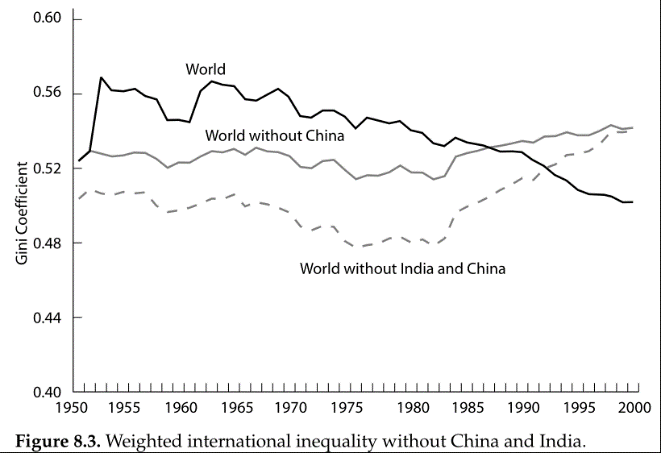
Global Inequality, Linguistic Distance, and Diffusion

Kyle Chan

**Abstract**

What explains patterns of economic development across countries? Research on supranational factors such as world-polity theory, colonial legacy, and the dynamics of inter-state competition have shed light on the sources of clustering patterns of present-day levels of development across countries. This study uses the most comprehensive linguistic dataset to date and introduces new methodological tools to sociology from population genetics to explore the relationship between inter-country linguistic distance and patterns of economic growth in the latter half of the twentieth century. After controlling for geography, income levels, and trade, linguistic distance is significantly associated with similarities in economic growth patterns.

Global income inequality is massive and persistent. While global income inequality appears to have declined in recent decades, this is primarily due to the economic rise of China (Firebaugh and Goesling 2004; Milanovic 2005). Excluding China from the analysis, the global population-weighted Gini coefficient remains above 0.52 and has in fact grown in recent decades (Milanovic 2005). Within-country inequality has recently become more salient in national politics and in social science research, driven in no small part by the work of Thomas Piketty (2013). However, between-country inequality remains the predominant driver behind global inequality today (Bourguignon and Morrisson 2002; Ravallion 2014; World Bank 2005) and deserves further investigation alongside within-country inequality.



**Figure 1**. Population-weighted Gini coefficients over time with and without China and India. Directly from Milanovic (2005).

What explains global income inequality across countries? Many conventional approaches to studying development emphasize the characteristics of individual countries. Most prominently, the Solow growth model, named after its inventor Robert Solow (1956) and known within economics as the neoclassical growth model, has long dominated studies of economic development among researchers and policymakers. While the Solow growth model makes no explicit assumptions about the nature of inter-country relationships, its design tends to emphasize individual country characteristics such as national labor supply, capital stock, and level of technological progress.

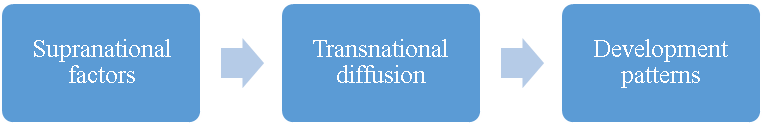
In contrast to this approach, this paper joins a broader shift in research focus within sociology and the social sciences more generally away from the characteristics of individual countries and toward an examination of supranational factors and their effects on relationships among countries. Dependency theory and world-systems theory, as pioneered by Wallerstein (1974) and others, utilizes a Marxist framework to understand a persistent power relationship between so-called “core” and “periphery” countries. World-polity theory, as developed by John Meyer and others (Meyer, Boli-Bennett, and Chase-Dunn 1975; Meyer 1980; Meyer et al. 1997), builds on a Weberian and new institutionalist emphasis on culture to explain the spread of certain institutional forms around the world. The enduring effects of colonial legacy on contemporary patterns of development has become the focus of a rapidly growing body of quantitative research that has identified important differences in colonizer identity, colonial regime structures, and settlement patterns (Acemoglu, Johnson, and Robinson 2001, 2002; M. K. Lange 2004; M. Lange, Mahoney, and Hau 2006). The effect of war and inter-state competition on the formation of nation-states and the motivations of political leaders to promote economic growth have been explored in detail by Tilly, Centeno, etc. Other super-country-level patterns have also been explored, driven by factors such as geographical proximity, religion and ideology (including communism), and so on.

**Transnational Diffusion**

One hypothesized mechanism by which supranational factors affect patterns of development across groups of countries is that of transnational diffusion. In the context of economic development, the diffusion of technologies—from physical technologies such as computers and mobile phones to organizational technologies such as double-entry bookkeeping and professional resumes—across countries has been uneven in terms of historical timing and levels of adoption. An examination of specific inter-country channels of diffusion shed light on the particularities of such linkages.

* E.g. rise of the Republic of Letters and scientific societies in Western Europe during Scientific Revolution
* E.g. Atlantic Crossing
* E.g. technology sharing: Britain to Meiji Japan, Japan to colonial Korea, postwar South Korea to Bangladesh textile industry
* E.g. military contact ((e.g. Napoleon, European ships in East Asia): Tilly, Centeno)

While this paper does not directly test for the effects of transnational diffusion as a mechanism *per se* (which Wejnert 2004 does), this paper takes an important step toward exploring the relationship between broader supranational factors that may influence diffusion networks and shape patterns of development. In particular, this paper investigates the relationship between linguistic distance—how similar are the dominant languages of any two countries?—and patterns of development.



Reduced form

**Figure 2**. Supranational factors influence patterns of transnational diffusion, which in turn affect development patterns across groups of countries. This paper examines the “reduced form” relationship between one supranational factor—linguistic distance—and patterns of economic growth.

Need to look up all their references as well!

Global spread of cricket:

(Kaufman and Patterson 2005)

Look up “diffusion” and “policy diffusion” literature in sociology journals! And cultural diffusion!

There is evidence to suggest that linguistic distance may influence—and thus serve as a good proxy for—the configurations of transnational diffusion networks. (need to cite evidence here)

In addition to contributing to the broader theoretical discussion over country clustering, diffusion, and development, this paper also addresses a number of issues that have arisen in the study of supranational patterns of economic development.

**Strict Causality vs. Patterns of Association in Cross-Country Studies**

First, this paper advocates a shift away from a focus on strict causal models that attempt to isolate and identify the effects of individual factors on patterns of economic development. There are two strongly interrelated reasons for this. First, given the limited sample size of any country-level analysis (there are currently 193 United Nations member states), even with country-year data, relative to the myriad potential explanatory factors, the system is statistically overdetermined.

Second, there exists too much endogeneity among potential explanatory factors to be able to sufficiently isolate a single factor and identify its causal effects. In the debate over whether institutions or capital accumulation (including human capital) better explain variation in development outcomes, the two are too intimately related to isolate the effects of one without the other. Greater levels of education likely strengthen pro-growth institutions, which in turn increases institutional capacity to improve levels of education. Rather than attempt to cut the Gordian knot of causal identification in country-level patterns of development, I argue that researchers should instead devote more resources to mapping out cross-country patterns and then examining the specific channels or mechanisms by which these patterns may arise.

**Levels of Income vs. Patterns of Growth**

Second, previous work on economic development has focused too much on income levels, which suffer from high temporal autocorrelation. A number of studies (cite here) have shown the long-term persistence of relative income levels over time. Of the top 10 wealthiest countries today by per capita GDP, all 10 of them were the wealthiest countries in 1800. Moreover, the historical mechanisms of accumulating returns and path dependence partly explain why there exists such high temporal autocorrelation in the long duree (cite here). Thus, recent studies attempting to explain contemporary variation in levels of development are merely picking up which countries were part of the first Industrial Revolution.

To address this problem, this paper examines patterns of economic growth rather than levels of income. While the relative ranking of countries by income would reasonably be fairly consistent over time, it is less clear that the specific temporal shapes of their economic development over time would be contingent on their prior levels of income. Moreover, I have gone one step further by standardizing growth rates: for any given country, their growth rates have been de-meaned and re-scaled to unit standard deviations. This further removes the effects of prior levels of income on patterns of economic growth. Without this step, one could expect a meaningful relationship between higher levels of income and lower average growth rates (following the beta-convergence of the neoclassical growth model) and with lower variance due to lower volatility in growth rates. As a final step to control for income levels, I have included current income levels as a control variable in my analysis.

This is not only important from a statistical methodology standpoint, but I would argue that differing patterns of development are an important dependent variable to investigate in itself. A number of scholars have examined variation in types of development:

* Peter Evans, Atul Kohli, Sandbrook, Kerala (Desai)
* Communist-style heavy industry approach
* Import-substitution in Latin America (w/ ECLA) vs. export-oriented mode in East Asia

This is also important in the context of environmental sustainability. Income levels alone mask important differences in environmental costs.

**Rethinking Geography and Regional Categories**

Third, many scholars, policymakers, and private sector actors take certain regional groupings of countries for granted. For example, multinational corporations often break down global markets into regional categories such as North America, Latin America, Europe, Middle East, Africa, and Asia Pacific. While this may be useful heuristically and may be sufficiently valid for certain purposes, not enough work has been done to question the validity of this geographical assumption. For one, we can see that there are a number of important exceptions to geographical groupings.

The Anglophone countries are perhaps the best example where the commonalities across Britain, the United States, Canada, Australia, and New Zealand[[1]](#footnote-1) stand in direct contradiction to their vast geographic spread. In terms of levels of income, these countries are much more closely related than they are to their geographical neighbors. Moreover, on a vast number of economic, political, social, and cultural dimensions, these countries constitute a fairly coherent grouping that defies their relative geographical positioning (cite examples).

Another example is sharp breaks in country clustering where there are none geographically. While countries across Western and Eastern Europe are distributed on a fairly continuous basis geographically, there exists a sharp break across the Iron Curtain that appears in contemporary economic, political, and social measures. In East Asia, difference between Northeast Asia and Southeast Asia. In Africa, difference between North Africa and Sub-Saharan Africa.

To address this problem, I look at linguistic distance and control for geographic distance between countries. As I will show, there are interesting relationships in patterns of economic growth that defy geographic proximity as a dominant factor.

In addition to these contributions to theory, this paper also makes data and statistical methodological contributions. In terms of data, this paper employs a high-quality state-of-the-art linguistic dataset based on the presence or absence of specific phonemes that allows for much more continuous measures of linguistic distance, particularly over the standard Ethnologue dataset. In terms of methodology, this paper introduces a non-parametric statistical procedure commonly used in genetics research that would be especially useful to sociologists given its emphasis on relations between data points rather than the standard variable characteristics of the individual observations. These two additional contributions will explained in greater detail in the data and methods section.

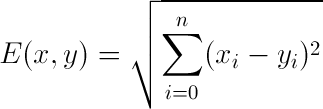
**Data and Methods**

Data for this study come from publicly available cross-country datasets. For each variable, distance matrices are generated which contain distance measures for every pair of countries, yielding a symmetric *n* x *n* matrix containing distinct distance values where *n* is the number of countries in the sample. Simple and partial Mantel tests, commonly used in population genetics, are used to calculate Pearson *r* correlations between these distance matrices and significance levels (i.e., *p*-values). Conventional statistical techniques such as OLS regression cannot be used due to complete autocorrelation within each distance matrix. The primary dependent variable is differences in patterns of economic growth and the primary independent variable is linguistic distance. Control variables include geographic distance, differences in income levels, and bilateral trade.

*Patterns of Economic Growth*

Annual per capita GDP data by country from 1950 to 2008 comes from the Maddison dataset. Per capita GDP values are in constant 2000 U.S. dollars adjusted for purchasing-power parity (PPP). Rolling annualized five-year averages are calculated producing economic growth values (trailing five years) from 1955 to 2008 for each country. Using five-year average annualized per capita GDP growth rates smooths out measurement error for any given year. These annualized five-year average economic growth values are then standardized within each country such that each country’s mean growth value is zero and standard deviation is one. Standardizing the economic growth values within countries reduces the effect of historical income levels, which are known to be associated with average growth rates and growth rate volatility (add citation).

These standardized values for each country-year are then used to generate a distance matrix for every country-pair based on the Euclidean distance across all years for any two countries. In the equation below, *x* and *y* are two different countries and *i* is each year’s five-year annualized growth rate for *n* years 1950 to 2008.



*Linguistic Distance*

While previous work has examined the relationship between linguistic distance and economic development, much of this work has relied on an overly coarse dataset. The Ethnologue dataset is widely used in the social sciences, particularly within economics and political science (cite here). It was created by linguistic experts who manually coded and produced a linguistic tree for all known languages. However, it suffers from being too coarse for measures of linguistic distance. The only way to derive linguistic distance between two languages is by counting the number of common nodes they share. As has been noted by others (cite here), this masks certain kinds of variation such as a potential log relationship between actual linguistic and number of common nodes. Thus, when Spoloare and Wacziarg (2009) run a regression with both linguistic distance using the Ethnologue data and genetic distance, they find genetic distance to be a better predictor of economic development. This is more likely due to the fact that the linguistic distance measure was coarser than the genetic distance measure, a problem they note.

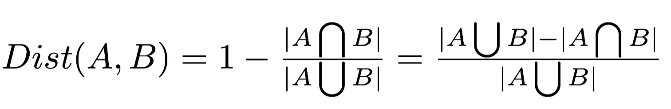
To address this problem, I use a state-of-the-art linguistic dataset of 2,082 languages with binary presence/absence data for 728 phonemes. This is taken directly from the Ruhlen database that was modified and used by Creanza et al’s (2015) paper in the *Proceedings of the National Academy of Sciences*. The primary advantage of this dataset is that one can use it to derive more continuous linguistic distance measures than the Ethnologue data. Moreover, because these languages are coded at the phonemic level, there is less margin for the interpretation of a language’s relative position due to factors that may not be orthogonal to other variables of interest such as level of development.

Each country is assigned a single “dominant” language corresponding to the language within that country with the greatest number of speakers. This raises some difficulties for certain countries. For example, for some countries, the plurality-spoken language may in fact constitute a relatively small fraction of the total population due to the existence of other widely spoken languages. In Kenya, the “dominant” language as determined by plurality is XXX but the second most widely-spoken language XXX has nearly as many speakers. The assignment of a single “dominant” language is a simplification that necessarily excludes some linguistic information. However, if anything, this should make the results from this study more conservative given that the use of a single “dominant” language is a coarser measure of cultural similarity.

In addition to this methodological issue, at a more data-specific level, there is an issue with a number of countries where the “dominant” language as determined by sheer plurality in the Ruhlen dataset does not appear to adequately correspond to the true plurality-spoken language. For example, within the Ruhlen dataset, Australia’s most widely-spoken language is an aboriginal language rather than Australian English. These omissions are due to the fact that the version of the Ruhlen dataset being used was originally designed to address indigenous population characteristics rather than modern population factors. As a result, a number of manual substitutions were made based on CIA data on most widely-spoken language. In the analysis, the manually edited and non-manually edited Ruhlen datasets were used and the findings did not change.

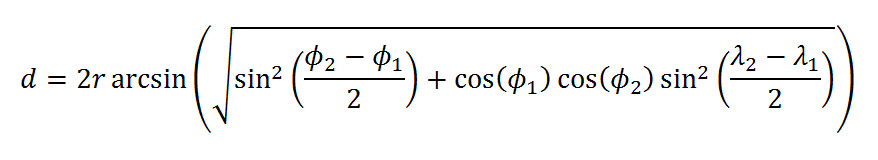
(show table of language manual edits)

Using this new dataset where each country is assigned a single “dominant” language, I created a linguistic distance matrix where the linguistic distance between any two countries is calculated as the Jaccard distance on the presence/absence data for each phoneme. The Jaccard distance measure is commonly used within linguistics (add citation) and scales phonemic differences by the number of phonemes present in both languages.



*Geographic Distance*

Geographic data for each country comes from Google. For each country, these consist of latitude and longitude values of the geographic centroid. Geographic distance between any two countries is calculated as the great-circle distance between these two sets of geographic coordinates using the haversine formula. This provides the shortest distance between two points on the surface of a sphere. All country-pair geographic distances are then logged.



*Income Levels*

*Bilateral Trade*

Bilateral trade data comes from UN COMTRADE. Using national import records, the UN COMTRADE dataset provides a monetary value of one-way trade flows in nominal US dollars for each importing country by exporter source. For example, the importing country USA received $x million in goods and services from exporting country XXX in 2000.

Rates of missing data are much higher for these bilateral trade data than for other distance measures. In order to increase the number of available observations and to reduce the particular effect of any single year, for each import-export country-pair the annual trade values over the decade 1991-2000 were averaged wherever data was available, yielding XX observations.

Both directions of trade flows for every country-pair were summed to produce a single total trade value for each country-pair.

(insert table for descriptive statistics for each variable)

**Simple and Partial Mantel Test**

Simple and partial Mantel tests are commonly used methodological tools in population genetics to evaluate the relationships between distance matrices (add citations). Conventional regression methods cannot be used effectively on distance matrices due to the complete autocorrelation inherent within any distance matrix. Within the standard multilevel model framework, autocorrelation is often addressed through the use of clustered standard errors. However, for distance matrices, autocorrelation does not exist at any sub-sample level but rather permeates the entire sample due to the triangle inequality property: for any three points in the underlying raw data A, B, and C, their three corresponding distance values A-B, B-C, and A-C in the distance matrix will be mutually correlated by transitivity. Thus, an application of conventional regression methods would overestimate the significance of any relationship by failing to adequately compensate for this inherent correlation across distance values.

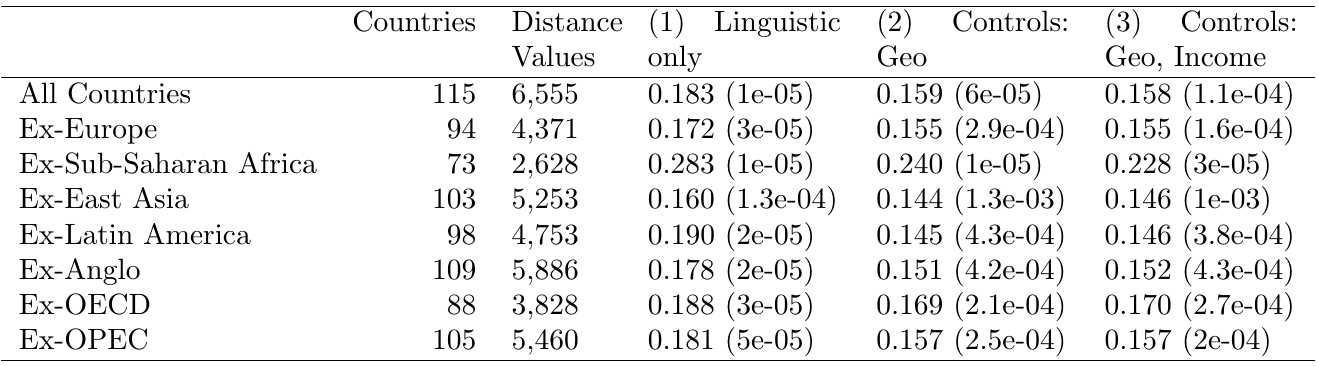
The Mantel test using randomized permutations offers one approach to addressing this autocorrelation problem. By randomly permutating rows (or columns) of one of the distance matrices DX and then calculating correlation coefficients between that randomized distance matrix DX and the other distance matrix DY, one can computationally produce null hypothesis correlation coefficients. Given a number of random permutations, a P-value is derived by checking how many of these randomized correlation coefficients exceeded the value of the original correlation coefficient. For example, if 99,999 random permutations were performed and 9 of them produced higher correlation coefficients than the original one, the P-value is therefore 9 plus 1 (for the original combination) out of 99,999 plus 1, in other words P = 10/100,000 or 1e104.[[2]](#footnote-2)

Examples of Mantel test:

Desmet et al 2011

**Table 1**. Mantel *r* Correlations between Linguistic and Growth Pattern Distances by Subset

Columns (1)-(3) contain Mantel *r* correlation coefficients with two-sided *p*-values in parentheses. Column (1) is a simple Mantel test between linguistic distance and growth pattern distance without controls. Column (2) and (3) add controls for log-geographic distance and both log-geographic distance and log-2008 income distance, respectively. All Mantel *r* coefficients are significant at the *p* < .05 level.



**Results**

There is a significant correlation between linguistic distance and growth pattern distance, even after controlling for geographic distance, income levels, and bilateral trade. This relationship is robust to the exclusion of certain subsets of countries that would likely significantly contribute to the relationship. Table 1 shows a summary of the Mantel correlation coefficients and their associated P-values across each set of controls and sample subsets (Appendix A includes a list of member countries for each subset). In all cases, the inclusion of additional controls lowers the Mantel *r* correlation coefficient but not to the degree that it nullifies the relationship between linguistic distance and growth patterns.

For the subsample analysis, Anglo countries drive the greatest amount of the overall language-growth relationship, which makes sense given their outlier status as a group on both linguistic distance and income level characteristics. OPEC countries also appear to drive a substantial portion of the relationship, likely due to the commonalities in their economic growth patterns due to oil prices. Table 2 shows the summary statistics for each subset. From this, we can see that OPEC countries have some of the closest growth patterns.

Within country subsets, there are positive Mantel *r* coefficients; however, none of these are significant, indicating that either this relationship does not hold within regions (i.e., driven primarily by inter-regional differences) or there is insufficient power to adequately assess this relationship. Figure XX shows the Mantel correlogram across different geographic distance thresholds, revealing significant positive relationships between growth patterns and geographic distance for geographic distances between XX and XX but not less than this.

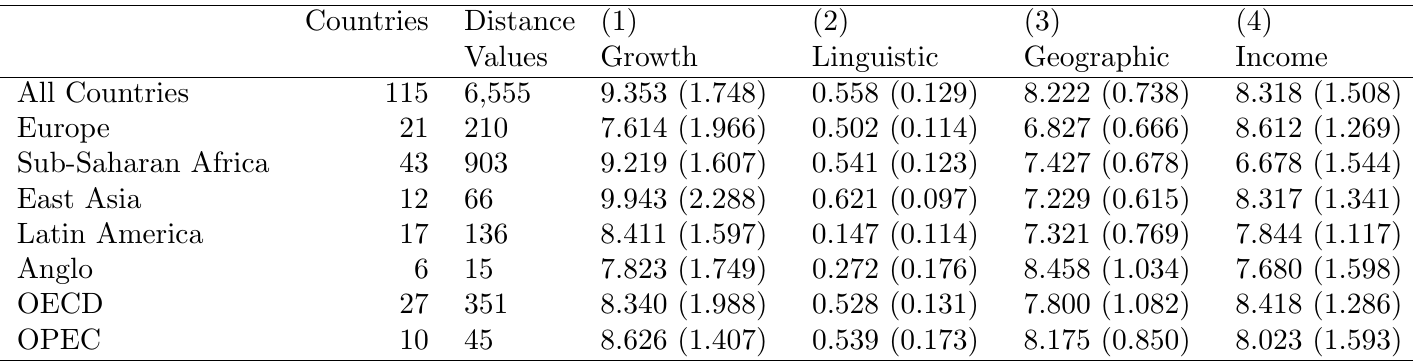
*Correlation table*

Linguistic distance is more strongly correlated with growth pattern distance than with income levels.

*Interesting areas of divergence*

Columns (1) and (2) are distance values with no interpretable units. Column (3) contains log-geographic distance in log-miles Column (4) contains log-2008 income level distance values in log-2008 U.S. dollars. Standard deviations are in parentheses.

**Table 2**. Average Distance Values by Distance Measure and by Subset



**Discussion**

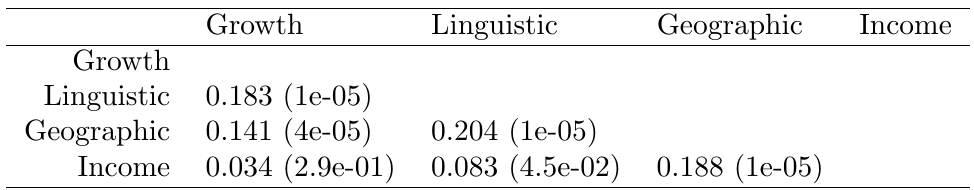
We should think about other supranational factors such as linguistic distance that may affect patterns of development.

Further work should be done. Detailed comparative historical work should be done to outline the particular diffusion channels by which technologies have flowed between countries and within groups of countries. For sufficiently comprehensive datasets, inter-country networks of diffusion should be mapped out.

Finally, more should be done to move away from looking at sheer levels of development or income and instead examine differing patterns of development. Differing approaches to trade strategies, differing approaches to dealing with labor and wages, and differing uses of environmental resources with differing effects should be examined.

Includes all countries (number of countries = 113; distance values = 6,555). P-values are in parentheses.

**Table 3**. Pair-wise Simple Mantel *r* Correlations between Distance Measures



1. South Africa is a special case [↑](#footnote-ref-1)
2. Note: There has been some controversy over the extent to which these randomized permutations produce an appropriate set of null hypotheses that sufficiently controls for spatial autocorrelation (add citations). [↑](#footnote-ref-2)