Empirical Seminar: Draft Paper

Global Inequality, Linguistic Distance, and Diffusion

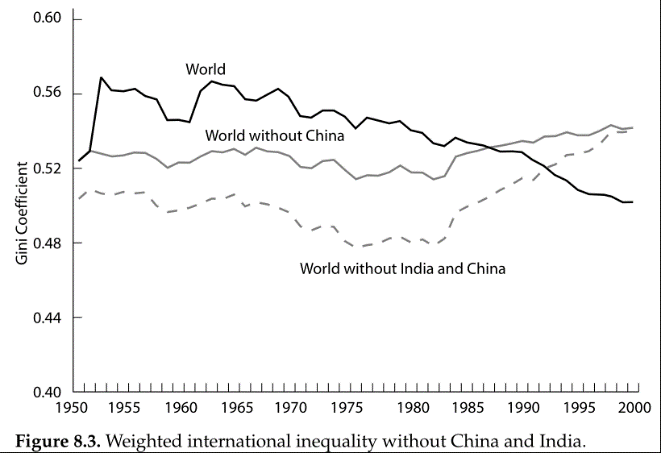
Kyle Chan

May 23, 2016

**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.

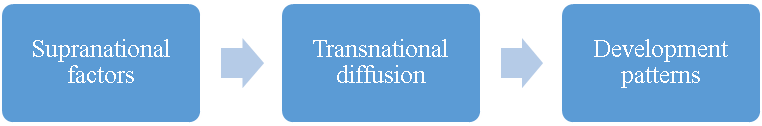
In contrast to this approach, this paper joins a broader shift in research focus within sociology and the social sciences more generally away from the characteristics of individual countries and toward an examination of supranational factors and their effects on relationships among countries. Dependency 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 (Centeno 2002; Tilly 1990). Other supranational 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 blast furnaces to organizational technologies such as double-entry bookkeeping and professional curricula vitae—across countries has been uneven in terms of historical timing and levels of adoption. The levels of adoption and types of utilization of certain technologies may in turn affect a country’s prospects for economic development.

Historically, certain networks or channels of technological diffusion have been particularly well explored. The so-called “Republic of Letters” across Western Europe from the early modern period through the Scientific Revolution and the Enlightenment was comprised of formal and informal exchanges of ideas, often by handwritten letter, among individual thinkers and later scientific societies. Leibniz alone had a network of over 400 correspondents across France, Italy, England and even China (via the Jesuits) through which he shared and discussed new developments in mathematics and natural philosophy (Harris 2006). Shortly after the Gramme dynamo, the first commercially viable electric power generator, was invented by in Belgium around 1871, an American physics professor at Cornell built a working replica using accounts published in the British journal *Engineering* (Brittain 1974). In the 1880s as Meiji Japan strove to modernize, Japanese cotton spinning firms imported British mill designs, engineers, and building materials wholesale, sometimes down to the very bricks from Lancashire (Saxonhouse 1974). A similar story occurred later in the 1970s when the South Korean conglomerate Daewoo invited entrepreneurs from Bangladesh to visit Kore and learn the textile trade. Bangladesh’s textile exports are now the second largest in the world after China’s. (South Korea, in turn, received industrial technology and know-how from the Japanese during a half-century period of colonization.)

While this paper does not directly test for the effects of transnational diffusion as a mechanism *per se* (which Wejnert (2005) does for diffusion and democracy), this paper takes a 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. While sharing similar languages may simplify the task of translation to some extent, this paper does not consider the direct effect of language on technological diffusion *per se* but rather uses linguistic distance as a proxy for broader forms of cultural distance that may in turn influence the formation of diffusion networks.



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—language—and patterns of economic growth.

A broad body of theoretical and empirical literature has emerged across sociology and the social science around transnational diffusion.

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!

Waves of war

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**

This paper advocates a shift away from an exclusive 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) relative to the myriad potential explanatory factors, the system is statistically overdetermined, even with country-year data.

Second, there exists too much endogeneity among potentially important explanatory factors to be able to sufficiently isolate a single factor and identify its causal effects. For example, in the debate over whether institutions or capital accumulation (including human capital) better explain variation in development outcomes, institutions and (human) capital are too intimately interrelated to make such a comparison useful. Greater levels of education likely strengthen pro-development institutions, and greater institutional capacity likely increases 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 versus Patterns of Growth**

Previous work on economic development has focused too much on income levels, which suffer from high temporal autocorrelation. A number of studies have shown the long-term persistence of relative income levels over time (Chanda, Justin Cook, and Putterman 2014; Comin, Easterly, and Gong 2010; Hibbs and Olsson 2004; Putterman and Weil 2010). Moreover, the historical mechanisms of accumulating returns and path dependence partly explain why there exists such high temporal autocorrelation over the longue durée (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 growth rates over time would be contingent on their prior levels of income. Moreover, I have taken the additional step of standardizing growth rates: for any given country, its 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, I have included current income levels as a control variable in my analysis.

An examination of patterns of economic growth is not only important from a methodology standpoint; understanding variation in patterns of development is an important sociological phenomenon to investigate in itself. Static income levels capture only one dimension of development, ignoring the important ways in which the experience of development as an evolving process over time differs across countries.

A number of scholars of development have examined the ways in which groups of countries have differed in their historical paths. Peter Evans (1995), Atul Kohli (2004), and Arthur Sandbrook et al. (2007) have charted various pathways of development experience outside the West from the state-led, export-oriented experiences of a number of East Asian countries to the social democratic political economies of Kerala and Mauritius. Any student of communism, particularly of the Soviet Union and Communist China, is familiar with the peculiar focus on heavy industry and economic autarky of these countries’ development experiences. The role of ECLA on Latin American countries’ approach to import-substitution. As scholars such as Frank Dobbin (1994) have shown, even among Western capitalist, there may be drastically different historical paths to development.

**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 stand in direct contradiction to their vast geographic dispersion. In terms of levels of income, these countries are much more closely related than they are to their geographical neighbors. Moreover, on a large number of economic, political, social, and cultural dimensions, these countries constitute a fairly coherent grouping that defies their relative geographical positioning (cite examples).

Another case where geographic distance may diverge from cultural distance is that of 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 former 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.

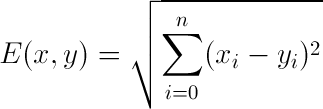
In addition to these contributions to theory, this paper also makes data and statistical methodological contributions. In terms of data, this paper employs a high-quality state-of-the-art linguistic dataset based on the presence or absence of specific phonemes that allows for more continuous measures of linguistic distance, particularly over the standard Ethnologue dataset. In terms of methodology, this paper introduces a non-parametric statistical procedure commonly used in population genetics research that would be especially useful to sociologists given its focus on relations between data points rather than the characteristics of individual observations. These two additional contributions will be discussed in greater detail in the next 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 correlation coefficients between these distance matrices and corresponding significance levels. 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 come from the Maddison Project (add citation). Per capita GDP values are in constant 1990 international Geary-Khamis dollars, which are adjusted for purchasing power (PPP). Rolling annualized five-year averages are calculated, yielding one economic growth rate value for each country and for each year (trailing five years) from 1955 to 2008. The use of five-year average annualized per capita GDP growth rates reduces the effects of measurement error for any single year. These annualized five-year average economic growth rates are then standardized within each country across years 1955 to 2008 such that each country’s mean annualized growth rate is zero with unit standard deviation. Standardizing the economic growth values within countries in this way reduces the effect of historical income levels, which are known to be associated with average growth rates and growth rate volatility (add citation).



**Equation 1**. Euclidean distance formula between any two points *x* and *y* for *n* dimension

These standardized values for each country-year are then used to generate a distance matrix of growth pattern distance values 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*

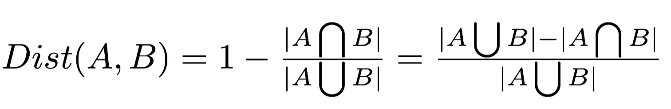
While previous work has examined the relationship between linguistic distance and economic development, much of this work has relied on an overly coarse dataset. The Ethnologue dataset is widely used in the social sciences, particularly within economics and political science (cite here). It was created 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 distance 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. Most of these edits occurred in English- and Spanish-speaking countries, and the closest language approximation was used. Ultimately, the use of non-edited versus edited linguistic datasets did not change the findings (See Appendix A for a list of manual edits to the linguistic data.)

**Equation 2**. Jaccard distance formula between any two points *A* and *B* based on phonemic presence/absence data

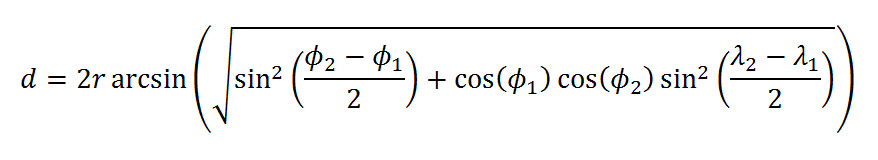


Using this new dataset where each country is assigned a single “dominant” language, I created a linguistic distance matrix where the linguistic distance between any two countries is calculated as the Jaccard distance on the presence/absence data for each phoneme. The Jaccard distance measure is commonly 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’s online set of canonical public-use databases for software developers (https://developers.google.com/public-data/docs/canonical/countries\_csv). For each country, geographical data consist of latitude and longitude values for the country’s geographic center. The geographic distance between any two countries is calculated as the great-circle distance between these two sets of latitude-longitude 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 to adjust for scale.

**Equation 3**. Haversine formula for calculating the great-circle distance between any two latitude-longitude pairs.



*Income Levels*

2008 per capita GDP income levels come from the Maddison Project. As with the economic growth data, these per capita GDP values are in constant 1990 international Geary-Khamis dollars, which are adjusted for purchasing power (PPP). A distance matrix for income levels is created using the logged absolute difference between the 2008 per capita GDP income levels for each country-pair.

*Bilateral Trade*

Bilateral trade data comes from UN Comtrade, which is the United Nations’ publically-available online database of official global trade statistics (http://comtrade.un.org/). Derived from national import records, the UN Comtrade dataset provides a monetary value of one-way trade flows in nominal U.S. dollars for each importing country by exporter source. For example, in 2000 the U.S. imported $233 million in goods and services from exporting country Canada.

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.

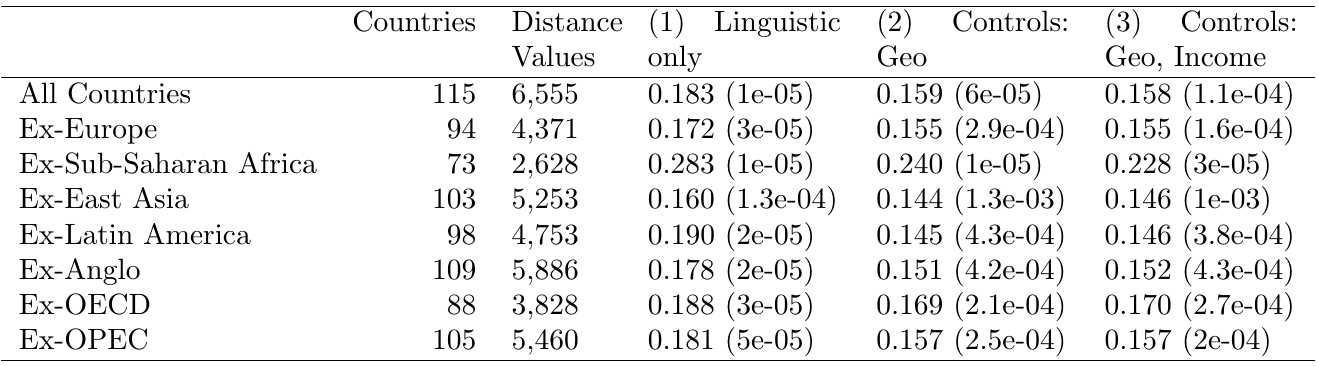
Both directions of trade flows for every country-pair were summed to produce a single total trade value for each country-pair. For example, U.S. exports to Canada and Canadian exports to the U.S. are summed to produce a single bilateral trade value for the U.S.-Canada country-pair.

**Simple and Partial Mantel Tests**

Simple and partial Mantel tests are commonly used methodological tools in population genetics to evaluate 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, conventional regression methods such as OLS would overestimate the significance of any relationship by failing to adequately compensate for this inherent correlation across distance values.

**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.

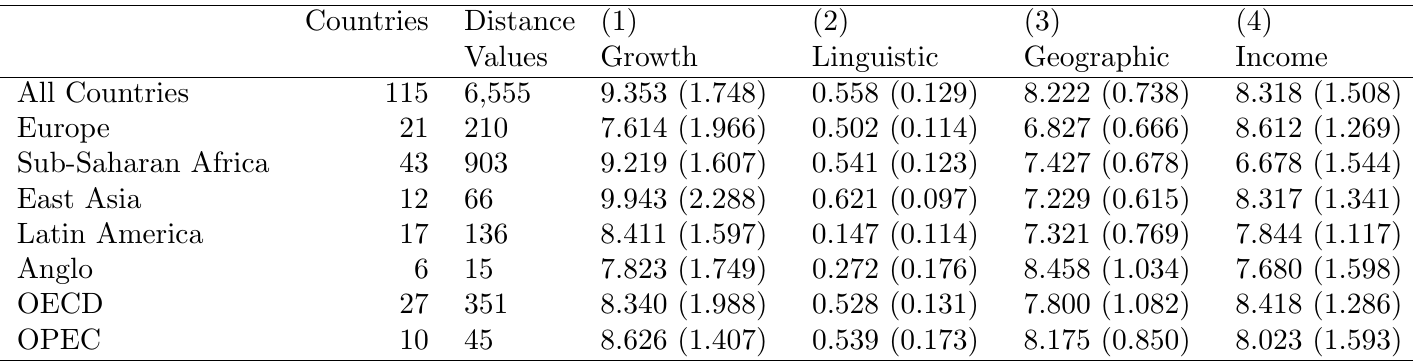


The Mantel test using randomized permutations offers one approach to addressing this autocorrelation problem. By randomly permutating rows (or columns) of one of the distance matrices DX and then calculating correlation coefficients between the randomized distance matrix DX and the comparison distance 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.[[1]](#footnote-1)

While the simple Mantel test as described above tests for the relationship between two distance matrices, the partial Mantel test allows for the assessment of such a relationship while controlling for one or more additional distance matrices. To evaluate the relationship between two distance matrices DX and DY while controlling for a third distance matrix DZ, the residuals from the partial correlations of DX on DZ and DY on DZ are calculated, yielding two new residual matrices DXZ and DYZ, respectively. Then the simple Mantel test with randomized permutations is performed on these residual matrices to generate a correlation coefficient and *p*-value that excludes the effect of the joint relationship of DZ on DX and DY.

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



Examples of Mantel test:

Desmet et al 2011

**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 *r* correlation coefficients and their associated *p*-values across each set of controls and sample subsets (Appendix B includes a list of member countries for each subset).

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 across Distance Variables*

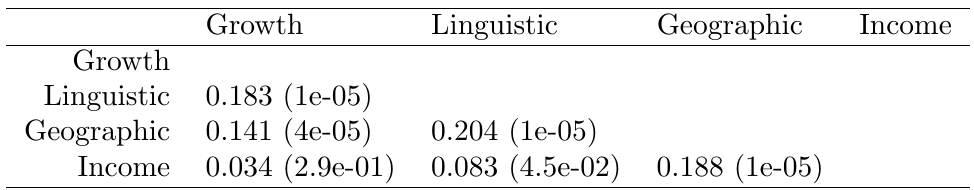
Table 3 shows the simple Mantel *r* correlation coefficients (and corresponding *p*-values) between each set of distance measures, providing a set of validity tests.

First, the relationship between distances for growth patterns and income levels is close to zero and not significant at the 95% level. This validates the approach taken to insulate growth rates from income levels via within-country growth rate standardization.

Second, the relationship between geographic distance and linguistic distance as well as the relationship between geographic distance and income level distance is relatively strong and statistically significant. This reconfirms prior findings of the underlying effect of geographic distance on linguistic distance and on income levels.

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

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



Third, linguistic distance is more strongly related to growth patterns than is geographic distance. This implies that language may be a more meaningful way of understanding cross-country clustering patterns in development than geography or simple regional groupings.

*Interesting areas of divergence*

**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.

1. 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 (Guillot and Rousset 2013; Legendre, Fortin, and Borcard 2015). [↑](#footnote-ref-1)