Global income inequality is massive and persistent when China’s recent economic rise is excluded from the analysis. While global income inequality appears to have declined in recent decades, the work of Milanovic (2005) as well as Firebaugh and Goesling (2004) has shown that this decline is primarily driven by the economic rise of China. Sans China, the global population-weighted Gini coefficient remains above 0.52 and has in fact grown in recent decades. 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.



Figure 1: Milanovic 2005

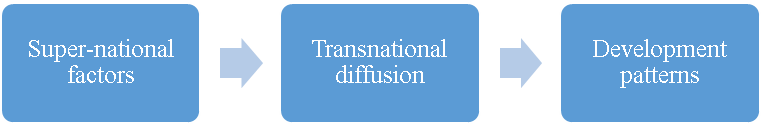
What explains global income inequality across countries? This paper joins a broader shift in research away from a focus on the characteristics of individual countries toward an examination of cross-country patterns and groupings including the relationships between countries. The Solow growth model, named 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 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 individual-country focus, a long line of research in sociology and across the social sciences has explored the relationships among countries, searching for broader clusters of development. Dependency theory and world-systems theory, as pioneered by Wallerstein (1974) and others, utilizes a Marxist framework to understand a persistent 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 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.

One meso-level social mechanism has received particular attention: 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 test for the specific effects of transnational diffusion as a mechanism (which Wejnert 2004 does), this paper takes a important step toward exploring the relationship between broader super-country factors that may influence diffusion networks and patterns of development. In particular, this paper investigates the effect of linguistic distance—how similar are the dominant languages of any two countries?—on patterns of development via its effect on diffusion networks.



Reduced form

Is there a relationship between linguistic distance and patterns of economic growth across countries, even after controlling for geography, trade, and income levels?

This paper addresses a number of issues that have arisen in the study of super-country patterns of economic development.

Strict Causality vs. Identifying Patterns of Association

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.

Income Levels 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 Categorization

Third, many scholars and policymakers as well as firms take for granted certain regional groupings of countries. For example, multinational corporations often break down global markets into the following regional categories: 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 and theoretical approach, 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.

High-Quality Linguistic Data

Fourth, 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 (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 X,XXX languages with binary presence/absence data for 739 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 political considerations.

New Methodological Tool: Mantel Test

Fifth, this paper introduces a new non-parametric statistical tool that is commonly used in genetic analysis but rarely used within sociology. Rather than simply testing for the relationship between two sets of variables per standard OLS or GLM regression models, the Mantel test allows for the examination of relationships between pairs of objects. Rather than

1. South Africa is a special case [↑](#footnote-ref-1)