the results with robustness checks. Section 3.6 concludes. Last, the appendix material contains extra tables and figures.

3.2 Data

To explore the link between firms and population of different groups, we use mainly three sources of data : U. S. Census, National Establishment Time Series (NETS), and Merged Outgoing Rotation Groups (MORG) data. We use the first for population location by race, the second for firm location by industry, and the third allow us to construct a measure of opportunities that we define as the firms that are likely to hire a specific group. We will detail later on how we proxy for these.

Census data. For population, we use the 1990, 2000 and 2010 US decennial census data for 25 counties that form the NYMSA.[2] It provides information on

2. We extract all the data from the National Historical Geographic Information System (Ruggles *et al.* (2016)).

population counts by race and education at the finest geographic level, i.e., the block level, for which we have the centroid coordinates that let us to look at the physical distance to employment. Most of the literature uses bigger geographic units (e.g., census tracts) which is convenient for data availability and stability over time. We will explain later the importance to proceed differently and use the smallest geographic levels, especially when it comes to look at micro geographic phenomena. Since we analyze employment and active population, we only keep blocks with no-zero counts of individuals between the age of 18 to 62 and aggregate race into four major groups : White, Black, Hispanic, and Asian. [3]

Table 3.1 provides summary statistics on population by groups for 1990, 2000 and 2010 respectively. One piece of evidence this table shows is the importance to look at other ethnic groups rather than focusing only on Black. Indeed, Black were the largest minority group in 1990, but this changed over time. In 2010, the Hispanic group is more important than any other groups. These tables also show implicitly the spatial distribution of groups. For instance, in 1990, the maximum value of Black count at the block level is 11147 compared to 6981 for White, who seem to be present in more blocks than any other groups. This suggests more dispersion and less segregation. Looking only at the three minority groups suggests also that Hispanic are the most dispersed. This is probably due to our aggregation of Hispanic that includes both White and Black Hispanic.

The Census also provides information on total income by race for the last 12 months of 2010, 2000 and 1990 censuses. They are before tax and inflation-adjusted for the release year and include all employment and self employment

---

3. The Census does not provide directly these groups. We aggregate the racial composition of blocks and define White from the Census table "White alone, not Hispanic or Latino". Similarly, Black from the "Black or African American alone" table, Asian from"Asian alone" table and Hispanic from "Hispanic or Latino" Census Table which may include White or Black Hispanic.

TABLE 3.1: Summary Statistics : Population

	2010							
	White		Black		Hispanic		Asian	
	all[a]	active[b]	all	active	all	active	all	active
mean	56.87	34.28	40.17	25.98	33.86	22.2	21.08	14.94
min	1	1	1	1	1	1	1	1
max	3378	2577	5276	4987	2801	2668	3511	2345
total pop.	9.7	5.72	3.43	2.1	4.42	2.77	1.89	1.27
# blocks	170736	166847	85368	81190	130705	124939	89996	85250
	2000							
mean	63.72	38.41	48.56	30.26	33.32	21.7	18.92	13.4
min	1	1	1	1	1	1	1	1
max	5599	4397	13969	8432	5530	3406	4485	3028
total pop.	10.28	6.07	3.38	2.01	3.62	2.25	1.36	0.92
# blocks	161353	158225	69697	66513	108750	103681	72287	69104
	1990							
mean	71.6	48.93	52.83	37.39	30.94	20.32	14.2	10.34
min	1	1	1	1	1	1	1	1
max	6981	5610	11147	9460	8637	8333	3005	2035
total pop.	10.89	7.33	2.91	1.97	2.71	1.69	0.82	0.54
# blocks	152187	149989	55032	52695	87721	83456	58317	55571

Notes : Zero counts are dropped and means are at the block level. Total population is per million for all NYMSA. [a] all age groups. [b] active population between 18 and 62 years old.

incomes. However, it is at the block group level. To break it down to the block level, we use racial population counts to obtain average income at the block level. [4]

NETS data. NETS [5] covers the universe of employment in NYMSA. It originally aimed to construct credit score indices, therefore it contains an extensive information at the establishment level. Since Census data are only available for decennial waves of 1990, 2000 and 2010, we will similarly focus on employment for those years to explore the population and firm dynamic. Thus, as shown in figure 3.1, we have access to 0.65, 0.87 and 1.27 million establishments with a unique Data Universal Numbering System (DUNS), and total employments of 9.69, 10.53 and 10.18 millions by up to 6 digit NAICS industry code for 1990, 2000, and 2010 respectively. Every plant is a commerce, service, or manufacturing unit at a single physical location with the exact latitude and longitude coordinates, total employment count and industry classification. NETS measures the number of jobs rather than the number of workers and it does not report any employee information, such as education, wages, or race. [6]

MORG data. The weakness of NETS is that it does not inform us about workers characteristics such as race, education and wages, which is needed in our analysis.

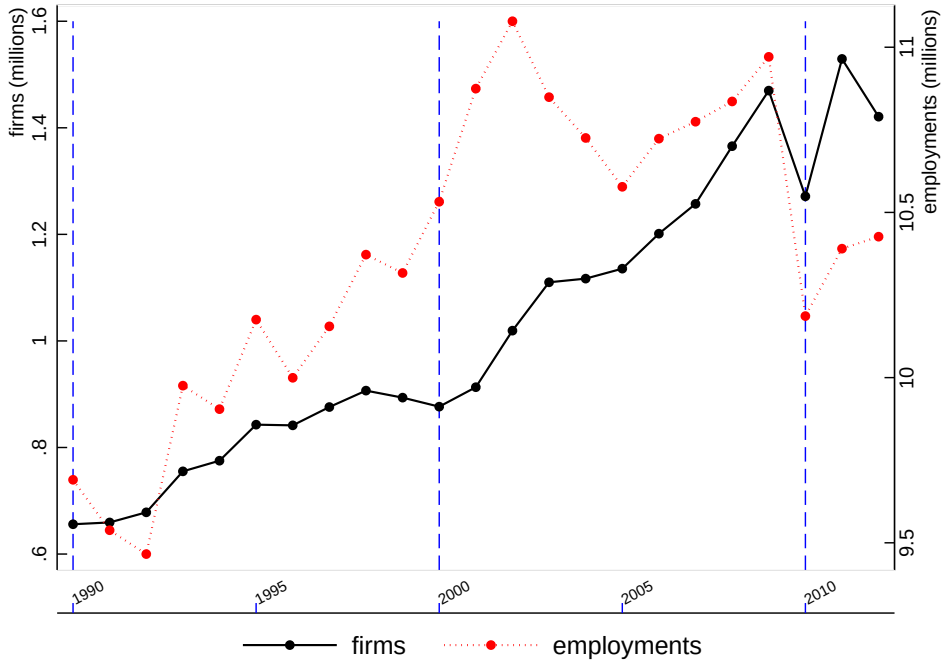
---

4. We also drop block groups where income is zero or negative, which may contain prisons, public building, etc.

5. This is a joint project where Walls & Associates link Dun & Bradstreet cross section data into longitudinal data that go from 1990 to 2012 (2015 version).

6. An employee at two establishments would be counted twice, and the employment counts do not separate full and part-time work. Also owners count as employees, for instance if a sole proprietor has two employees, NETS counts it as three employees. This might overestimate employment but we don't think that it is too problematic.

FIGURE 3.1: Employments and plants count (1990 to 2012)



Thus, we make use of a third source of data that is the Current Population Survey (CPS) that give us monthly information on labor market characteristics for 60 000 households, stratified to better represent the U.S. population. Each respondent is solicited to provide data on itself and other individuals in the dwelling that have 16 years old or over. Every household is followed monthly for 4 months, then not observed for 8 months and finally observed for another 4 months. In the 4th and 8th month of interview, information such as earnings, education and race is assembled into the Merged Outgoing Rotation Groups (MORG) by the Bureau of Labor and Statistics (BLS). We aggregated the MORG from 2000-2007 and 2011-2015[7] which gives us approximately 4 millions observations.

Ideally, we could use this database for population since it is disaggregated to the

7. We do so to keep industry classifications stable and avoid the 2008 financial crisis.

household level. However, for a spatial analysis, it has a major weakness since the household location is only reported at the county level rather than census block or point level. Instead, we use this data to proxy for opportunities that we define as firm of a particular industry that are likely to hire an individual of a particular group. To this end, we take all U.S. counties except the ones that form NYMSA and construct a wide average of characteristics by race and industry. Precisely, we use MORG and compute nationwide race-industry-shares for each group and obtain a ranking of 276 industries that range from 4 to 6 NAICS digits. We will explain later the thresholds we choose to define opportunities and how it might be problematic to take the White group as they are the majority group.

3.3 Empirical Strategy

In this section we motivate how decentralization and housing segregation are potential ingredients for adverse labor outcomes of minorities. We then explain how we proceed to measure spatial mismatch and discuss benchmarks and opportunities.

3.3.1 Decentralization and Spatial Mismatch

In the early 20^{th} century, workers and firms were located near one another and people walked to their jobs. The introduction of cars has shifted people's location who lived in suburban areas and worked in central cities in the mid 20^{th} century.[8] With the amplification of public transit networks and the democratization of

8. There are numerous other reasons peculiar to the postwar period such us home insurance mortgage by the US administration as well as racial tensions.

cars [9], both people and firms have more incentives to move out of the city. In the U. S., 57% and 70% of residents and jobs were located in central cities in 1950, and this decreased to 37% and 45% in the 1990s (Mieszkowski et Mills, 1993). At the end of the 20^{th} century, most U.S. metropolitan employees worked more than 5 miles from the central city (Glaeser *et al.*, 2001).

Decentralization is a result of the trade-off between the benefits and costs of density. On the one hand, consumers will choose to locate closer to the city that offers higher amenities, public goods and less commute, but this comes with its downside of higher housing prices, crime and pollution. On the other hand, firms locate in Central Business Districts (CBD) and gain from agglomeration economies, flows of ideas, and access to consumers, but pay high land prices, constrained by transportation networks, and its cost. [10]

There is a general consensus of employment decentralization [11], and it is slightly more pronounced than population decentralization. Yet, the gradient is still strong where population and employment track each other continuously over time and space. Nevertheless, in this paper we argue that there are heterogeneous colocation patterns across firms of different industries and populations of different groups, and this might mitigate the general picture of the gradient. While intensive land-use industries, i.e., manufacturing, are doubtlessly localized farther downtown, this is not the case for consumer-oriented industries, i.e., services, that will ultimately choose core downtown. Similarly, on the population side, it has been observed that poor, immigrants and minorities tend to sort into CBDs, whereas suburban

---

9. By 1990, car commutes account for 83% of total commute in the U.S.

10. See Krugman (1991) on the interaction of economies of scale with transportation cost.

11. See Mieszkowski et Smith (1991) for Houston, Giuliano et Small (1991) for Los Angles, McMillen et McDonald (1998) for Chicago, and Macauley (1985) for 18 other US cities.

areas attract high income and the majority group.[12] For instance in 2000, 64% of Black lived in CBDs whereas it is only 28% for White in the largest U. S. MSAs (Gobillon *et al.*, 2007).

Therefore, this rises the question of "who is close to what ?" This paper aims to explore this heterogeneity in the suburbanization of economic activity and population sorting to analyze the joint distribution. Precisely, we investigate two key aspects. We first look at employment dynamics between 1990 and 2010 and provide a bird's view of decentralization in NYMSA by testing whether this affects specific groups of population. In a second step, we apply a measure of colocation to test and explore how jobs and workers follow each other. In other words, we analyze the physical distances between people and firms of industries that are likely to hire a specific group. Our motivation comes from the possible interactions of spatial segregation, that creates spatial inertia for minorities, and the decentralization of some specific firms that might be potential employers for these groups. We define spatial mismatch as the physical disconnection of people of some racial groups from their potential employer, and this physical disconnection might lead to adverse local labor outcomes such us unemployment and lower income.

This idea first originated in Kain (1968), credited as the father of SMH. He advanced that segregation of African Americans and decentralization within metropolitan areas lead to a physical disconnection between places of residence and those of work. Kain asserts that in some places, likely segregated, there is a surplus of workers of some groups (say African Americans). Outside of these locations, likely where firms decentralized, the housing discrimination towards minorities (say

12. There are a few exceptions in the U. S. case (e.g., New York City hosts the poorest and the richest). Western European cities are also structured differently, where high income sort into CBDs and low income further away from the city.

African Americans) generates a labor supply shortage. In a nutshell, the idea is that cities, like suburban areas, have centrifugal and centripetal forces, and the net sum drives the location of firms and population. However, these two forces are not the same for all industries and ethnic groups. The pull and push factors for firms, combined with higher inertia of some ethnic groups' location, might create a spatial skill mismatch. This paper aims to explore this heterogeneity in the location choice and shed light on the role that race and poverty play.

3.3.2 Measurement

The magnitude of spatial mismatch is measured typically by indices that capture the 'imbalance' between population and jobs.[13] These measures encompass generally, in specific areas, the composition in terms of jobs and population, commuting time, distance, or cost. Nevertheless, this class of indices suffer from serious issues. Computing them at given areas (counties, tracts or even MSAs) automatically rises the question of which geographic unit to use, and brings up the so called Modifiable Areal Unit Problem. The high level of aggregation and lack of 'realistic' benchmarks that enable us to test the SMH are also a major challenge.

In this paper, we adopt the Duranton et Overman (2005, 2008) framework of firm coagglomeration to firm and population colocation, and look at the spatial mismatch in the physical or geographical sens. Precisely, we use Behrens et Mousssouni (2018) continuous measures of segregation and replace one of the groups by employment. This measure allows us to overcome partially the problems mentioned previously. The idea is intuitive : we explore the between-groups dimension of segregation and substitute jobs for one of the racial groups, which allows us

13. Variety of Dissimilarity indices have been widely used to measure both segregation and spatial mismatch. See Martin (2004).

to measure the degree of the *physical isolation-exposure of certain jobs of a given sector with the location of people of a given race.*

Concretely, consider two agents : population of characteristics x (e.g., Black, Asian, etc.), and firms of characteristics y (say industry type). There are n_i^x population counts of group x located at the centroid of a census block i with $i = \{1, 2, ...L_x\}$, and n_j^y employments counts of a firm belonging to an industry y located at a point j with $j = \{1, 2, ...L_y\}$. We then compute all $L_x L_y$ bilateral distances d_{ij}, which we kernel smooth using the following formula :

$$\widehat{k}_{ij}^{xy}(d) = \frac{1}{h \sum_{i=1}^{L_x} \sum_{j=1}^{L_y} n_i^x n_j^y} \sum_{i=1}^{L_x} \sum_{j=1}^{L_y} n_i^x n_j^y f\left(\frac{d - d_{ij}}{h}\right), \tag{3.1}$$

where $f(\cdot)$ is a Gaussian kernel and h is the bandwidth parameter set using Silverman's rule-of-thumb.

Recall, by doing so, we assume implicitly that population of group x are located at the centroids of census blocks L_x [14], whereas the exact latitudes and longitudes L_y of firms with employment counts are known. The estimator in (3.1) gives us, for each distance d, the share of bilateral distances between people of groups x and jobs of industries y.

To obtain an aggregated 'absolute' measure of spatial mismatch, we naturally use the cumulative distribution as follows :

$$\widehat{K}_{ij}^{xy}(d) = \int_0^d \widehat{k}_{xy}^{ij}(\zeta) \mathrm{d}\zeta. \tag{3.2}$$

The measure (3.2) suggests what share of bilateral distances between group x and their opportunities y is smaller than d. For instance, a $\widehat{K}_{ij}^{xy}(1km) = 0.3$ for $x =$

14. This is not a major issue since block centroids are population weighted.

Blacks and y = their employment opportunities means that 30% of their bilateral distances are less than 1 kilometer. Or there is 30% chance that a random draw of one Black and one potential job yields a pair that is less than 1 kilometer from one another. The larger $\widehat{K}_{ij}^{xy}(1km)$, the less Blacks are spatially mismatched with their opportunities.

Discussion. Despite the flexibility and richness of our measure to overcome challenges of existing indices, a few observations are in order :

1. Routes and distance. A bilateral straight line distance d_{ij} is an underestimate of the actual route that individuals take, and understates actual travel distance even more for shorter trips on the ground. But, we argue that the measurement error tends to decrease for longer trips where one can deviate less from a strait line path and detours tend to be smaller.
2. Time and distance. Distance is not always evenly correlated with time spent traveling. While in denser areas small distances might take longer time duration, suburban resident are likely to experience the inverse and commute longer distance but faster. Consequently, we face a negative bias in the core downtown area where congestion is high, and a positive bias in the peripheries and suburban areas. This depend obviously on the transportation mode of the worker.
3. Costs and distance. This is one of the crucial mechanisms of SMH since the distance will matter if it is costly.[15] For the same trip, going to work (or seeking a job) by public transportation will cost less than taking a car (fuel, parking, insurance, etc.). Distance will then fail to proxy for cost. For

15. Importantly, poor workers may not earn enough to support long commutes or even move to the suburbs.

example, if someone takes a train, a flat fare applies over a considerable distance and this is not the case for a car trip. Thus, our measure might overestimate spatial mismatch for individuals that have access to a good transportation network, although, one can argue that public transportation is also time and energy consuming. Think about waiting and walking time or some radial axes that require traveling downtown to connect to another line. This might offset the direct pecuniary cost.

In a nutshell, straight line distance captures differently time, route system, and financial costs, which at the end of the day depend on individual's locations. One way to overcome these issues is by subsampling to homogeneous agents that face similar constraints. However, in our case, a one-direction bias can be also addressed by an appropriate benchmark. Indeed, comparing our measures against counterfactual distributions might help partially to reduce these type of measurement error since both distributions have similar location set constraints. In other words, comparing the observed distributions to the appropriate benchmarks that carry similar route, time, and cost constraints will allow us to overcome the 'strait line distance' issue, and reveal more accurately who is more spatially mismatched. We discuss more this matter in the next section.

Benchmarks

The measures (3.1) and (3.2) are absolute measures that translate the observed joint distribution of opportunities and groups. As in Duranton et Overman (2005, 2008) for firms and Behrens et Moussouni (2018) for populations, setting out the benchmark against which we measure empirical distributions is of paramount importance. Stating if a group is spatially mismatched and what is the magnitude of such is closely related to how we define counterfactuals, which is not an easy

task. While for firms the idea of benchmarks is to define random distributions that occur in the absence of 'any type' of physical constraint, and similarly for population the random counterfactuals that occur in the absence of any type sorting (by race, income, etc.), combining both firms and population makes it even trickier to set out a 'reasonable' benchmark.

Indeed, what is the benchmark in our case ? It should be the distributions that generate random spatial mismatch. At first glance, it would be the distributions that we observe if jobs and people are randomly picking up locations with no constraint, and this might generate 'random' spatial mismatch if people (or firms) pick locations that happen to be far from one another. [16] In this case, we interpret the deviation from this benchmark as spatial mismatch due to location constraints or choices of both agents.

For instance, if minorities have a free location choice and if being far from place of work is harmful, then they will choose to be closer to their opportunities. In this case, any deviation from this counterfactual, and that result in spatial mismatch could be interpreted as a consequence of segregation. While this is likely true, decentralization of jobs and firms' location, as summarized in section 3.3.1, is also a potential ingredient of SMH. Ultimately, put together, it is likely caused by a sum up of firms and peoples' location choices. Thus, from a methodological point of view, even if there is no clear consensus on a specific definition or what counterfactual to use, the magnitudes and interpretations of SMH are highly dependent on the benchmark.

In our empirical analysis, and given the flexibility of our measures, we will rely on empirical-to-empirical comparisons and test several type of benchmarks. We

16. In a random world, there are strong reasons to assume that firms' and peoples' locations are not driven by the same factors.

first compare results of our measure in equation (3.1) and (3.2) between White, Black, Hispanic and Asian, and test who is more mismatched than others. Second, we make comparisons of these groups against the total population-employment distribution as another benchmark. We also compare minorities against White and their opportunities as another benchmark since they are the majority group. [17] Third, we compute d_{ij} between groups and total employment that we compare to d_{ij} between same groups and their opportunities. Fourth, we test for each group if poverty amplifies the spatial mismatch, since there are strong reasons to believe that the poor are more constrained in their locations. Finally, we also use a time-benchmark reference by comparing results of (3.1) and (3.2) for 1990, 2000, 2010, which allows us to look at the dynamic of spatial disconnection over time.

Opportunities

For individuals, we chose four major groups (Black, Asian, Hispanic, and White) in the NYMSA as well as the poor within each group. In a similar manner, yet less straightforward, we need to define firms that are potential employers. Again, as for the benchmark, the magnitude of spatial skill mismatch will depend on the definition of opportunities.

First, in a perfect world, we would have information on job vacancies that are skill-attainable for these groups. Unfortunately, at finer spatial scale, such an information is very limited. We then use information on job counts per firm location and assume implicitly that jobs are vacancies, which overstates the availability of employment since job seekers apply for vacancies not jobs. Nevertheless, job counts can proxy for potential employers and the likelihood of finding a job, especially

17. White might constitute a good benchmark since they are unlikely to face housing discrimination and are pretty flexible to move

for sectors that have a high turn-over.

Second, who are the potential employers for each group ? Most studies focus on all jobs, retail sector, blue collar, or unskilled jobs without setting out a rationale on these choices. In this paper, we dig more into this issue and use the U.S. national distribution of jobs by race and sector and back out a ranking of potential employers. We take the top 5%, 10% and 25% of each group job distribution and compute d_{ij} for all the employment of firms that belong to each of these thresholds. In the same manner, we obtain within each group the ranking of potential employers for poor which we proxy by high school drop out or college without degree. Another way is to estimate a predicted probability for each group and use industry fixed effect to back out another ranking. However, we might face endogeneity problems with the latter, we therefore chose the former.

Finally, there is another complication in constructing the opportunities : the level of competition is spatially uneven. Few people who are 20 kilometers from many vacancies is different than many people who are 20 kilometers from few vacancies. In other words, for the same number of vacancies, looking for a job is more difficult in locations (e.g., central city) where there are more seekers than in suburban locations where there are less seekers. A measure that does not take into account the locally competing markets might underestimate spatial mismatch in the central cities where there is a greater competition for jobs than suburban areas. However, we can control for the segmentation of labor markets in two different ways. On the one hand, since our measure are continuous and flexible, one can choose to only focus on central areas and its surrounding, and do robustness checks for different cut-offs. On the other hand, an appropriate counterfactual that controls for the overall distribution of jobs and population will allows us to assess the 'net' spatial mismatch conditional on local competition.

3.4 Results

In this section, we begin by presenting facts on decentralization. We then zoom in to different racial groups as well as the poor. In a second step, we compare how these groups and their opportunities are colocated in NYMSA from 1990 to 2010 using our measure of spatial mismatch.

3.4.1 Employment and population decentralization

While there is no general consensus on decentralization measurement, we make use of CDFs to explore whether people or firms have shifted away from the central city. We compare population and job decentralization, explore the effect of race and poverty, and then look at the dynamic between 1990 and 2010. Precisely, we use rings of different radius going from 0 to 50 km with 100 m steps, and in which we compute the share of firms and individuals with different characteristics. We then plot the cumulative distribution of each share against their distance from the CBD. [18]

Figure 3.2 shows the CDFs for total employment which we compare to that of all population (panel (a)) and that of poor population (panel (b)). To simplify, we focus on 20 km radius for our baseline results. [19] A vivid picture that we observe is that the monocentric model is a fairly good representation of NYMSA. Indeed, panel (a) shows that 40% of jobs are located less than 10 km from the CBD for approximately both years. [20]

---

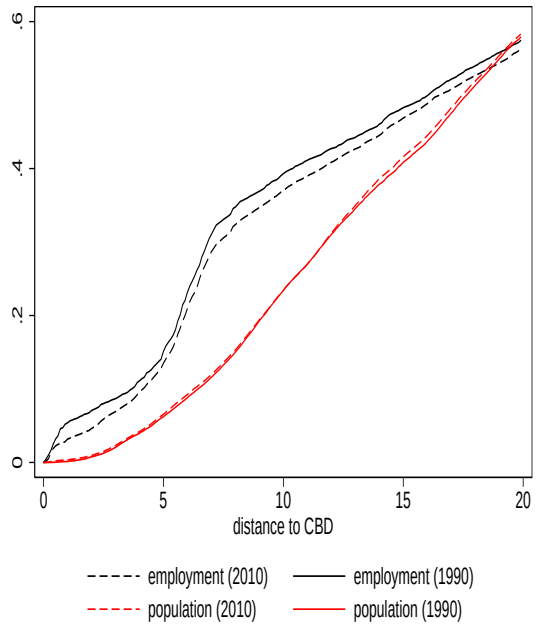
18. We choose the Wall Street centroid as the our reference for NYMSA CBD.

19. Results for above this threshold are presented in the Appendix.

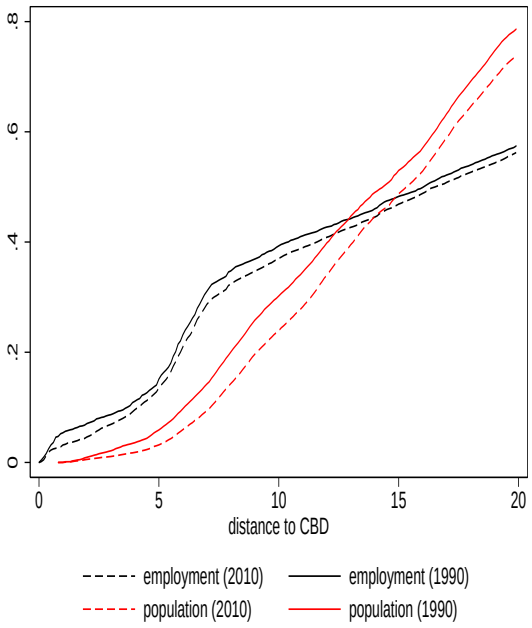
20. We do not report the 2000 year since there is no drastic change.

FIGURE 3.2: Employment and population shift (1990-2010).

(a) Population and employment.



(b) Poor population and employment.

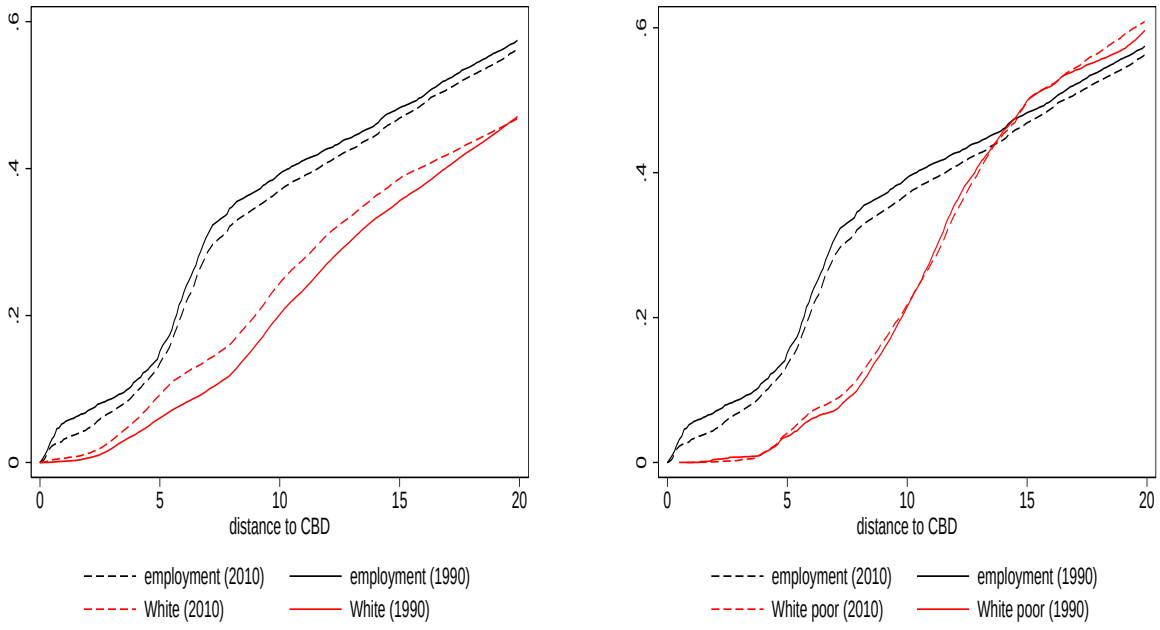


Notes : Black lines are for employments and red for populations. Dashed are for 2010 and continued for 1990. See figure 3.10 in the in the Appendix for bigger radius. Poor groups are defined as the bottom quartile of income distribution.

However, the picture changes somewhat when we look at the distribution of population. First, over the two decades, while there is a slight decentralization of firms (shown by a right shift of the black solid line), population seems to be me more stable over time in panel (a). But in panel (b), interestingly the poor seem to shift away for the center city as well as from jobs. Second, comparing the red and black lines show unambiguously that jobs don't follow people for both years. At smaller distance, the figure illustrates an excess of employment relatively to individuals while at bigger distance it is the opposite.

FIGURE 3.3: Employment and White shift (1990-2010).

(a) White. (b) Poor White.



Notes : Black lines are for employments and red for populations. Dashed are for 2010 and continued for 1990. See figure 3.11 in the in the Appendix for bigger radius. Poor White are defined as the bottom quartile of their income distribution.

Is this similar for all groups? The short answer is no. There is heterogeneity across racial groups, with a bigger shift for Black and Hispanic, and to a lesser

extent for Asian and White. Indeed, figure 3.3 shows that White shifted closer to jobs while poor White remained more or less constant over time, and far from general employment. If we consider White as the Majority group, and unlike the minorities they face less discrimination, and restricted location choice, this might explain partially their moving closer to jobs.

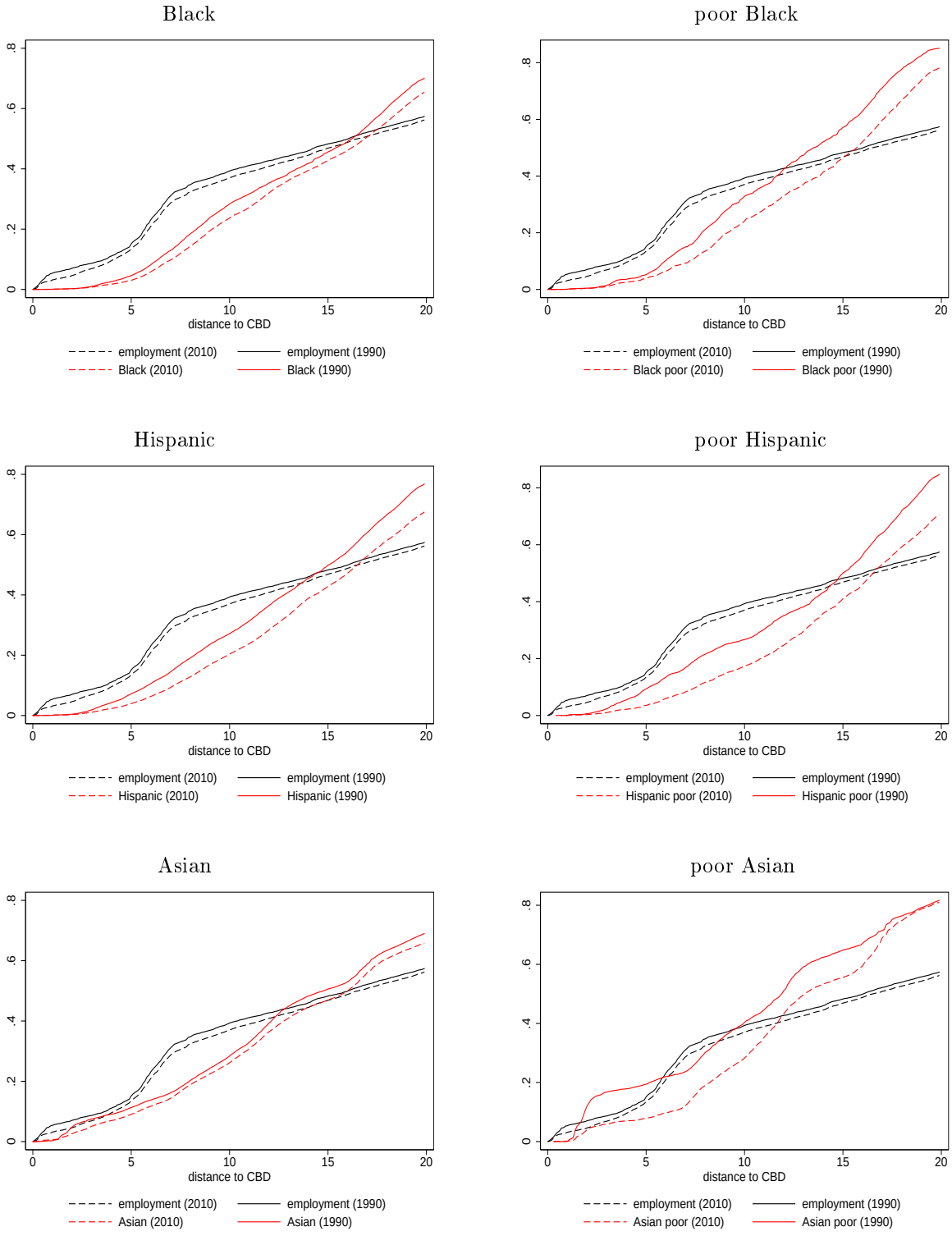
We now turn to the minorities. Figure 3.4 shows that Black, in opposition to White, shifted away from jobs. Similarly, but with a bigger magnitude, Hispanic also shifted away from employment between 1990 and 2010. Comparing panels (a) and (b) shows again that the shifts are stronger for the poor. One exception is for Asian, they tend to have similar patterns as White, and are closer to employment. Nevertheless, even if they do better in terms of closeness to jobs, the poor Asian have shifted significantly away from jobs in 2010. Historically, the Asian group was small and concentrated around the so called 'China Towns', yet it has grown and our figures show that it started to spread out.[21]

3.4.2 Employment and population disconnection

The previous section shows that between 1990 and 2010 employment locations shifted towards the overall population of NYMSA. Yet, when we focus on specific groups, we find that the poor shifted substantially away from general employment with a significant differences across race. One critical issue to keep in mind is the extent to which this is accurate at bigger radius. In other words, when looking at bigger rings, even if we have same shares of people and firms, we might have jobs on one side and people on the other side of the ring. Thus, to now explore

21. Table 3.6 in the in the Appendix shows a summary ranking of population of each group in a ring of 10 km radius that contains around 40 percent of total NYMSA jobs. The proportion of both Asian and poor Asian decreased over the two decades.

FIGURE 3.4: Employment and minorities shift (1990-2010).



Notes : Black lines are for employments and red for populations. Dashed are for 2010 and continued for 1990. See figure 3.12 in the in the Appendix for bigger radius. Each poor group is defined as the bottom quartile of its income distribution.

more precisely the 'real' disconnection, we turn to our measure in equation (3.1) and (3.2) and compute the bilateral distances d_{ij} between jobs and people. This allows us to explore the physical disconnection of each group with their suitable employment and opportunities.

We first look at the dynamics of our measures between the total employment and population, and explore again the effect of race and poverty. Second, rather than looking at total jobs, we explore opportunities of each group and their physical distances to people. We simplify by looking at a 2010 'snapshot' and only the top quartile opportunities of each group. [22] Last, within each group, we test again how poverty, on top of race, gives different perspectives.

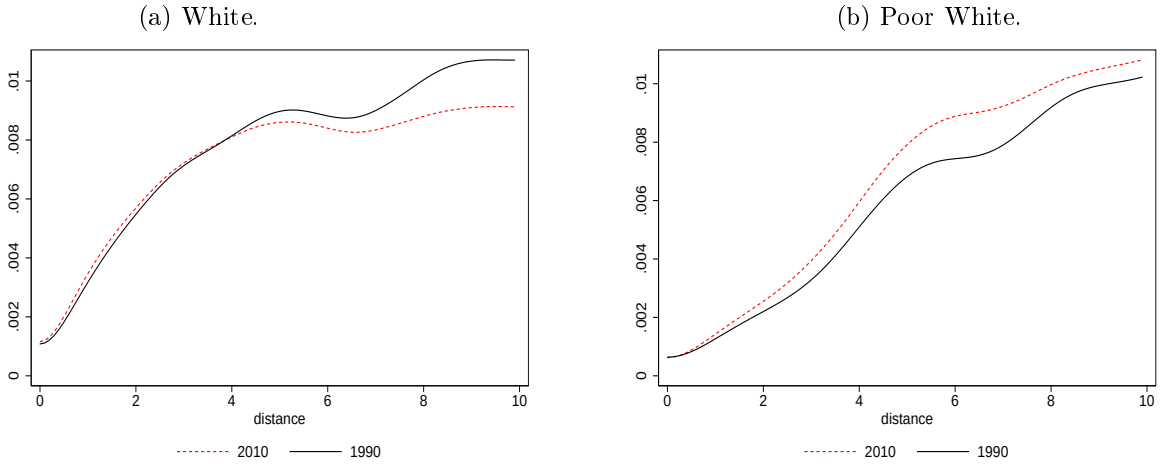
Employment and race

We look at White separately since they are the majority group and their patterns are likely to be similar to total population. Figure 3.5 looks at the equation (3.1) over time between total employment and White group in panel (a), and poor White group in panel (b). One fact that stands out is that White seems to get more close to jobs between 1990 and 2010 and this is more pronounced for the poor White. Indeed, the poor are more colocated with total jobs in 2010 than two decades before. One interpretation of this is that employment is likely concentrated in the core areas of New York and where simultaneously the the majority group tend also to locate since they are unlikely to face racial redlining.

How about minorities ? figure 3.6 shows clearly different patterns for minorities as well as the poor minorities. Both got more disconnected from total jobs over the two decades. Black, Hispanic, and Asian are closer to employment in 1990 than in

22. Results for top 5% and 10% as well as for 2000 and 1990 are available upon request.

FIGURE 3.5: $\widehat{k}_{ij}^{xy}(d)$ for employment and White (1990 to 2010).



Notes : See figure 3.13 in the in the Appendix for bigger radius.

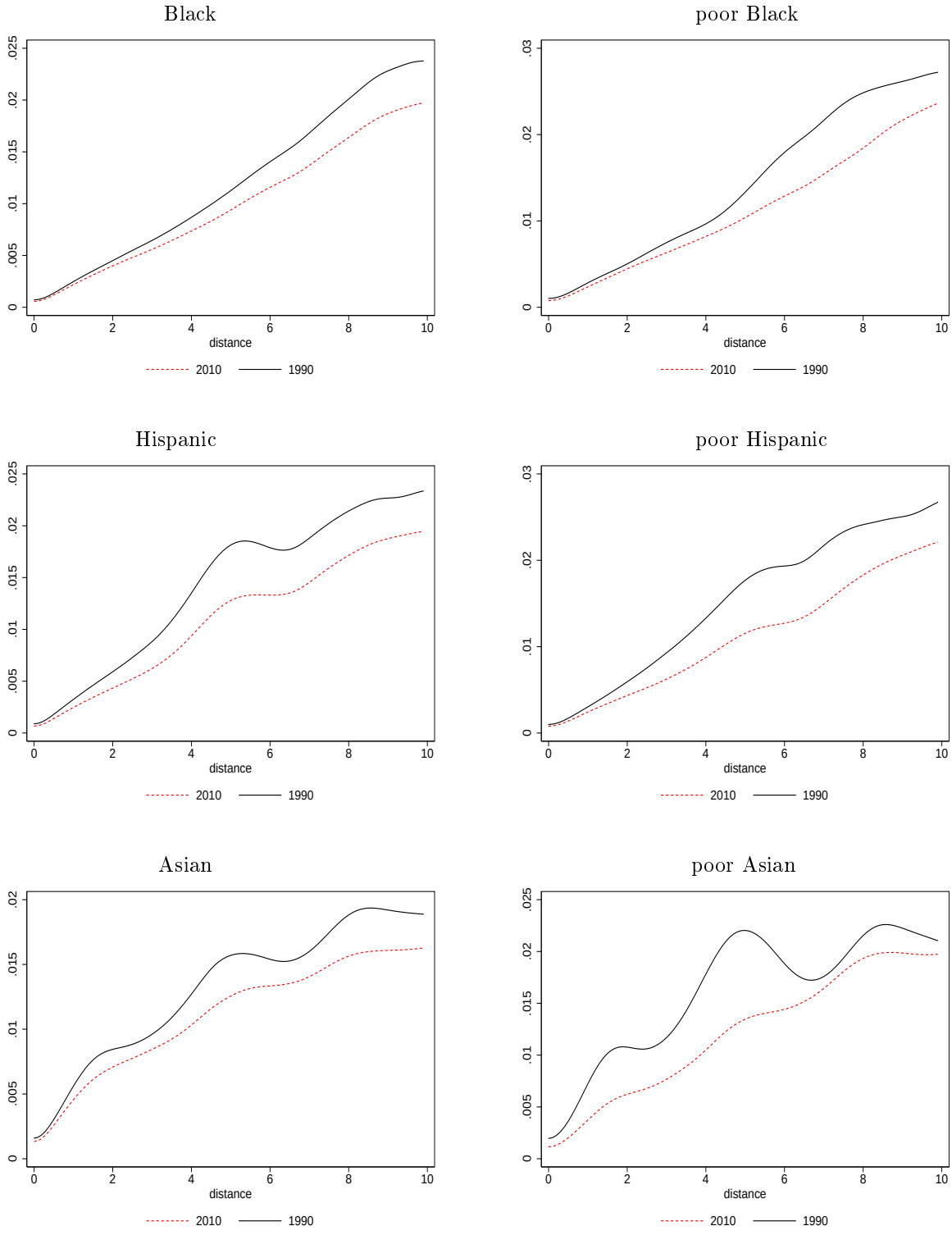
2010. The disconnection from jobs affects all minority groups as well as the poor of each group

Put together, figure 3.3 and 3.5 show that White shifted towards the city, and closer to total jobs. This suggests that even when we look at the actual physical distance between jobs and population, White and poor White seem to be better off and got closer to total jobs over the two decades.

Opportunities and race

Recall that in the previous section, we look at jobs in general. One valid concern, as for classical measures of SMH, is that all jobs are not potential employers and looking how groups are closer from them is not that relevant. Instead, it makes more sense to look at the potential employers only. To this end, we re-compute our measure of equation (3.1) and (3.2) and look how close each group is to its opportunities. In this section, we define the latter as the top quartile of national employers of each group. To have an idea about the ranking, table 3.2 shows the

FIGURE 3.6: $\widehat{k}_{ij}^{xy}(d)$ for employment and minorities (1990 to 2010).

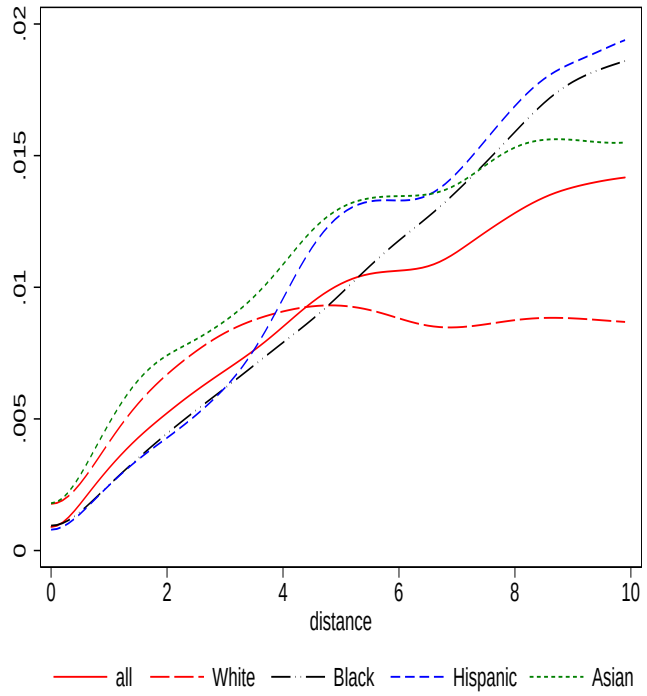


Notes : See figure 3.14 in the in the Appendix for bigger radius.

top five industries that are likely to hire each group.

Figure 3.7 shows the PDFs for total population and employment as well as each racial group with their respective opportunities. Exploring the PDFs, at small distance, we find that White and Asians are closer to their opportunities than Black and Hispanic. In other words, at smaller distance, White and Asians are the less spatially mismatched then Black and Hispanic. Table 3.3 shows the difference between the CDFs of each group with their respective opportunities, and CDFs of total population with total employment, for $d < 5\ km$. For both years, it shows clearly that White are closer to their opportunities (i. e., excess colocation), and that minority groups are father from their opportunities (i. e., excess dispersion), than the total population from employment.

FIGURE 3.7: Top quartile opportunities and race (2010).



Notes : See figure 3.15 in the in the Appendix for bigger radius.

TABLE 3.2: Industry Ranking : All

Race	Top 5 industries	Relative share[a]
Black	Tobacco manufacturing	2.85
	Taxi and limousine service	2.84
	Barber shops	2.75
	Bus service and urban transit	2.74
	Fiber, yarn, and thread mills	2.63
White	Coal mining	1.28
	Farm product raw materials, merchant wholesalers	1.25
	Lawn and garden equipment and supplies stores	1.23
	Other motor vehicle dealers	1.22
	Fuel dealers	1.22
Hispanic	Animal slaughtering and processing	3.50
	Cut and sew apparel manufacturing	3.26
	Fruit and vegetable preserving and food manufacturing	2.97
	Not specified metal industries	2.93
	Landscaping services	2.88
Asian	Nail salons and other personal care services	7.10
	Electronic component and product manufacturing	4.54
	Computer systems design and related services	3.63
	Computer and peripheral equipment manufacturing	3.46
	Software publishing	3.41

Notes : [a] We divide the share of each group in each industry by its overall share in the MORG sample. The higher the ratio, the more over represented the group is in a given industry.

TABLE 3.3: $\Delta\ \widehat{K}_{ij}^{xy}(d)$

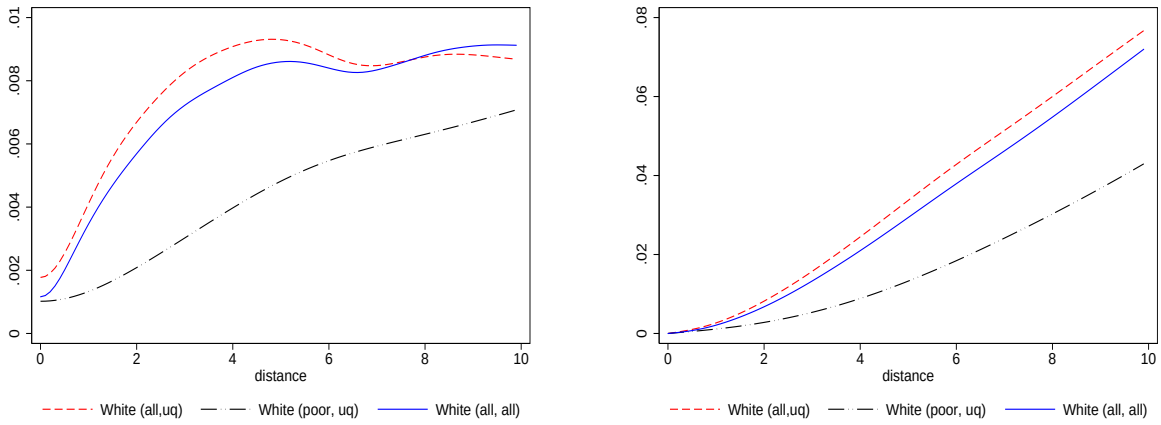
	2010				1990			
$d =$	White	Black	Hispanic	Asian	White	Black	Hispanic	Asian
0.5 km	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.0 km	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000
1.5 km	0.001	-0.001	-0.001	-0.001	0.001	0.000	0.001	0.001
2.0 km	0.002	-0.001	-0.001	-0.001	0.001	-0.001	0.001	0.001
2.5 km	0.003	-0.001	-0.002	-0.002	0.002	-0.001	0.002	0.002
3.0 km	0.004	-0.002	-0.002	-0.002	0.002	-0.002	0.002	0.002
3.5 km	0.004	-0.002	-0.002	-0.002	0.003	-0.002	0.004	0.004
4.0 km	0.005	-0.002	-0.002	-0.002	0.003	-0.003	0.007	0.007
4.5 km	0.005	-0.003	-0.001	-0.001	0.002	-0.003	0.010	0.010
5.0 km	0.005	-0.003	0.000	0.000	0.002	-0.003	0.015	0.015

Notes : For instance, $\Delta\ \widehat{K}_{ij}^{xy}(d)$ for White is the difference between $K_{ij}^{xy}(d)$ computed for White and their opportunities, and $K_{ij}^{xy}(d)$ computed for total population and employment.

One last exercise, we compare within each race whether poverty is a push or a pull factor. Precisely, in figures 3.8 and 3.9, we look at spatial mismatch between each race with its upper quartile opportunities (red small dashed lines), and the poor of each race with their upper quartile opportunities (black long dashed lines) and all population of that race with total employment (blue solid line). As for White, the figure shows that they are more close to their opportunities than total employment, but also more close to their potential employer than poor White to their opportunities. As for minorities, there is no clear patterns of the effect of poverty on top of race. Poor Black tend to be slightly more colocated than Black, poor Asian slightly more dispersed than Asian, while Hispanic and poor Hispanic have similar distributions.

FIGURE 3.8: Top quartile opportunities, White and poverty (2010).

(a) $\widehat{k}_{ij}^{xy}(d)$. (b) $\widehat{K}_{ij}^{xy}(d)$.



Notes : See figure 3.16 in the in the Appendix for bigger radius.

3.4.3 Robustness Check

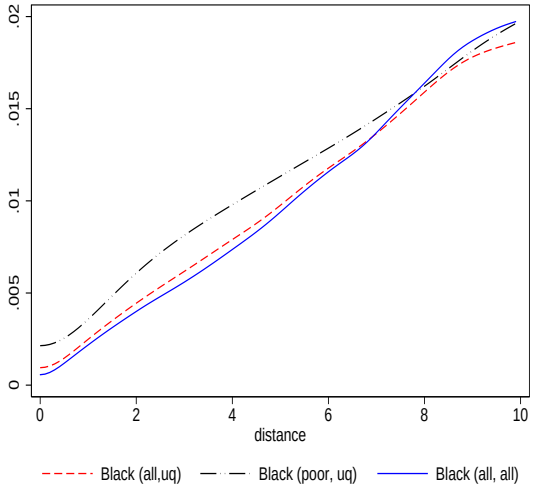
In this section, we present alternative measures that we apply to assess SMH. We implement two different sets of indices that are mostly used for segregation and firm concentration. We once again focus on top quartile opportunities and explore the effects of poverty of each group on the spatial disconnection in 1990, 2000 and 2010.

First, we make use of the Exposure index that measures the degree of potential contact between minority and majority group. In our case, it reflects the probability that a group shares an areal unit with its potential employers which we proxy similarly by the top quartile opportunities, and compute using the following formula :

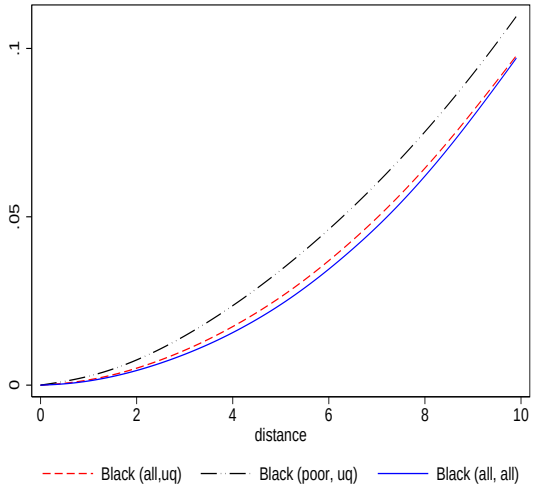
$$Expo = \sum_{i=1}^{n} \left[x_i/X\right] \left[y_i/t_i\right], \tag{3.3}$$

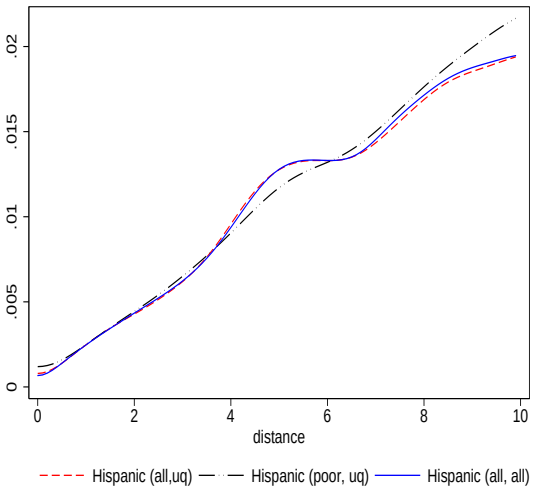
FIGURE 3.9: Top quartile opportunities, Minorities and poverty (2010).

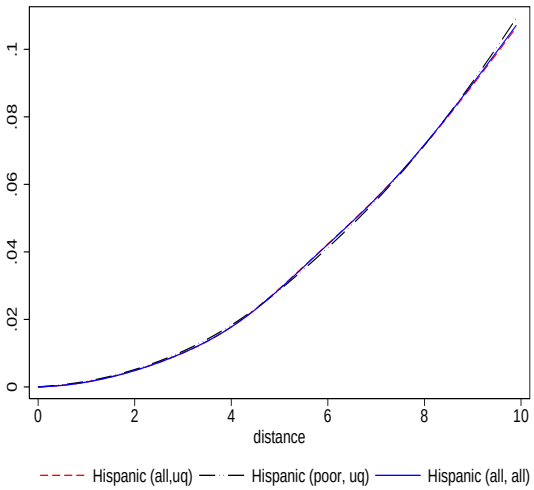
(a) $\widehat{k}_{ij}^{xy}(d)$.

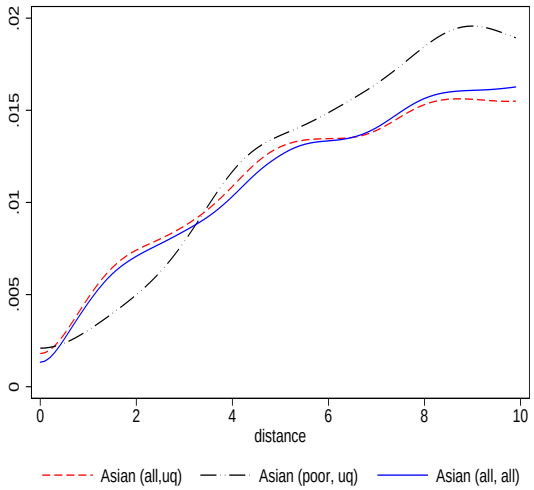


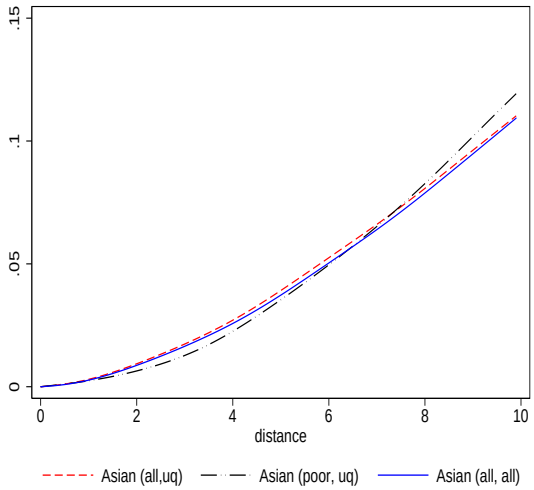
(a) $\widehat{K}_{ij}^{xy}(d)$.











Notes : See figure 3.17 in the in the Appendix for bigger radius.

where x_i and y_i are the total count in a unit i of respectively population of a given group and its employments opportunity, t_i is the total of both employment and population of that group at the same subarea i and X is the total population count of the same group in the whole NYMSA.

Second, we also adopt the Ellison *et al.* (2010) measure (henceforth, EG) that is used to exploit industrial colocation. We view this measure as a spatial covariance between each racial groups and its potential employers, and extend it to test the SMH using the following formula :

$$EG^{xy} = \frac{\sum_m (s_i^x - s_i)(s_i^y - s_i)}{1 - \sum_i (s_i)^2}, \tag{3.4}$$

where s_i^x is the share of racial group x located in geographic unit i ; s_i^y the share of employment opportunities y of a group x and where s_i is the share of both total active population and employments in i.

Exposure index range from zero to one. If it takes higher values then groups are more "exposed" to their job opportunities, i.e., less spatial mismatch. The EG index is slightly different since it takes negative values, but like Exposure index a higher values means more closeness to jobs.

Two comments need to be kept in mind. First, these two alternative measures suffer from the so called Modifiable Areal Unit Problem. This means that a random permutation of the units i will result in the same values of Exposure, and the EG index. In other words, any employment opportunity outside of i, even if it is contiguous, is not taken into account. This was not a problem for our baseline measure. Second, when it comes to compare over time, as we do over three waves, the unit i has to be stable. Otherwise, results maybe different over time for any 'resplit' of census blocks. This is also not of a big problem for our baseline measures since the kernel smoothing suffers less from the geographic unit spliting.

To address partially this issue for our alternative measures, we will use a concordance algorithm developed by Behrens *et al.* (2019) that allows us to obtain a stable unit.

TABLE 3.4: Exposure Index

	2010		2000		1990	
	all	poor	all	poor	all	poor
White	0.34	0.34	0.27	0.30	0.24	0.24
Black	0.29	0.22	0.26	0.17	0.20	0.11
Hispanic	0.32	0.29	0.31	0.24	0.29	0.24
Asian	0.49	0.40	0.45	0.33	0.46	0.39

Notes : Poor groups are defined as the bottom quartile of their income distribution. We use stable geographic units to perfom our comparaison over the three census waves.

Table 3.4 shows results of Exposure index for each group as well as the poor. Similarly to previous results, the table shows that Asian and White are more 'exposed' to their opportunities than Hispanic and Black. Comparing withing each group reveals that the poor seem to be less exposed to their opportunities than the rest of the population. Note that, there is no threshold above which we can state that there is a spatial mismatch, and in the segregation literature numbers above 0.50 are considered to be high. This rises again the geographic unit problem that does not take into account the 'real' spatial proximity and tend to overestimate segregation, or in our case the SMH.

In the same manner as table 3.4, table 3.5 shows the result for EG index. Comparing between races shows similar patterns. White and Asian seem to be less disconnected than Black and Hispanic, and this become more evident for the poor White and poor Asian who show more spatial connection to their jobs then other poor groups.

TABLE 3.5: Ellison-Glaeser Index

	2010		2000		1990	
	all	poor	all	poor	all	poor
White	0.00001	0.00178	0.00000	0.00144	0.00001	0.00157
Black	-0.00002	0.00089	0.00000	0.00066	-0.00001	0.00034
Hispanic	-0.00003	-0.00018	0.00003	0.00018	0.00002	0.00015
Asian	-0.00001	0.00262	-0.00004	0.00216	-0.00003	0.00351

Notes : Poor groups are defined as the bottom quartile of their income distribution. We use stable geographic units to perfom our comparaison over three census waves.

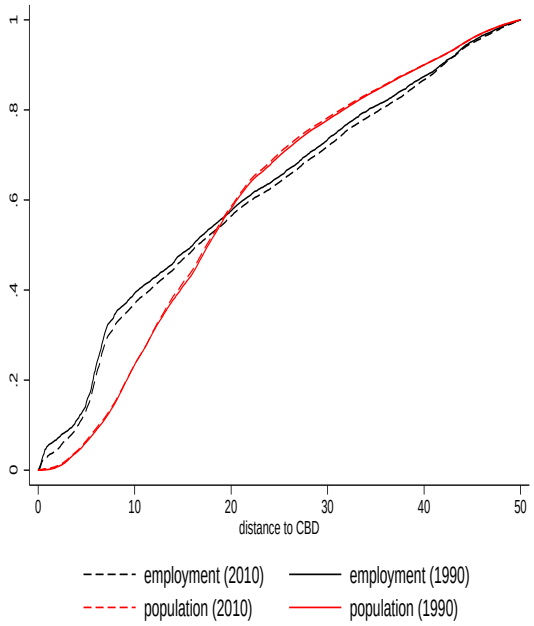
3.5 Appendix

TABLE 3.6: CDFs of 10 km ring.

	All		Poor	
		2010		
1.	Asian	0.26	Asian	0.28
2.	White	0.24	Black	0.24
3.	Black	0.23	White	0.21
4.	Hispanic	0.20	Hispanic	0.17
		1990		
1.	Asian	0.28	Asian	0.40
2.	Black	0.28	Black	0.33
3.	Hispanic	0.27	Hispanic	0.27
4.	White	0.20	White	0.22

FIGURE 3.10: Employment and population shift (1990-2010).

(a) Population and employment.



(b) Poor population and employment.

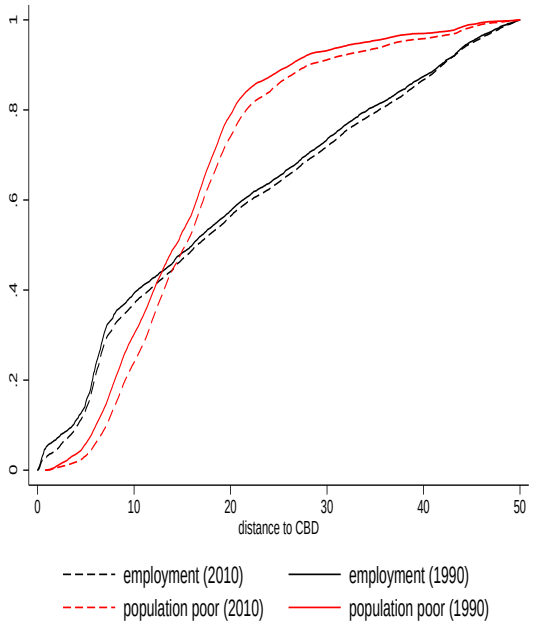
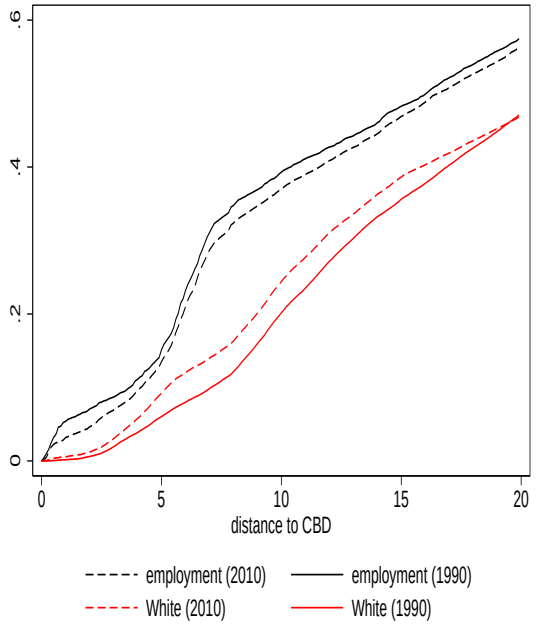


FIGURE 3.11: Employment and White shift (1990-2010).

(a) White.



(b) Poor White.

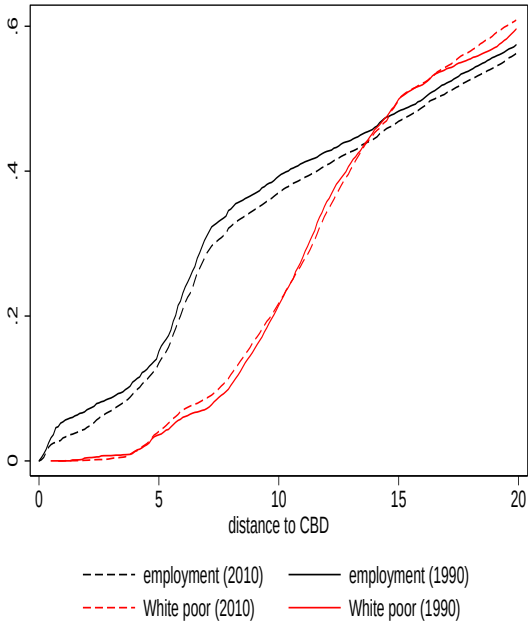
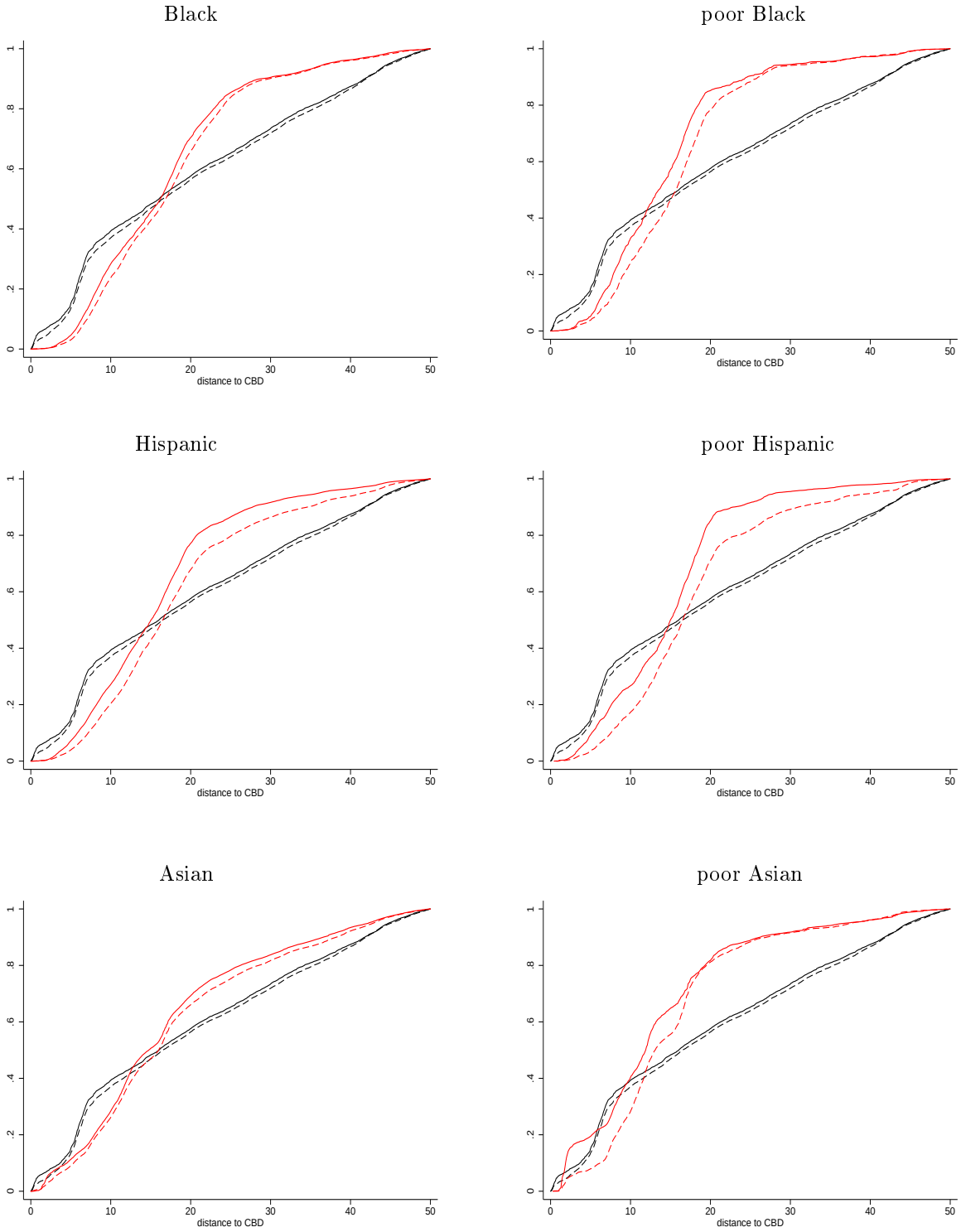


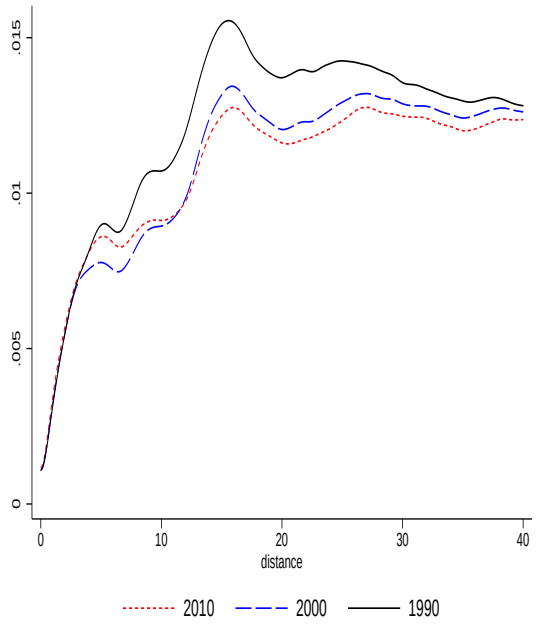
FIGURE 3.12: Employment and minorities shift (1990-2010).



Notes : Black lines are for employments and red for populations. Dashed are for 2010 and continued for 1990.

FIGURE 3.13: $\widehat{k}_{ij}^{xy}(d)$ for employment and White (1990 to 2010).

(a) White.



(b) Poor White.

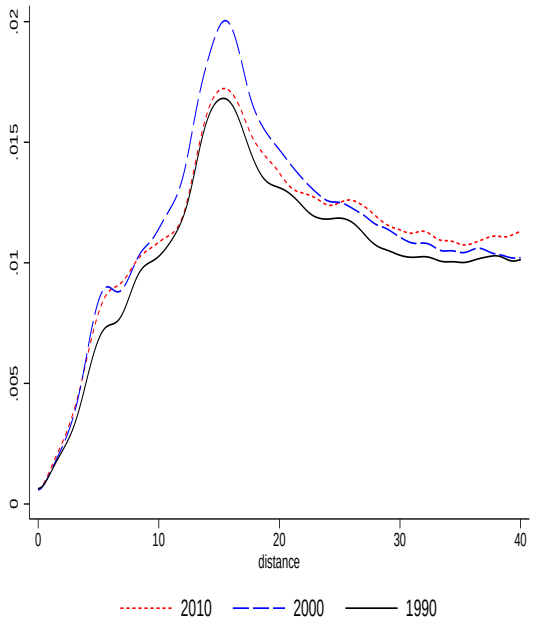


FIGURE 3.14: $\widehat{k}_{ij}^{xy}(d)$ for employment and minorities (1990 to 2010).

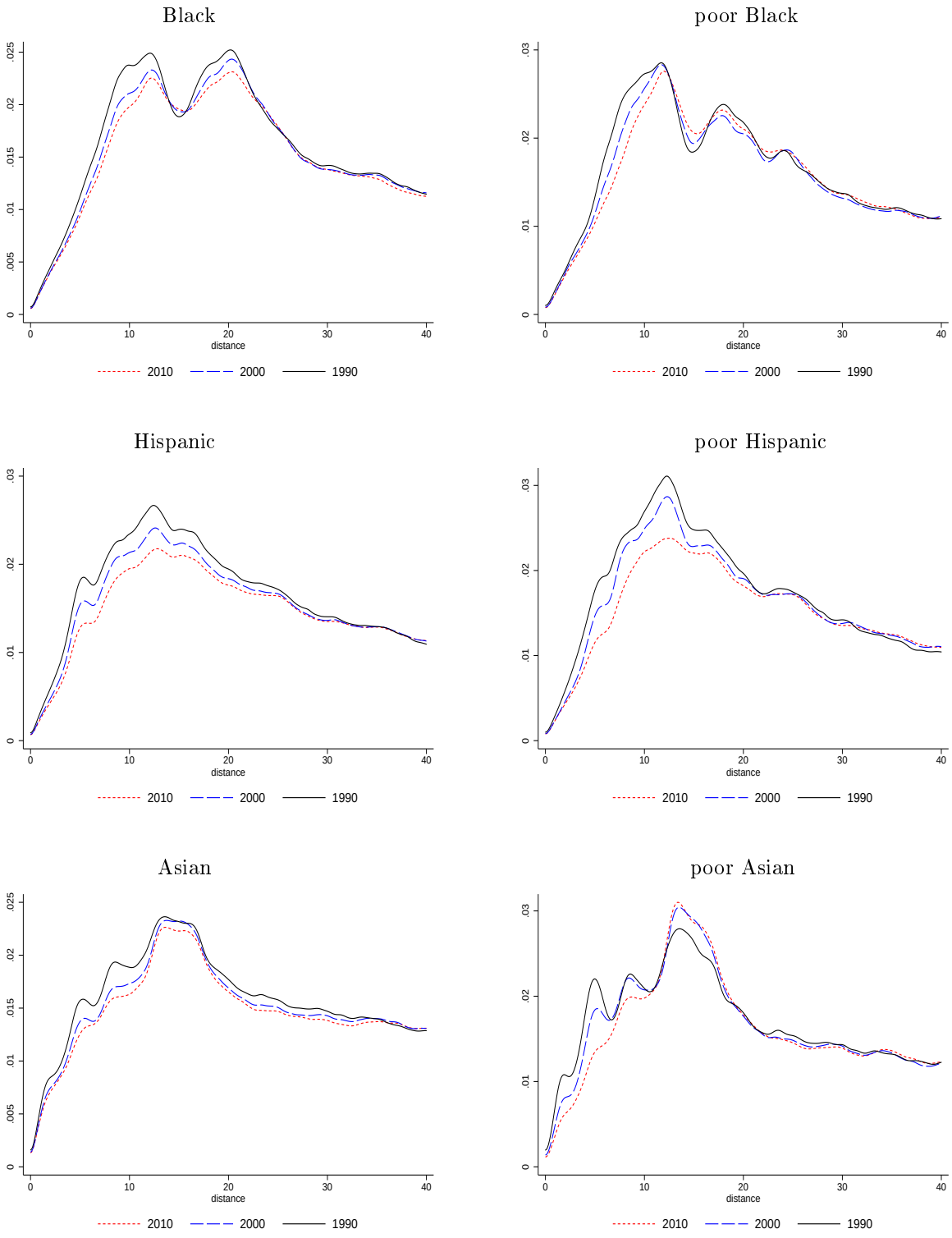
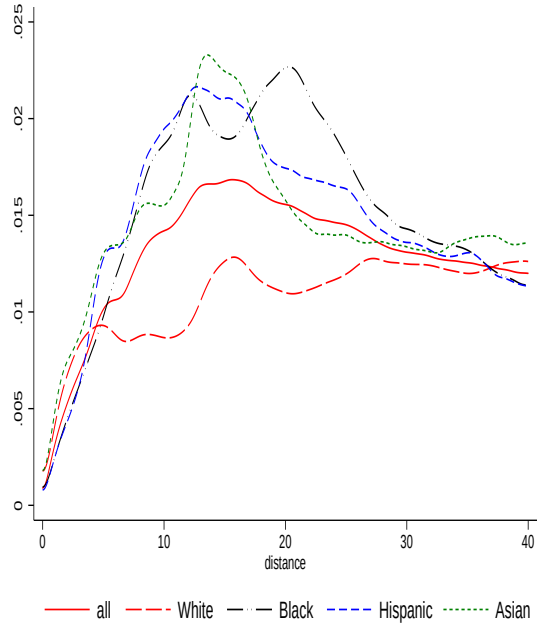


FIGURE 3.15: Top quartile opportunities and race (2010).

(a) $\widehat{k}_{ij}^{xy}(d)$.



(b) $\widehat{K}_{ij}^{xy}(d)$.

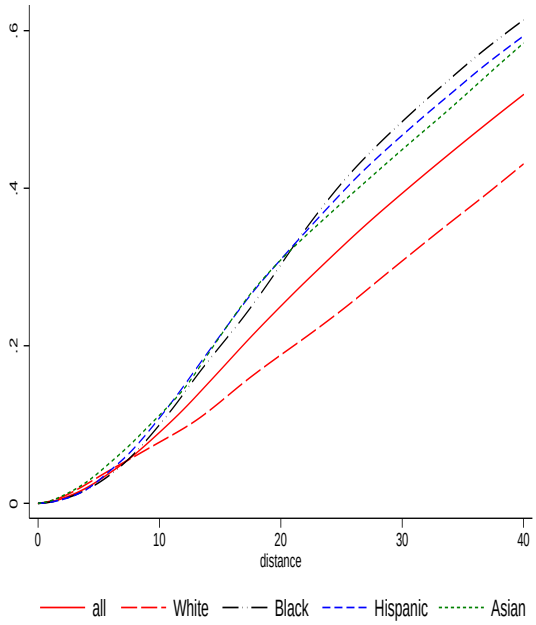
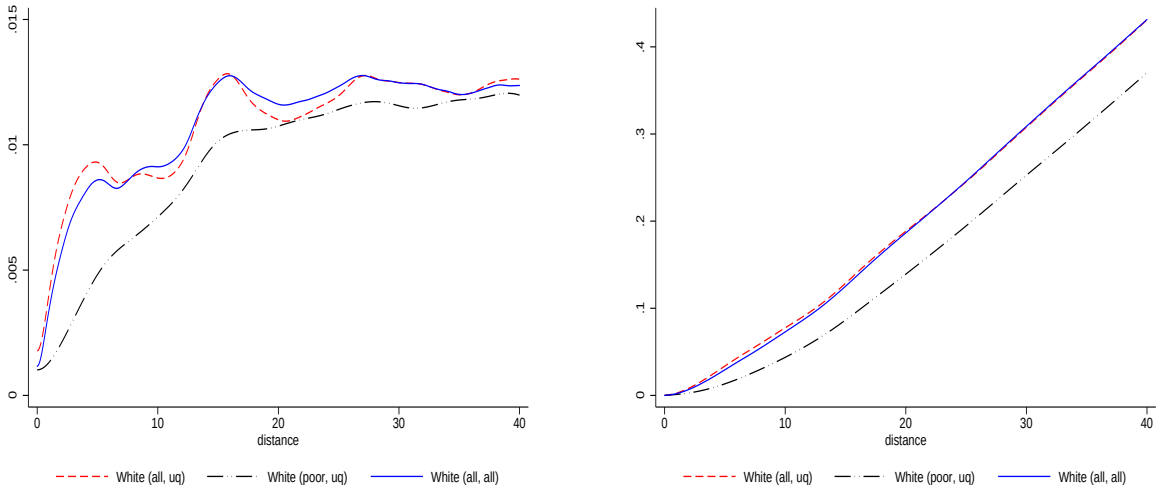


FIGURE 3.16: Top quartile opportunities, White and poverty (2010).

(a) $\widehat{k}_{ij}^{xy}(d)$. (b) $\widehat{K}_{ij}^{xy}(d)$.



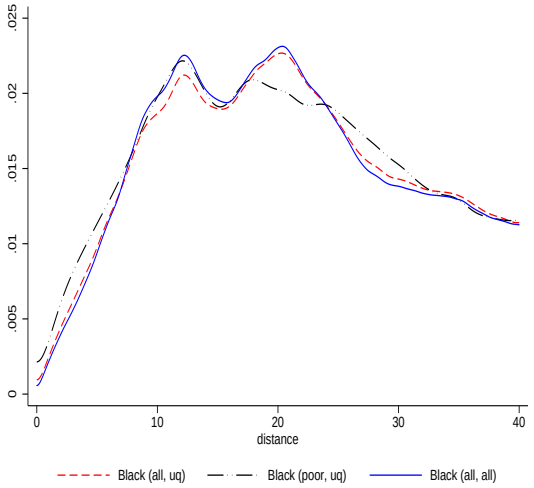
3.6 Conclusion

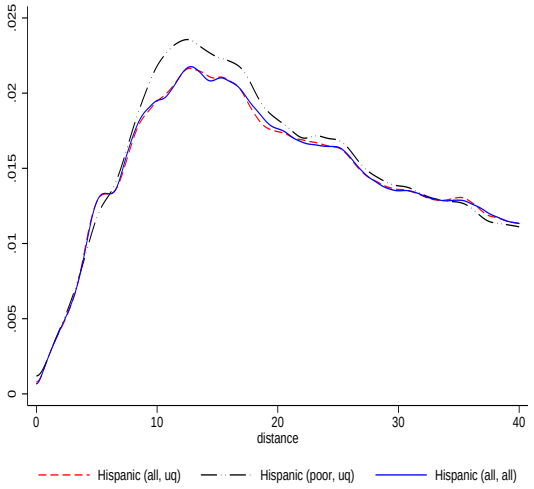
This paper investigates the spatial distribution of employment and population in the NYMSA. We first look at decentralization and then at the heterogeneity across different groups. To test the spatial mismatch hypothesis, we adopted a continuous measure of firm colocation that is rich and flexible . We found that race and poverty are two major drivers of the joint distribution of jobs and people, and that the poor and minorities tend to have different patterns than White and all population.

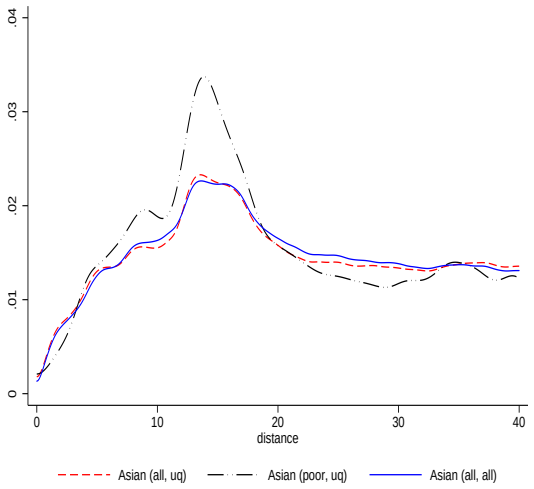
First, there is a shift of jobs towards population in general, but not for the poor who shifted away from jobs between 1990 and 2010. Second, race shapes also the patterns of decentralization. White shifted towards employment while Black, Hispanic and Asian moved away form jobs over the two decades, and this is again more substantial once we look at the poor within minority groups. As a second

FIGURE 3.17: Top quartile opportunities, Minorities and poverty (2010).

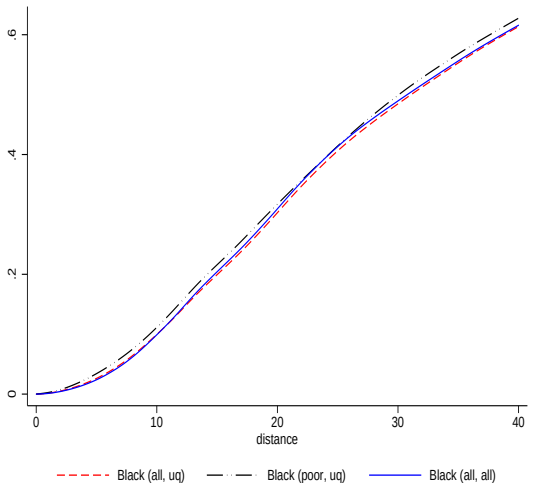
(a) $\widehat{k}_{ij}^{xy}(d)$.

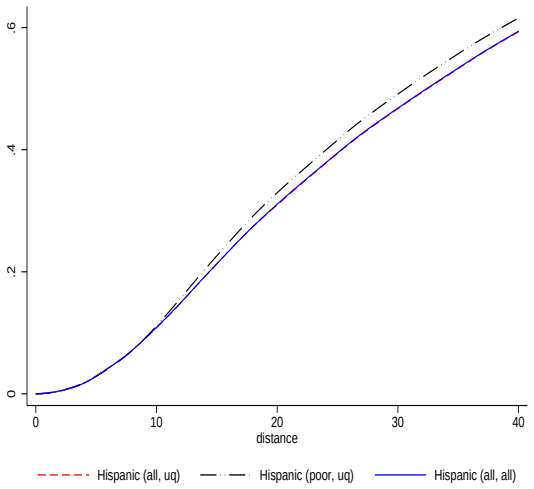


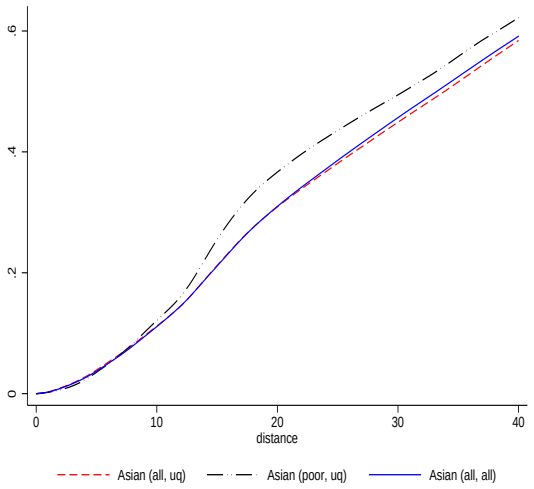




(a) $\widehat{K}_{ij}^{xy}(d)$.







step, and since a heterogeneity in the decentralization is potential ingredient of SMH, we applied a measure of colocation between groups and their potential employers, and found that spatial disconnection increased increased for White and decreased for minorities between 1990 and 2010. When we look at opportunities, White and Asian are closer to their opportunities than Black and Hispanic.

This paper attempts to make progress on the so called SMH, and highlights the importance of race and poverty. The suggestive evidence we provide might shed a light for policy makers to tackle the physical disconnection between jobs and people by targeting the right groups and jobs. There are three policies that can help to reduce spatial disconnection : Bring jobs to people, bring people to jobs, or connect jobs to people. Thus, knowing which jobs and which groups of population is crucial for an efficient public policy. This paper stresses mainly this point even if we did not provide any causal link of the SMH to the racial gap unemployment rates discussed in the introduction. While this question is technically challenging, and difficult in measurement, looking at the direct labor outcomes for poor and minorities is undoubtedly a fruitful area for future research.

CONCLUSION

Cette thèse démontre l'importance de la race et de la pauvreté dans l'analyse du choix de localisation et de co-localisation des individus.

Dans le premier chapitre, nous avons développé des mesures novatrices pour remédier aux problèmes auxquels font face les mesures classiques. Elles nous aident à voir de manière distincte l'importance de la race et de la pauvreté dans les patterns de ségrégation. Une application sur des données de la métropole de New York révèle que les Noirs, Hispaniques et Asiatique sont ségrégés de manière significative, et les revenus amplifient encore plus cette ségrégation.

Par ailleurs, dans le deuxième chapitre, en appliquant ces mesures dans le contexte canadien, nous avons trouvé aussi que l'origine ethnique est un facteur important dans le choix de co-localisation. En effet, nous avons prouvé que les ethnies qui sont similaires linguistiquement, culturellement, génétiquement, et qui partagent un passé politique et colonial ont tendance à être plus proches les unes des autres, avec des effets plus importants dans l'Est que l'Ouest canadien. Nous avons prouvé ainsi l'existence d'un corollaire de la Première Loi de la Géographie qui stipule que « les choses proches sont plus similaires que les choses distantes ».

Dans le dernier chapitre, nous avons analysé la distribution jointe firme-individu et vu comment le choix de localisation des individus et des firmes sont corrélés. Nous avons trouvé que les minorités sont loin de l'emploi en général, mais aussi de leurs opportunités, et la pauvreté a tendance à accentuer davantage cette déconnexion physique. Quand certains employeurs potentiels sont loin de certains

groupes d'individus, cela pourrait affecter ces derniers dans leur recherche d'emploi et leur niveau de revenus.

Cette thèse pourrait servir d'outil dans l'élaboration de politiques publiques ayant pour objectif la réduction de la ségrégation et l'accroissement de la diversité. Pouvoir mesurer et comprendre les origines des choix de localisation est crucial pour l'efficacité de ces politiques. En d'autres mots, comprendre les racines d'un problème est un premier pas pour régler ce problème. C'est ce que nous avons essayé de faire à travers cette thèse.

BIBLIOGRAPHIE

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. et Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth*, *8*(2), 55–194.

Angrist, J. et Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton : Princeton University Press.

Arbia, G., Benedetti, R. et Espa, G. (1996). Effects of the maup on image classification. *Geographical Systems*, *3*(2), 123–141.

Barlet, M., Briant, A. et Crusson, L. (2013). Location patterns of service industries in france : A distance-based approach. *Regional Science and Urban Economics*, *43*(2), 338 – 351. `http://dx.doi.org/https://doi.org/10.1016/j.regsciurbeco.2012.08.004`. Récupéré de `http://www.sciencedirect.com/science/article/pii/S0166046212000749`

Bayer, P., McMillan, R. et Rueben, K. S. (2004). What drives racial segregation ? new evidence using census microdata. *Journal of Urban Economics*, *56*(3), 514–535.

Behrens, K. (2016). Agglomeration and clusters : Tools and insights from coagglomeration patterns. *Canadian Journal of Economics*, *49*(4), 1293–1339.

Behrens, K., Boualam, B., Martin, J. et Mayneris, F. (2019). Gentrification and pioneer businesses. Document de travail 2019–02, Université du Québec à Montréal.

Behrens, K. et Moussouni, O. (2018). Distance-based segregation measures. Mimeographed, Université du Québec à Montréal.

Bell, W. (1954). A probability model for the measurement of ecological segregation. *Social Forces*, *32*(4), 357–364. Récupéré de `http://www.jstor.org/stable/2574118`

Billings, S. B. et Johnson, E. B. (2016). Agglomeration within an urban area. *Journal of Urban Economics*, *91*(C), 13–25. `http://dx.doi.org/10.1016/j.jue.2015.11.004`. Récupéré de `https://ideas.repec.org/a/eee/juecon/v91y2016icp13-25.html`

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *American Economic Review*, *77*(4), 531–553.

Boustan, L. (2013). Racial residential segregation in american cities. NBER Working Paper No. 19045.

Bridgman, B. (2008). What does the atlas narodov mira measure ? *Economics Bulletin*, *10*(6), 1–8.

Brown, C., Holman, E., Wichmann, S. et Velupillai, V. (2008). Automatic classification of the world's languages : A description of the method and preliminary results. *Language Typology and Universals*, *61*(4), 285–308.

Bruk, S. I. et Apenchenko, V. S. (1964). *Atlas Narodov Mira*. Moscow : Miklukho-Maklai Ethnological Institute, Department of Geodesy and Cartography of the State Geological Committee of the Soviet Union.

Burton, J., Nandi, A. et Platt, L. (2010). Measuring ethnicity : challenges and opportunities for survey research. *Ethnic and Racial Studies*, *33*(8), 1333–1349.

Carillo, P. E. et Rothbaum, J. L. (2016). Counterfactual spatial distributions. *Journal of Regional Science*, *56*(5), 868–894.

Carrillo, P. et Rothbaum, J. L. (2016). Counterfactual spatial distributions. *Journal of Regional Science*, *56*(5), 868–894. Récupéré de `https://EconPapers.repec.org/RePEc:bla:jregsc:v:56:y:2016:i:5:p:868-894`

Cassey, A. J. et Smith, B. O. (2014). Simulating confidence for the ellison–glaeser index. *Journal of Urban Economics*, *81*, 85–103.

Cavalli-Sforza, L. L., Menozzi, P. et Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton : Princeton University Press.

Combes, P.-P., Mayer, T. et Thisse, J.-F. (2008). *Economic Geography : The Integration of Regions and Nations*. Université Paris1 Panthéon-Sorbonne (Post-Print and Working Papers) hal-00311000, HAL

Coulson, N. E., Laing, D. et Wang, P. (2001). Spatial Mismatch in Search Equilibrium. *Journal of Labor Economics*, *19*(4), 949–972. Récupéré de `https://ideas.repec.org/a/ucp/jlabec/v19y2001i4p949-72.html`

Cutler, D. M., Glaeser, E. L. et Vigdor, J. L. (1999). The rise and decline of the american ghetto. *Journal of Political Economy*, *107*(3), 455–506.

Duranton, G. et Overman, H. G. (2005). Testing for localization using micro-geographic data. *Review of Economic Studies*, *72*(4), 1077–1106.

Duranton, G. et Overman, H. G. (2008). Exploring The Detailed Location Patterns Of U.K. Manufacturing Industries Using Microgeographic Data. *Journal of Regional Science*, *48*(1), 213–243.

Echenique, F. et Fryer, R. G. (2007). A measure of segregation based on social interactions. *The Quarterly Journal of Economics*, *122*(2), 441–485.

Ellison, G. D. et Glaeser, E. L. (1997). Geographic concentration in U.S. manufacturing industries : A dartboard approach. *Journal of Political Economy*, *105*(5), 889–927.

Ellison, G. D. et Glaeser, E. L. (1999). The geographic concentration of industry : Does natural advantage explain agglomeration ? *American Economic Review*, *89*(2), 311–316.

Ellison, G. D., Glaeser, E. L. et Kerr, W. R. (2010). What causes industry agglomeration ? Evidence from coagglomeration patterns. *The American Economic Review*, *100*(3), 1195–1213.

Faggio, G., Silva, O. et Strange, W. C. (2017). Heterogeneous agglomeration. *Review of Economics and Statistics*, *99*(1), 80–94.

Falck, O., Heblich, S., Lameli, A. et Suedekum, J. (2012). Dialects, cultural identity, and economic exchange. *Journal of Urban Economics*, *72*(2–3), 225–239.

Fearon, J. (2003). Ethnic and cultural diversity by country. *Journal of Economic Growth*, *8*, 195–222.

Fearon, J. et Laitin, D. (2003). Ethnicity, insurgency, and civil war. *American Political Science Review*, *97*(1), 75–90.

Feitosa, F. F., Câmara, G., Monteiro, A. M. V., Koschitzki, T. et Silva, M. P. S. (2007). Global and local spatial indices of urban segregation. *International Journal of Geographical Information Science*, *21*(3), 299–323. `http://dx.doi.org/10.1080/13658810600911903`. Récupéré de `https://doi.org/10.1080/13658810600911903`

Giuliano, G. et Small, K. A. (1991). Subcenters in the los angeles region. *Regional Science and Urban Economics*, *21*(2), 163 – 182. `http://dx.doi.`

org/https://doi.org/10.1016/0166-0462(91)90032-I. Récupéré de http://www.sciencedirect.com/science/article/pii/016604629190032I

Glaeser, E. L., Kahn, M. E., Arnott, R. et Mayer, C. (2001). Decentralized employment and the transformation of the american city [with comments]. *Brookings-Wharton Papers on Urban Affairs*, 1–63. Récupéré de http://www.jstor.org/stable/25058782

Glaeser, E. L. et Vigdor, J. (2012). The end of the segregated century : Racial separation in america's neighborhoods, 1890–2010. CIVIC Report #66.

Gobillon, L., Selod, H. et Zenou, Y. (2007). The Mechanisms of Spatial Mismatch. *Urban Studies*, *44*(12), 2401–2427. Récupéré de https://ideas.repec.org/a/sae/urbstu/v44y2007i12p2401-2427.html

Guiso, L., Sapienza, P. et Zingales, L. (2009). Cultural biases in economic exchange ? *The Quarterly Journal of Economics*, *124*(3), 1095–1131.

Head, K. et Mayer, T. (2014). In *Gravity Equations : Workhorse, Toolkit, and Cookbook.*, volume 4 de *Handbook of International Economics (G. Gopinath and E. Helpman and K. Rogoff, eds.)* chapitre 3, 131–195. Elsevier.

Head, K., Mayer, T. et Ries, J. (2011). The erosion of colonial trade linkages after independence. *Journal of International Economics*, *81*(1), 1–14.

Hidalgo, C. A. et Castañer, E. E. (2016). The amenity space and the evolution of neighborhoods. arXiv :1509.02868v2 [physics.soc-ph].

James, D. R. et Taeuber, K. E. (1985). Measures of segregation. *Sociological Methodology*, *15*, 1–32. Récupéré de http://www.jstor.org/stable/270845

Kain, J. (1968). Housing segregation, negro employment, and metropolitan decentralization. *The Quartely Journal of Economics*, *82*(2), 175–197.

Klier, T. et McMillen, D. (2008). Evolving agglomeration in the u.s. auto supplier industry. *Journal of Regional Science*, *48*(1), 245–267. Récupéré de https://EconPapers.repec.org/RePEc:bla:jregsc:v:48:y:2008:i:1:p:245-267

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Political Economy*, *99*(3), 483–99. Récupéré de https://EconPapers.repec.org/RePEc:ucp:jpolec:v:99:y:1991:i:3:p:483-99

Lazear, E. P. (1999). Culture and language. *Journal of Political Economy*, *107*(6), 95–126.

Lewis, M. P. (2009). Lewis, m. paul (ed.), "ethnologue : Languages of the world, sixteenth edition", dallas, texas : Sil international. (on line link : http ://www.ethnologue.com/16/web/).

Macauley, M. K. (1985). Estimation and recent behavior of urban population and employment density gradients. *Journal of Urban Economics*, *18*(2), 251–260. Récupéré de https://ideas.repec.org/a/eee/juecon/v18y1985i2p251-260.html

Marcon, E. et Puech, F. (2017). A typology of distance-based measures of spatial concentration. *Regional Science and Urban Economics*, *62*, 56 – 67. http://dx.doi.org/https://doi.org/10.1016/j.regsciurbeco.2016.10.004. Récupéré de http://www.sciencedirect.com/science/article/pii/S0166046216302782

Martin, P., Mayer, T. et Thoening, M. (2008). Make trade, not war ? *Review of Economic Studies*, *75*(3), 865–900.

Martin, R. W. (2004). Can black workers escape spatial mismatch ? employment shifts, population shifts, and black unemployment in american cities. *Journal of Urban Economics*, *55*(1), 179 – 194. http://dx.doi.org/

https://doi.org/10.1016/j.jue.2003.09.003. Récupéré de http://www.sciencedirect.com/science/article/pii/S0094119003001165

Massey, D. S. et Denton, N. A. (1988). The dimensions of residential segregation. *Social Forces*, *67*(2), 281–315. Récupéré de http://www.jstor.org/stable/2579183

Maurel, F. et Sedillot, B. (1999). A measure of the geographic concentration in french manufacturing industries. *Regional Science and Urban Economics*, *29*(5), 575–604. Récupéré de https://EconPapers.repec.org/RePEc:eee:regeco:v:29:y:1999:i:5:p:575-604

McMillen, D. P. et McDonald, J. F. (1998). Suburban Subcenters and Employment Density in Metropolitan Chicago. *Journal of Urban Economics*, *43*(2), 157–180. Récupéré de https://ideas.repec.org/a/eee/juecon/v43y1998i2p157-180.html

McPherson, M., Smith-Lovin, L. et Cook, J. M. (2001). Birds of a feather : Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

Mecham, R. Q., Fearon, J. et Laitin, D. (2006). Religious classifcation and data on shares of major world religions. Mimeographed, Stanford University.

Mele, A. (2013). Poisson indices of segregation. *Regional Science and Urban Economics*, *43*(1), 65–85.

Melitz, J. et Toubal, F. (2014). Native language, spoken language, translation and trade. *Journal of International Economics*, *93*(2), 351–363.

Mieszkowski, P. et Mills, E. S. (1993). The causes of metropolitan suburbanization. *Journal of Economic Perspectives*, *7*(3), 135–147. http://dx.doi.org/10.1257/jep.7.3.135. Récupéré de http://www.aeaweb.org/articles?id=10.1257/jep.7.3.135

Mieszkowski, P. et Smith, B. (1991). Analyzing urban decentralization : The case of houston. *Regional Science and Urban Economics*, *21*(2), 183 – 199. http://dx.doi.org/https://doi.org/10.1016/0166-0462(91)90033-J. Récupéré de http://www.sciencedirect.com/science/article/pii/016604629190033J

Morgan, B. S. (1975). The segregation of socio-economic groups in urban areas : a comparative analysis. *Urban Studies*, *12*(1), 47–60. http://dx.doi.org/10.1080/00420987520080041. Récupéré de https://doi.org/10.1080/00420987520080041

Mori, T. et Smith, T. E. (2015). On the spatial scale of industrial agglomerations. *Journal of Urban Economics*, *89*, 1 – 20. http://dx.doi.org/https://doi.org/10.1016/j.jue.2015.01.006. Récupéré de http://www.sciencedirect.com/science/article/pii/S0094119015000078

Nechyba, T. J. (2006). In *Income and peer quality sorting in public and private schools*, volume 2 de *Handbook of the Economics of Education (E. Hanushek and F. Welch, eds.)* chapitre 22, 1327–1368. Elsevier.

Patacchini, E. et Zenou, Y. (2006). Search activities, cost of living and local labor markets. *Regional Science and Urban Economics*, *36*(2), 227–248. Récupéré de https://ideas.repec.org/a/eee/regeco/v36y2006i2p227-248.html

Pemberton, T. J., DeGiorgio, M. et Rosenberg, N. A. (2013). Population structure in a comprehensive genomic data set on human microsatellite variation. *G3-Genes/Genomes/Genetics*, *3*, 903–919.

Piketty, T. et Goldhammer, A. (2014). *Capital in the Twenty-First Century*. Harvard University Press. Récupéré de http://www.jstor.org/stable/j.ctt6wpqbc

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W. et Cavalli-Sforza, L. L. (2005). Support from the relationship between genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences*, *102*(44), 15942–15947.

Reardon, S. F. et Firebaugh, G. (2002). 2. measures of multigroup segregation. *Sociological Methodology*, *32*(1), 33–67. http://dx.doi.org/10.1111/1467-9531.00110. Récupéré de https://doi.org/10.1111/1467-9531.00110

Reardon, S. F. et O'Sullivan, D. (2004). Measures of spatial segregation. *Sociological Methodology*, *34*, 121–162. Récupéré de http://www.jstor.org/stable/3649372

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J. et Sobek, M. (2016). Ipums usa : Version 8.0 [dataset]. minneapolis, mn : Ipums, 2018. https ://doi.org/10.18128/d010.v8.0. Récupéré de https://doi.org/10.18128/D020.V7.1

Schelling, T. C. (1969). Models of segregation. *American Economic Review*, *59*(2), 488–493.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, *1*((2)), 143–186.

Scholl, T. et Brenner, T. (2015). Optimizing distance-based methods for large data sets. *Journal of Geographical Systems*, *17*(4), 333–351. Récupéré de https://EconPapers.repec.org/RePEc:kap:jgeosy:v:17:y:2015:i:4:p:333-351

Spolaore, E. et Wacziarg, R. (2009). The diffusion of development. *The Quarterly Journal of Economics*, *124*(2), 469–529.

Spolaore, E. et Wacziarg, R. (2016). Ancestry, language and culture. In *The Palgrave Handbook of Economics and Language* 174–211. Springer.

Spolaore, E. et Wacziarg, R. (2018). Ancestry and development : New evidence. Mimeographed, Tufts University.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Annual Review of Sociology*, *46*((Supplement)), 234–240.

Wasmer, E. et Zenou, Y. (2006). Equilibrium search unemployment with explicit spatial frictions. *Labour Economics*, *13*(2), 143–165. Récupéré de `https://ideas.repec.org/a/eee/labeco/v13y2006i2p143-165.html`

WCD (2007). World christian database. http ://www.worldchristiandatabase.org/wcd/.

Weidmann, N. B., Rød, J. K. et Cederman, L.-E. (2010). Representing ethnic groups in space : A New Dataset. *Journal of Peace Research*, *47*(4), 491–499.