Behrens, Kristian; Moussouni, Oualid

Working Paper

What Matters for Choosing your Neighbors: Evidence from Canadian Metropolitan Areas

Document de travail, No. 2019-03

Provided in Cooperation with:
Department of Economics, School of Management Sciences (ESG UQAM), University of Quebec in Montreal

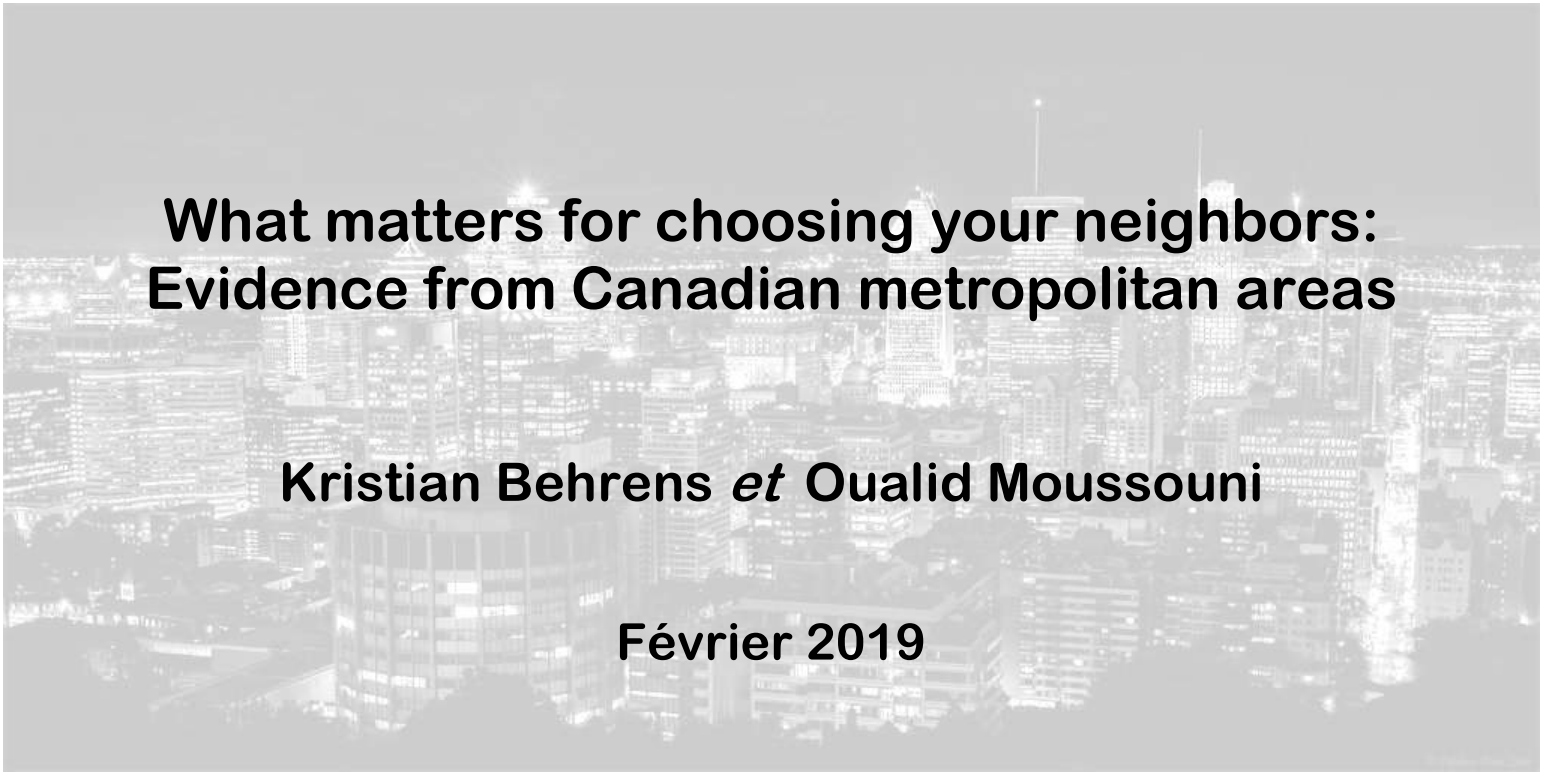
Suggested Citation: Behrens, Kristian; Moussouni, Oualid (2019) : What Matters for Choosing your Neighbors: Evidence from Canadian Metropolitan Areas, Document de travail, No. 2019-03, Université du Québec à Montréal, École des sciences de la gestion (ESG UQAM), Département des sciences économiques, Montréal

This Version is available at:
https://hdl.handle.net/10419/234788

ESG RECHERCHE

DOCUMENT DE TRAVAIL / WORKING PAPER

No. 2019-03



What matters for choosing your neighbors: Evidence from Canadian metropolitan areas

Kristian Behrens *et* Oualid Moussouni

Février 2019

ESG UQÀM

Département des sciences économiques
École des sciences de la gestion
Université du Québec à Montréal

What matters for choosing your neighbors: Evidence from Canadian metropolitan areas

Kristian Behrens, Université du Québec à Montréal, Canada; National Research University Higher School of Economics, Russie; et CEPR
Oualid Moussouni, Université du Québec à Montréal, Canada.

Document de travail No. 2019-03

Février 2019

Département des Sciences Économiques
Université du Québec à Montréal
Case postale 8888,
Succ. Centre-Ville
Montréal, (Québec), H3C 3P8, Canada
Courriel : brisson.lorraine@uqam.ca
Site web : http://economie.esg.uqam.ca

What matters for choosing your neighbors? Evidence from Canadian metropolitan areas*

Kristian Behrens† Oualid Moussouni‡

February 15, 2019

Click here for the latest version

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.

*We thank our discussants, Jordi Jofre-Monseny and Carlez Méndez-Ortega, as well as Mark Brown, Julien Martin, Florian Mayneris, Giordano Mion, Yasusada Murata, Vincent Rebeyrol, and seminar and workshop participants at the Journées de la Société Canadienne de Sciences Économiques, the 2018 UEA Meetings in New York, the National Housing Conference in Ottawa, the winter workshop of CMSSE in Saint Petersburg, and the 2018 NARSC Meetings in San Antonio for valuable comments and suggestions. Behrens and Moussouni gratefully acknowledge financial support from the CRC Program of the Social Sciences and Humanities Research Council (SSHRC) of Canada for the funding of the *Canada Research Chair in Regional Impacts of Globalization*; and from SSHRC's Insight Grants Program ('Cities in Motion'; grant #435-2016-1246). This article was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'. Any remaining errors are ours.

†Université du Québec à Montréal, Canada; National Research University Higher School of Economics, Russian Federation; and CEPR, UK. E-mail: behrens.kristian@uqam.ca

‡Université du Québec à Montréal, Canada. E-mail: moussouni.oualid@courrier.uqam.ca

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 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.

[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."

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 and 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]

Our measures of similarity—linguistic, religious, cultural, and genetic—and of historico-

[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 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.

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 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

[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.

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 lays out the methodology, and explains our data and measurements. Section 3 explains the empirical strategy and discusses identification concerns. Section 4 presents our results. It also contains many extensions and a battery of robustness checks. Last, Section 5 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 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.

2.1 Geographic proximity between groups

2.1.1 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 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 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."

[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, 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.

2.1.2 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.

2.1.3 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 and 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:

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

[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 (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:

$$\widehat{K}_c^{ij}(d) = \int_0^d \widehat{k}_c^{ij}(\zeta)\mathrm{d}\zeta. \tag{2}$$

The measure (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 (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 and 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 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

[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.

[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.

city-ethnic pair combination). We will provide robustness checks using an alternative measure of colocation—the Ellison-Glaeser index (Ellison and Glaeser 1997)—later in the paper.

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 12 there provides the full list of our variables.

2.2.1 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 and Toubal (2014) and Spolaore and 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 and 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 and 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 and 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 and 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 and 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.

2.2.2 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 and 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.

2.2.3 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 and 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.

2.2.4 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).

2.2.5 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.

3 Empirical strategy

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

3.1 Estimating equation

Our basic specification is the following linear model:

$$\widehat{K}_c^{ij}(\overline{d}) = \alpha + X^{ij}\beta + \delta_c^i + \delta_c^j + \varepsilon_c^{ij}, \tag{3}$$

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 concentration). 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 (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 and 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] 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 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

[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.

[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 and Pischke 2009).

may partially capture the same underlying characteristics.

3.2 Identification concerns

Our explanatory variables X^{ij}, described in Section 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 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

[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.

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 red-lining'). 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

$$\widehat{K}_c^{ij}(\overline{d}) = \alpha + X^{ij}\beta + \delta_c^i + \delta_c^j + \left[\widetilde{K}_c^{ij}(\overline{d}) + \varepsilon_c^{ij}\right], \tag{4}$$

a classic case of omitted variables.[16]

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,

[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 and 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 and Glaeser 1999, Klier and McMillen 2008, Carillo and Rothbaum 2016, and Behrens and 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 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 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 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) are again more likely to be informative of the true effects we are looking for.[18]

[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.

[18] Alternative potentially informative sample splits would be in terms of housing consumption (apartments vs

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.

4 Results

Tables 1 and 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 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

detached or semi-detached units), or in terms of occupations and jobs. Unfortunately, we do not have those data for our small geographic units.

degrees of colocation if there are strong patterns in where housing is available. We will control for that aspect of simultaneity in arrival later.

Table 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).

Which pairs are the most coagglomerated in Canadian cities? Table 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 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. 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: 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.

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 3.

Table 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 3, has a strong negative effect on colocation patterns: ethnic groups that are genetically more distant tend to colocate less.[19]

Table 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.

We next include all variables jointly into our baseline specification. Table 4 summarizes our results, where we progressively add the economic, historico-political, cultural, and genetic variables to our basic geographic variables. As Table 4 shows, the coefficients on the geographic variables progressively decrease as we add our economic, linguistic, historic, religious, and genetic variables. As expected, 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 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 14 and 15 in Appendix C show the same results as Tables 3 and 4 for the 2006 Census. Our qualitative results are stable across censuses.

[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 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 groups)					-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.

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.

4.2.1 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 5 summarizes results for our baseline specification (5) from Table 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 Head et al. 2010): 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.

4.2.2 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 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 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.

Table 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.

(2010), given by:

$$EG_c^{ij} = \frac{\sum_m (s_{m,c}^i - x_{m,c})(s_{m,c}^j - x_{m,c})}{1 - \sum_m (x_{m,c})^2}, \tag{5}$$

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 (5) 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 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 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.

4.2.3 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 6 summarizes our

[20] Figure 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).

[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.

results. As shown, they change little compared to the baseline results in Table 4. Actually, the results in column (5) of Table 6 are almost identical to the correponding results in column (5) of Tables 4. We further show in Table 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 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 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 6.

4.2.4 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 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 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.

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

[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 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.

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).

4.3.1 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 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 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

Table 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.

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.

4.3.2 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 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 17 in Appendix C shows our results, along the same lines as Table 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.

4.3.3 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 8 and Figure 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 and genetic distance,

[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.

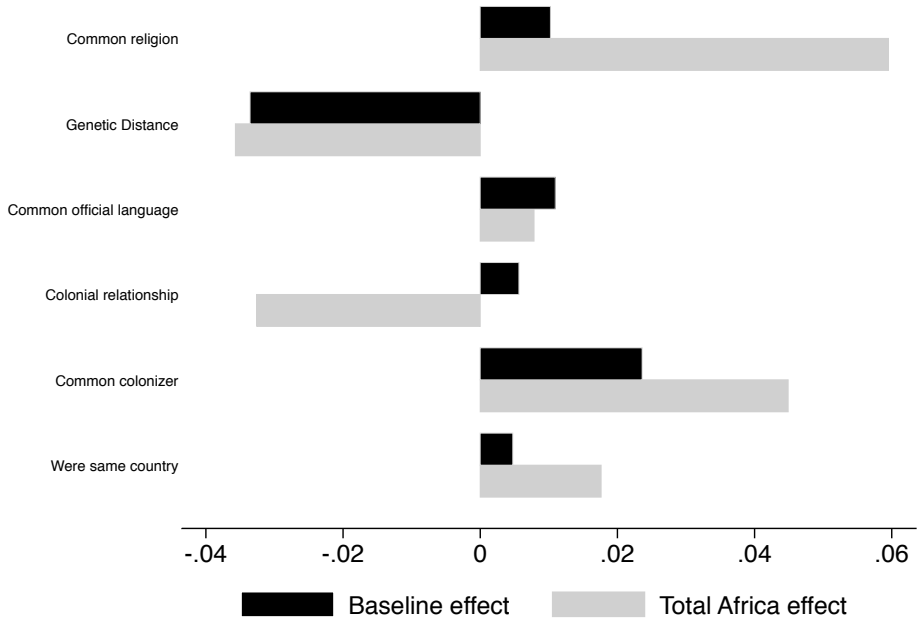
[24] We also estimated the specification for pairs ij that originate both from Asia. The results are similar, except that religion matters less whereas language matters more for these couples. In any case, our results suggest that the effects do not vanish when looking at these groups.

Table 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 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).

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.

4.4 Extensions: Heterogeneity by city and mean reversion

4.4.1 Heterogeneity across cities

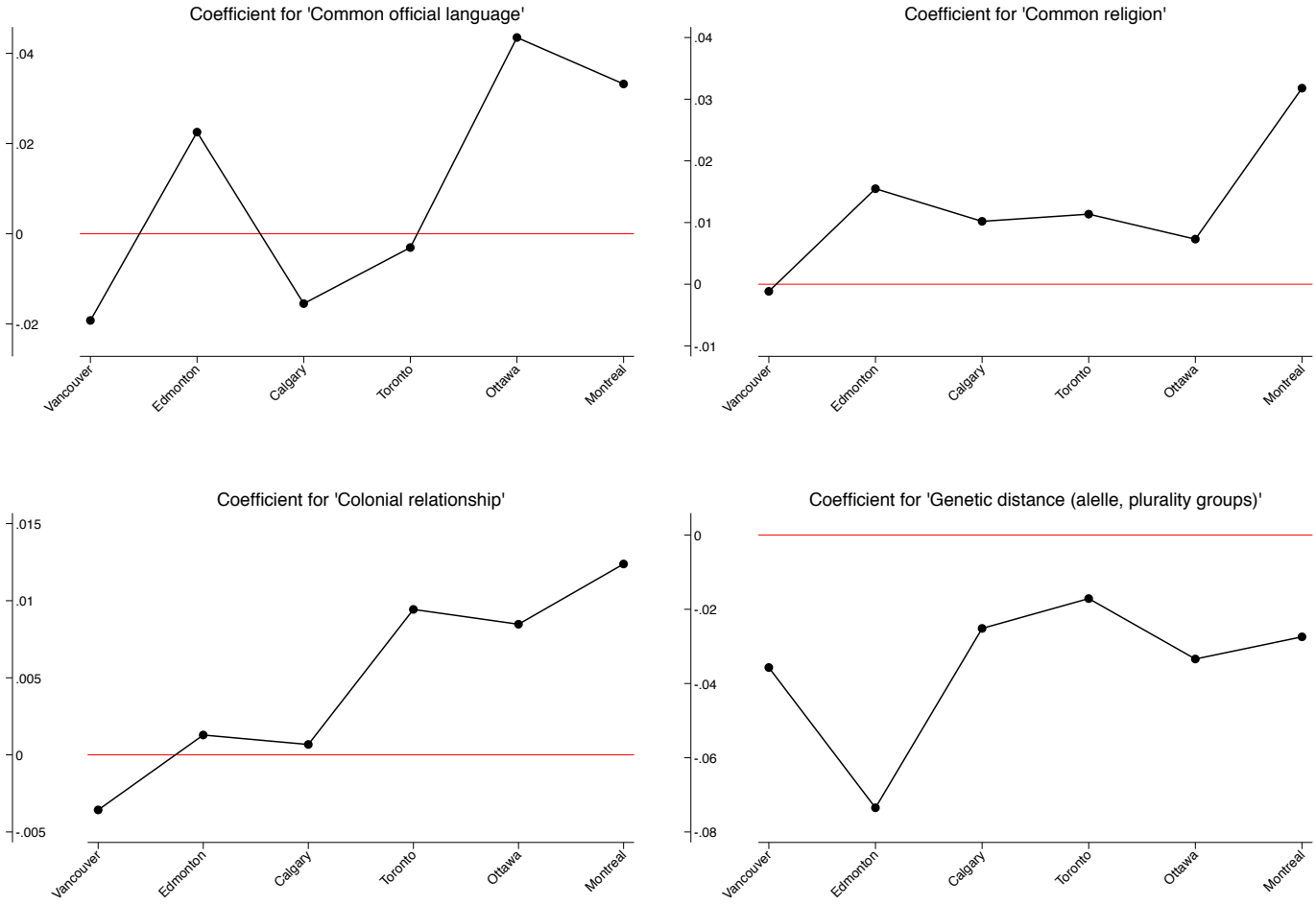
Until now, we have considered common coefficients across all cities. Yet, there may be heterogeneous effects across cities. First, historic immigration patterns differ across cities in Canada. Thus, language may be more important in some cities whereas religion may be more important in others. Second, 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 (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 9 and Figure 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 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

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



Notes: See Table 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.

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).

4.4.2 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

Table 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.

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 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 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 and 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.

5 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 popula-

Table 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.

tions 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.

References

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

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

Arbia, G., R. Benedetti, and G. Espa, "Effects of the MAUP on Image Classification," *Geographical Systems*, 1996, *3* (2), 123–141.

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

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

Behrens, K. and O. Moussouni, "Distance-based segregation measures," Mimeographed, Université du Québec à Montréal 2018.

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

Boustan, L., "Racial Residential Segregation in American Cities," NBER Working Paper No. 19045 2013.

Bridgman, B., "What does the Atlas Narodov Mira measure?," *Economics Bulletin*, 2008, *10* (6), 1–8.

Brown, C., E. Holman, S. Wichmann, and V. Velupillai, "Automatic Classification of the World's Languages: A Description of the Method and Preliminary Results," *Language Typology and Universals*, 2008, *61* (4), 285–308.

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

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

Carillo, P. E. and J. L. Rothbaum, "Counterfactual Spatial Distributions," *Journal of Regional Science*, 2016, *56* (5), 868–894.

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

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

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

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

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

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

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

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

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

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

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

Glaeser, E. L. and J. Vigdor, "The end of the Segregated Century: Racial separation in America's neighborhoods, 1890–2010," CIVIC Report #66 2012.

Guiso, L., P. Sapienza, and L. Zingales, "Cultural Biases in Economic Exchange?," *The Quarterly Journal of Economics*, 2009, *124* (3), 1095–1131.

Head, K. and T. Mayer, in "Gravity Equations: Workhorse, Toolkit, and Cookbook.," Vol. 4 of *Handbook of International Economics (G. Gopinath and E. Helpman and K. Rogoff, eds.)*, Elsevier, 2014, chapter 3, pp. 131–195.

Head, K., T. Mayer, and J. Ries, "The erosion of colonial trade linkages after independence," *Journal of International Economics*, 2010, *81* (1), 1–14.

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

Hidalgo, C. A. and E. E. Castañer, "The Amenity Space and The Evolution of Neighborhoods," arXiv:1509.02868v2 [physics.soc-ph] 2016.

Klier, T. and D. P. McMillen, "Evolving agglomeration in the U.S. auto supplier industry," *Journal of Regional Science*, 2008, *48* (1), 245–267.

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

Lewis, M P, "Lewis, M. Paul (ed.), "Ethnologue: Languages of the World, Sixteenth edition", Dallas, Texas: SIL International. (On line link : http://www.ethnologue.com/16/web/)," 2009.

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

McPherson, M., L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 2001, *27*, 415–444.

Mecham, R. Q., J. Fearon, and D. Laitin, "Religious Classifcation and Data on Shares of Major World Religions," Mimeographed, Stanford University 2006.

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

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

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

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza, "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*, 2005, *102* (44), 15942–15947.

Schelling, T. C., "Models of Segregation," *American Economic Review*, 1969, *59* (2), 488–493.

Schelling, T. C., "Dynamic Models of Segregation," *Journal of Mathematical Sociology*, 1971, *1* (2), 143–186.

Spolaore, E. and R. Wacziarg, "The diffusion of development," *The Quarterly Journal of Economics*, 2009, *124* (2), 469–529.

Spolaore, E. and R. Wacziarg, "Ancestry, language and culture," in "The Palgrave Handbook of Economics and Language," Springer, 2016, pp. 174–211.

Spolaore, E. and R. Wacziarg, "Ancestry and Development: New Evidence," Mimeographed, Tufts University 2018.

Tobler, W., "A computer movie simulating urban growth in the Detroit region," *Annual Review of Sociology*, 1970, *46* ((Supplement)), 234–240.

WCD, "World Christian Database," http://www.worldchristiandatabase.org/wcd/ 2007.

Weidmann, N. B., J. K. Rød, and L.-E. Cederman, "Representing ethnic groups in space: A New Dataset," *Journal of Peace Research*, 2010, *47* (4), 491–499.

Appendix material

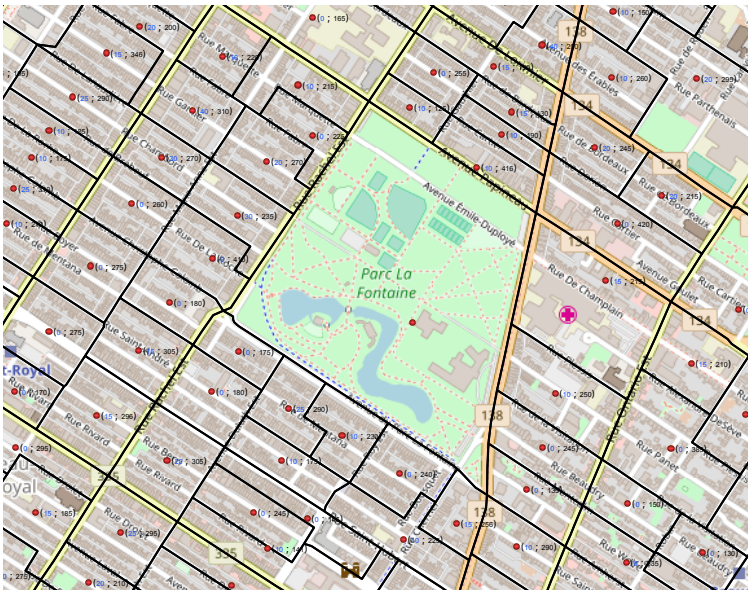
This set of appendices is structured as follows. Appendix **A** presents additional details and information on our data. Appendix **B** explains in more detail our procedure for mapping ethnic groups to countries. Last, Appendix **C** contains additional tables and results.

Appendix A. Addititional information on the data

A.1. Census data

Figure 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 11 reports summary statistics by city, including population figures and the number of dissemination areas.

Figure 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. Fig-

ure 5 and Table 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.

Table 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.

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 households, with a lower response rate (68.6%, or around 7 million individual responses). The

estimated data, if any, 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 and Mayer 2014). Table 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 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 from Melitz and 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

Table 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.

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

[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.

Table 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	

$\texttt{lang\_csl}_{ij}$ 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 and 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 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.

[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.

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

Linguistic distances (`lang_lingdist_dom_formula`**,** `lang_lingdist_weighted_formula`**,** `lang_cognate_dominant` **and** `lang_cognate_weighted`**).** Our source of data is Spolaore and Wacziarg (2009, 2016). The first measure of linguistic distance 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 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 "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 20 in the supplemental online appendix provides more detailed correlations within the language measures.

[30] For instance, Spolaore and 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` $= \sum_{i=1}^{I} \sum_{j=1}^{J} \left(S_{1i} \times S_{2j} \times c_{ij} \right)$, 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.

A.2.2. Measures of religious and cultural distance.

Common religion (`cult_comrelig`). This measure comes from Melitz and 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 and 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 and 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 and 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 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.

[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.

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 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

[33] See www.cepii.fr/anglaisgraph/bdd/distances.htm

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 and Apenchenko, 1964). It delineates the territories of ethnic groups associated with more than 8,900 polygons worldwide.[34] To understand how the procedure works, consider Figure 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 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 18 in the online appendix for a detailed breakdown of the mapping from ethnic groups to countries).

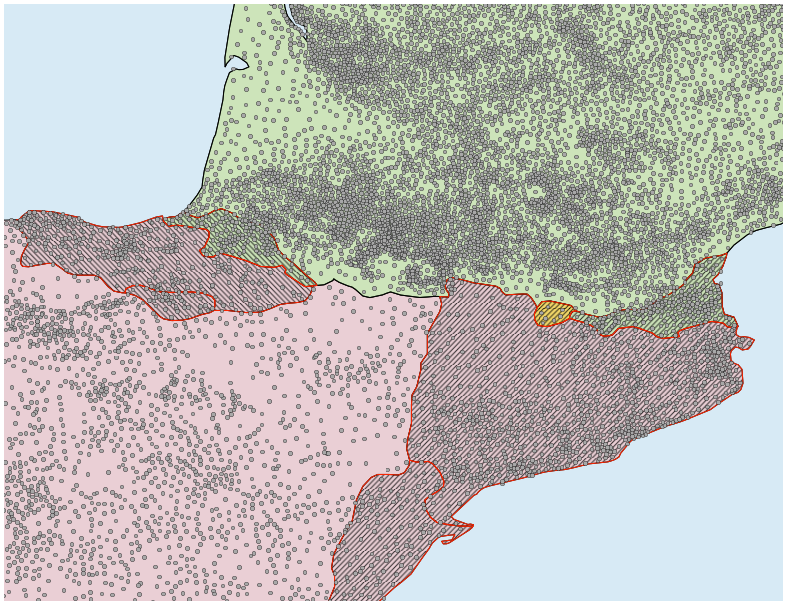
Formally, let θ_i^c denote the share of ethnic group i in country c, with $\sum_c \theta_i^c = 1$. We use these shares to split out ethnic groups in the different DAs among countries. For example, a dissemination area in city c that reports 100 residents of Flemish ethnicity will be split into

[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.

Figure 4: Mapping ethnic groups—for example, Basques and Catalans—to countries.



$100 \times \theta^{BEL}_{Flemish} = 96$ people from Belgium and $100 \times \theta^{FRA}_{Flemish} = 4$ people from France, using the shares summarized in Table 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 'split'.

Appendix C: Additional results

This appendix contains additional tables and results.

2006 Census. Tables 14 and 15 show the same results as Tables 3 and 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

[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.

of ethnic groups, which explains why we have smaller sample sizes and why the results are generally less precise.

Table 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.

Robustness to large enough ethnic groups. Table 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 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 17 depicts our results where we estimate the K-densities for the rich DAs and for the 'owner' DAs as defined in the main text.

Table 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 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 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 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

[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 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.

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]

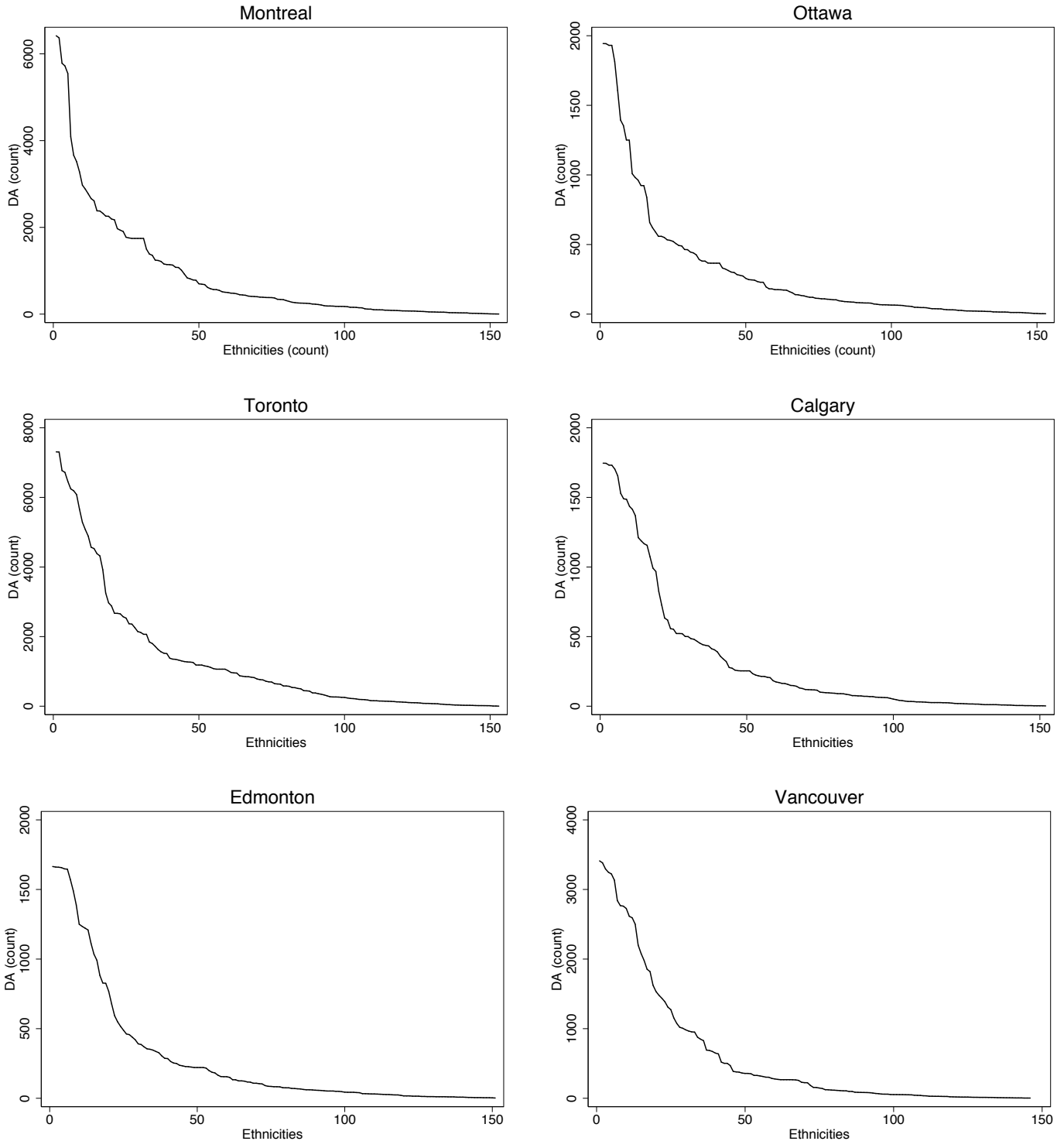
To summarize, there are two types of potential section biases: into migration, and across cities. While we cannot control the former, we think that presenting results that include all pairs of ethnic groups into the analysis will help to mitigate the latter.

Appendix S.2. Additional tables and results

- Table 18 shows the mapping of ethnic groups to countries, including the different population shares.
- Figure 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 19 summarizes results for the 2016 and 2006 censuses for coagglomeration patterns measured at 100m and 1km distance thresholds, respectively.
- Table 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 21, provides a detailed breakdown of the largest and smallest ethnic groups by cities.

[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.

Figure 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 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 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 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 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.'