LINGUISTIC DISTANCE AND MARKET INTEGRATION IN INDIA

JAMES FENSKE[†] AND NAMRATA KALA[⋆]

Abstract. The role of cultural distance in market integration, particularly in the developing world, has received relatively little attention. Using prices from over 200 South Asian markets spanning 1861 to 1921, we show that linguistic distance correlates negatively with market integration. A one-standard-deviation increase in linguistic distance predicts a reduction in the price correlation between two markets of 0.121 standard deviations for wheat, 0.181 for salt, and 0.088 for rice. While factors like genetic distance, literacy gaps, and railway connections are correlated with linguistic distance, they do not fully explain the correlation between linguistic distance and market integration.

[†]University of Warwick
[⋆]MIT Sloan School of Management
E-mail addresses: J.Fenske@warwick.ac.uk, kala@mit.edu.
Date: June 24, 2020.
We are grateful to Latika Chaudhary, Martin Fiszbein, Marc Klemp, Alan Taylor, Romain Wacziarg, and to audiences at the Association for the Study of Religion, Economics, and Culture, George Mason University, Pontificia Universidad Católica de Chile the University of Manchester, the University of Toulouse, and the University of Warwick for their comments. Extra thanks are due to Marlous van Waijenburg for sharing additional price data with us, and to Paradigm Data Services (inquire@pdspl.com), Connie Yu and Mina Rhee for their assistance in data entry.

LINGUISTIC DISTANCE AND MARKET INTEGRATION IN INDIA

Abstract. The role of cultural distance in market integration, particularly in the developing world, has received relatively little attention. Using prices from over 200 South Asian markets spanning 1861 to 1921, we show that linguistic distance correlates negatively with market integration. A one-standard-deviation increase in linguistic distance predicts a reduction in the price correlation between two markets of 0.121 standard deviations for wheat, 0.181 for salt, and 0.088 for rice. While factors like genetic distance, literacy gaps, and railway connections are correlated with linguistic distance, they do not fully explain the correlation between linguistic distance and market integration.

1. Introduction

Economic historians use market integration as a key measure of economic development (Shiue and Keller, 2007; Studer, 2008). Although language barriers have been stressed in the macroeconomic literature as inhibiting trade and the diffusion of technology (Guiso et al., 2009; Spolaore and Wacziarg, 2009), the role of these variables in market integration within countries, particularly in the developing world, has received comparatively little attention, despite the sizable economic impacts that these barriers can have in other contexts (Ashraf and Galor, 2013; Spolaore and Wacziarg, 2018). In this paper, we consider the economy of colonial India, in which a large number of dissimilar languages prevail. In particular, we ask: do market pairs that are more linguistically distant display less market integration, conditional on physical distance and other measures of dissimilarity?

We collect data from *Wages and Prices in India* on grain and salt prices for 206 South Asian markets between 1861 and 1921. These markets span the territories of modern-day Bangladesh, Burma, India, and Pakistan. We merge these markets to populations by language collected from the 1901 colonial census of India. We map these languages into 257 ISO language codes from *Ethnologue*, which also provides us with language trees. Taking the correlation coefficient between the price series at a pair of markets i and j, we show that, conditional on physical distance, religious distance, dissimilarities in geography, and fixed effects for markets i and j, prices at i and j are less correlated if i and j are more linguistically distant. Our estimates suggest that two markets with unrelated languages will, compared to two markets sharing a common tongue, have correlation coefficients that are 0.067 less in the case of wheat, 0.189 less in the case of salt, and 0.035 less in the case of rice, relative to means of 0.81 (wheat), 0.54 (salt) and 0.81 (rice) across all market pairs

in the data. These are large relative to the coefficients we estimate for physical distance, and suggest a possible role for cultural distance in raising trade costs, even for relatively low-value, homogenous goods.

In assessing the mechanisms that link linguistic distance to market integration, we turn to both the economic literature and to the history of colonial India. Linguistic distances need not matter exclusively for market integration through language; that is, language itself is one of many imperfect measures of broader ancestral distance. This concept may include shared history, institutions, culture, and norms, among other characteristics (Spolaore and Wacziarg, 2016). Language barriers may represent more general barriers to the transmission of vertical traits (Spolaore and Wacziarg, 2009, 2018). They may capture differences in tastes, and hence the presence or absence of certain markets (Atkin, 2013, 2016). They may affect the costs of information transmission and coordination (Gomes, 2014). They may otherwise affect trade costs through interaction, migration, business connections, conflict, or xenophobia (Bai and Kung, 2017; Laval et al., 2016; Rauch and Trindade, 2002). They may work through costs of language or education acquisition (Isphording and Otten, 2014; Jain, 2017; Laitin and Ramachandran, 2016; Shastry, 2012). They may correlate with common preferences for public goods, redistribution, and infrastructure (Desmet et al., 2020, 2012, 2017).

To assess which of these explanations may account for our results, we assemble data from a wide range of primary and secondary sources. We show that market pairs that are more linguistically distant from each other are also more genetically distant, but that this summary measure of barriers to the diffusion of technological and institutional innovations is not itself a sufficient statistic for the coefficient on linguistic distance. We find little evidence that linguistic distance predicts missing markets or fewer shared trading communities. Historical differences in literacy across market pairs do correlate with linguistic distance, but do not fully account for its correlation with price integration. Though more linguistically similar market pairs evidence longer periods of time connected to the colonial railway system, this fails to explain away the correlation. Thus, while linguistic distance may have operated in part as a marker of other population differences, as a barrier to the acquisition of similar levels of human capital, and as a barrier to the co-acquisition of public goods that facilitated trade, no one of these mechanisms can fully account for the barriers of linguistic cleavages.

Our paper contributes principally to two literatures. The first investigates the role of linguistic distance, in particular, and cultural distances, more broadly, in shaping economic outcomes. Linguistic similarity predicts greater trade between countries (Anderson and Van Wincoop, 2004; Egger and Lassmann, 2012; Hutchinson, 2005; Melitz and Toubal, 2014). More generally, linguistic, religious, and cultural distances across societies correlate with ancestral distance and predict a wide range of economic outcomes (Spolaore and Wacziarg,

2018). Within Indian economic history, social divisions of language, caste, and religion have been particularly salient. Industrial segregation was driven by information sharing within ethnolinguistic communities (Gupta, 2014). Caste and religious divisions, as well as the preferences of caste, ethnic, and religious elites contributed to reduced spending on schooling, which had effects that persisted until the 1970s (Chaudhary, 2009; Chaudhary and Garg, 2015; Chaudhary et al., 2012).

Second, we contribute to a literature on market integration and trade. Following on works such as Persson (1999) and Shiue and Keller (2007), several contributions in economic history have measured price integration across markets to compare levels of economic development across regions (Federico, 2011; O'Rourke and Williamson, 2002; Studer, 2008).[1] In the study of Indian economic history, Persaud (2019) has shown that price volatility mattered by spurring international migration. More generally, our work is related to a broader literature on the evolution of trade and market integration throughout history (Estevadeordal et al., 2003; Jacks et al., 2008; Pascali, 2017).

We also make a substantial data contribution, digitizing both detailed language data from the colonial census and price data spanning a wider set of markets and commodities (68,181 observations) than addressed by the work of Allen (2007), Andrabi and Kuehlwein (2010) or Studer (2008).

The most similar studies to ours, Falck et al. (2012) and Lameli et al. (2015), use dialect similarity within Germany to predict intra-regional trade and migration. Our work differs from these in several respects. Notably, the linguistic cleavages existing in India are greater than those between the often mutually-intelligible dialects of German. We consider possible roles of genetic distance[2] and transport investment. Finally, we provide evidence from a large and multilingual developing country, cover a longer time period, examine price integration as an outcome, and use a more spatially disaggregated unit of analysis.

2. Historical background

2.1. **Language in South Asia.**

There are four language families prominently represented in South Asia: Indo-European, Dravidian, Sino-Tibetan, and Austro-Asiatic (Asher, 2008). Prior to the arrival of Indo-European languages roughly 3500 years ago, the sub-continent was predominantly Dravidian-speaking (Asher, 2008).

Almost half the world's population speaks an Indo-European language descended from the protolanguage that originated at least 6,000 years ago in eastern Anatolia (Gamkrelidze and Ivanov, 1990). These spread throughout Europe and South Asia through both population

[1]Other studies have used historical price series to measure the responsiveness of prices and welfare to variables such as weather shocks and transportation infrastructure (Andrabi and Kuehlwein, 2010; Jia, 2014; Waldinger, 2014).

[2]See Giuliano et al. (2014) as an example for trade between countries.

movement and replacement of languages used by existing populations (Haak et al., 2015; Renfrew, 1989). Most speakers of Indo-European languages in South Asia speak Indo-Aryan languages such as Hindi and Bengali. Indo-Aryan languages date at least as far back as 100 BCE (Asher, 2008; Emeneau, 1956). The principal Dravidian languages became separated no later than 1000 CE, the main literary languages being Telugu, Kannada, Tamil, and Malayalam (Asher, 2008). Tamil cave inscriptions date to the second century BC, Malayalam inscriptions to the ninth century AD, Kannada inscriptions to 450 AD, and Telugu place names to the second century AD (Krishnamurti, 2003). Austro-Asiatic languages, divided primarily into the Mon-Khmer and Munda branches, predate the Indo-European languages in South Asia, and may have been present as long as the Dravidian languages (Asher, 2008). The small number of Sino-Tibetan speakers in South Asia speak primarily Tibeto-Burman languages (Asher, 2008).

Within India, the presence of multiple languages has been shaped by population movements and divergence of relatively isolated speakers (Asher, 2008). The rapid adoption of Indo-European languages suggests these had been adopted by the broader Dravidian-speaking community as a lingua franca (Krishnamurti, 2003), though the Dravidian boundary has been shifting southwards for a very long time, and Dravidian languages were largely absent from the Gangetic valley by 0AD (Emeneau, 1956). Languages in close proximity to each other have influenced each other (Montaut, 2005, p. 91). Malayalam uses several Sanskrit words, inflected words, and phrases (Krishnamurti, 2003). Indian languages borrow from each other through extensive bilingualism, and Indo-European and Dravidian languages have had grammatical impacts on each other (Emeneau, 1956; Krishnamurti, 2003). A particular feature of India is the durability of migrant languages, for example the continued use of Gujurati by communities that have lived in Tamil Nadu for several centuries (Montaut, 2005, p. 94).

2.2. **Markets in Colonial India.** The secondary literature on Indian history provides some information on how local prices of foodgrains were determined. Andrabi and Kuehlwein (2010) cite figures demonstrating that production was regionally concentrated, and that most food grains were largely consumed within India. For example, in 1919, the Punjab and the United Provinces accounted for 70 percent of the acreage devoted to growing wheat, while Bengal, Bihar, Orissa, and Madras accounted for 70 percent of the acreage devoted to growing rice. Only 5 percent of wheat and 7 percent of rice was exported beyond India in 1895. Exchange even within India was limited. The non-monetary sector of the economy was large (Kumar, 1983), even in 1950 (Chandavarkar, 1983).

At the start of our period, 1861, trade costs were high. Land transport was expensive and slow, with food grains largely hauled by oxen walking along dilapidated roads and carrying loads on their backs or in carts (Bhattacharya, 1983). In Western India, for example, where

few roads existed, trade relied on donkeys, camels and bullocks (Divekar, 1983). Intra-regional trade in low-value commodities was possible along rivers, but access to this trade was spatially limited (Derbyshire, 1987). Bullocks required a year to travel the distance that a railway would later cover in a week (McAlpin, 1974). Where a lack of roads made wheeled transportation difficult, caravans carried cotton and grain (Roy, 2012). Large-scale, long-distance shipments of grain were generally unprofitable (Hurd, 1975). The costs of overland transport limited market integration (Kessinger, 1983). Migration rates were low and wage convergence between districts over the nineteenth century was slow (Collins, 1999). Speed, cost, and seasonality constrained the geographical scope of the commercial orbit of the United Provinces (Derbyshire, 1987).

These costs fell during the 60-year time period of our analysis. The telegraph network spread through India in the 1850s and 1870s (Collins, 1999). Increasing commercialization benefitted from the replacement of the fragile military occupation with settled governance, a growing market for raw materials in Europe, and infrastructural improvements such as canal irrigation, metalled roads, and railway construction (Derbyshire, 1987; Kumar, 1983). The railways in particular reduced price dispersion across markets (Hurd, 1975), increased incomes (Donaldson, 2018), and reduced famines (Burgess and Donaldson, 2010); they are likely to have also increased price co-movement across districts. Price dispersion fell more rapidly for cash crops such as cotton than for food grains (McAlpin, 1974). Andrabi and Kuehlwein (2010) find evidence of trade in grain from districts that lacked railroads to neighboring districts with rail connections.

How did markets themselves work? Bhattacharya (1983) describes prototypical local market places in Eastern India in which farmers sold directly to consumers and middlemen in small quantities, and itinerant traders made small profits exploiting price differences within limited areas. Large farmers served as links between village markets and larger towns by buying grain from smaller farmers through credit contracts, holding stock while waiting for a favorable market, and taking grain to the mart or river mart offering the best price. Merchants' agents played a similar role. Larger towns gave rise to a stratified system of retail sellers, wholesale merchants, and those who bought from wholesalers and sold to retailers. Divekar (1983), Kumar (1983), and Kessinger (1983) provide similar descriptions for other regions of India in the first half of the nineteenth century.

Later in the century, commission agents and buyers' agents operated in towns that contained railway stations and banks (Roy, 2014). They owned capital such as carts, grain pits and warehouses. Commission agency and auction-type sales were prevalent. Company agents contracted with farmers in the villages, while landlords and others lent money to these farmers and were repaid in grain that they also sold to the commission and buyers' agents. In more remote areas, itinerant traders, including peasants, brought crops to bazaars. At

this time, forward trade seldom occurred. Europeans were largely absent from this trade, particularly from local transactions, though they were occasionally company agents and commission agents in railway towns. This helps explain why Europeans, sharing a common language, did not do more to drive market integration and may help explain our results.

Generally, prices in local markets correlated with fluctuations in the overall Indian money supply (Adams and West, 1979). Prices were typically lower in producing regions (Andrabi and Kuehlwein, 2010). On average, prices rose slowly through the 19th century and rapidly during the First World War (McAlpin, 1983).

2.3. **Language in Markets in Colonial India.** The languages used in trade varied from market to market, depending on which trading castes were dominant in each location. These are often described in the Imperial Gazetteers for each province.[3] In the Punjab, for example, the multilingual Banias, Khatris, the Aroras who spoke local languages such as Punjabi and Gujarati were dominant in different parts of the province. Predominantly Urdu-speaking Shaikhs and largely Gujarati-speaking Khojas were also important (p. 49). In wheat markets, cultivators themselves traded directly with exporters (p. 87). In Bengal, much of the trade was in the hands of Marwari Agarwals and Oswals, who might often speak local languages. Hindi-speaking Rauniars and Kalwars were more prominent in Bihar (p. 91). In Madras, the Tamil-speaking Chettis and Telugu-speaking Komatis controlled trade in the districts where these languages dominated. Traders themselves were, however, often multilingual, and changed the language used depending on the market. As Montaut (2005, p. 94), drawing on Pandit (1977), puts it:

> The classic example is of the Gujarati merchant one century ago, who uses Kacchi (a dialect of Gujarati) in the local market, Marathi for wider transactions in the region, standard Gujarati for readings, Hindustani when he travels (railway station), Urdu in the mosque, with some Persian and Arabic, but also *sant bhasha* in devotional songs, his variety of Gujarati for family interaction, English when dealing with officials.

3. Empirical strategy and Data

3.1. **Empirical strategy.** In this paper, we use price data covering M South Asian markets. Each observation is a market-pair, indexed ij. For product p, traded between markets i and j, we estimate:

$$\rho_{ij}^{p} = \beta^{p} LinguisticDistance_{ij} + x_{ij}^{p\prime}\gamma^{p} + \delta_{i}^{p} + \eta_{j}^{p} + \epsilon_{ij}^{p}. \tag{1}$$

[3]Imperial Gazetteer of India, Provincial Series, Vol 1. Bengal (1909), Madras (1908), and Punjab (1908). Superintendent of Government Printing.

In (1), ρ_{ij}^{p} is the correlation coefficient for the price of p between markets i and j. $LinguisticDistance_{ij}$, described below, captures linguistic distance between the two markets. x_{ij}^{p} is a vector of controls. We use this to account for a wide set of dissimilarities between i and j that may correlate with linguistic distance and with the degree of price integration. In our baseline estimations, x_{ij}^{p} includes a constant, as well as controls for *proximity* (log distance in kilometers between the markets, whether both markets are coastal, and whether both markets are connected by the same river), *geographic similarity* (the correlations in precipitation and temperature between the markets, and their absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, and terrain slope), *agricultural similarity* (absolute differences in suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, or tomato), *other measures of similarity* (whether the markets are in the same province, and their religious distance), and *characteristics of the data* (first year, last year, and number of years in which the price is available for both markets).

One limitation of our empirical strategy is the possibility that our control variables are measured with greater error than our principal right hand side variable of interest, i.e. $LinguisticDistance_{ij}$. This could lead to our estimates of β^{p} being overstated. We note, then, that linguistic distance may be interpreted more broadly, for example as a measure of greater ancestral distance. δ_{i}^{p} and η_{j}^{p} are fixed effects for market i and market j. The sample is all market pairs ij such that $i \neq j$, $i > j$, and there are sufficient observations to compute ρ_{ij}^{p}. That is, we have at most $\frac{M^2-M}{2}$ observations in any one regression. We cluster standard errors by both market i and market j in the baseline (Cameron et al., 2011). Because of the possible spatial dependence induced by forming every pairwise combination of markets, we show results in the online appendix in which we cluster at alternative levels and compute Conley (1999) standard errors.

3.2. **Data.** We use several sources of data. Below, we discuss our sources for prices in colonial India, for linguistic distance across markets, and for our additional controls.

3.2.1. *Prices.* Our data on prices are taken from three editions (1921, 1907, and 1885) of *Wages and Prices in India*. These are initially in reported in sers ($\sim$ 1.15 kg) per rupee: we invert this measure to obtain nominal prices. For 206 markets in modern-day Pakistan, India, Bangladesh, and Burma, these data provide prices for more than a dozen crops: Arhar Dal, Bajra, Barley, Gram, Jawar, Kangni, Maize, Marua, Rice, Salt, Wheat, Bulrush Millet and Similar, Great Millet and Similar, and Lesser Millets. The data covers both British India and the Princely States. These do not represent all markets in India – almost every populated place would have a market of some sort. Rather, these are markets in which the colonial government collected price data. More populous districts and districts in British

India are more likely to appear in the data, and, in provinces such as Coorg that have few districts, at least one district is likely to be present.

In most of our results, we focus on the three most commonly reported prices: rice, wheat, and salt. The data do not allow us to consider differences between different varieties of wheat or salt. However, we also show that estimates of (1) with several other crops produce similar estimates. The price data cover the period 1861 through 1921, with many markets entering our data for the first time in 1869. While the data-collection methods differed across markets in early years, from 1872 onwards uniform fortnightly returns of retail prices were used.[4] So long as there are at least three years in which a price is reported in both markets i and j, we can compute a correlation coefficient for that product for the ij pair. This quantity, ρ_{ij}^{p}, is our principal dependent variable.

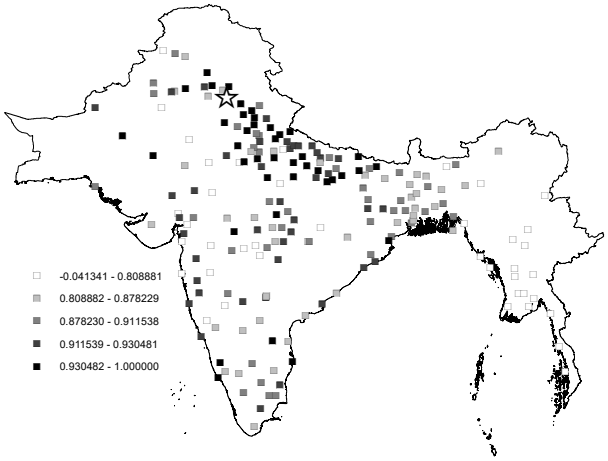
In Figure 1, we provide intuition for our results by mapping the correlation between the price of rice in a single market, the largely Punjabi-speaking city of Ludhiana, with the price of rice in all other markets in our data. It is clear from the figure that rice prices track those in Ludhiana more closely in regions that speak more closely-related languages such as Hindi and Gujarati and less closely in regions that speak more distantly-related languages such as Burmese and Telugu. These regions are, however, also closer in physical proximity to Ludhiana, and many of the markets that most closely track prices in Ludhiana lie on the Indo-Gangetic Plain. Thus, our analysis relies on estimation of (1) to demonstrate that the correlation between linguistic distance and price integration cannot be explained away by other observable differences in proximity or geography.

3.2.2. *Linguistic distance.* To compute linguistic distances between the markets in our data, we use two additional data sources. These are the 1901 Census of India and version 19 of the *Ethnologue* Global Dataset. For each district that existed in 1901, the census data report the number of speakers of each language. For example, the three most commonly spoken languages reported for Ludhiana District are "Punjabi" (665,476), "Hindostani" (2,970), and "Kashmiri" (1,224). We assign each market to the language composition of the district that contained it in 1901. For consistency with the *Ethnologue* data on distances, we aggregate these to the level of ISO language codes. For Ludhiana, the three most commonly spoken languages become *pan*, *hin*, and *kas*. The data do not, unfortunately, mention second languages.

To compute the distances between these languages, we turn to *Ethnologue*. Every language in this source is categorized using a language tree with a maximum number of 15 branches. These classifications are based on several sources, the most important of which is Frawley

[4]We show below that results are similar when we use only the period after 1891 (the midpoint of the price data) to compute our dependent variable: see section A.1. We are not worried, then, that differences in how data were collected before and after 1872 drive our results.

FIGURE 1. Ludhiana: Rice price correlations



(2003). Such "cladistic" measures have become widely used in economics (Desmet et al., 2012; Gomes, 2014).[5]

Following Esteban et al. (2012), we take the distance d_{mn} between any two languages m and n as:

$$d_{mn} = 1 - \Big(\frac{SharedBranches}{15}\Big)^{\delta}. \tag{2}$$

Similarly following Esteban et al. (2012), we choose $\delta = 0.05$ as a baseline and use $\delta = 0.5$ for robustness. To aggregate these to distances between markets, given population shares of languages m and n in each district i and j of s_{mi} and s_{nj}, we follow Spolaore and Wacziarg (2009) and compute linguistic distance between districts as:

$$LD_{ij} = \sum_{m}\sum_{n}(s_{mi} \times s_{nj} \times d_{mn}). \tag{3}$$

In Figure 2, we map the linguistic distances between every district in our data and Ludhiana. While it is evident that the markets at which languages more closely related to Punjabi are spoken are geographically close to Ludhiana, it is also clear that this correlation of linguistic and physical distance is not perfect. Distances change relatively rapidly over space when the linguistic composition of the population similarly changes rapidly. Further,

[5]Although alternative distance measures exist based on phonetic similarity of languages (Dickens, 2018), these would be measured with considerable error in our data, given the large number of languages in our data for which the phonetic word lists of the Automated Similarity Judgment Program are either missing or incomplete. (We do, however, report results using these as an alternative measure in section A.2).[6] Under this classification system, for example, Punjabi is coded as Indo-European, Indo-Iranian, Indo-Aryan, Intermediate Divisions, Western, Panjabi.

regions that are relatively similar in physical distance can be quite dissimilar in their linguistic distance. Punjabi and Bengali, for example, both share the branches Indo-European, Indo-Iranian, and Indo-Aryan. Punjabi and Tamil, by contrast, share no branches, as Tamil is a Dravidian language. And yet the distance between the Punjab and Bangladesh is not markedly different than the distance between the Punjab and Tamil Nadu. The log distance in kilometers between Ludhiana and Dacca is 7.40, whereas it is 7.76 between Ludhiana and Madurai.

3.2.3. *Additional controls.* Some of our control variables are computed directly. Distance in kilometers is computed using the latitude and longitude of the market. "Both coastal" and "both connected by the same river" indicators are computed in ArcMap using a shapefile of district boundaries. "Minimum year," "maximum year," and "number of common observations" are computed directly from the price data.

The "same province" indicator is based on the provinces that contained each market in 1901. The "religious distance" variable is computed using the same equation as (3), taking the religious composition of each district as reported in Table 8 of the 1921 Census (Literacy By Religion). We assume that the distance d_{qr} between any religion q and r is 1 if $q \neq r$ and 0 if $q = r$.[7]

Data on land quality are taken from Ramankutty et al. (2002) and have been used in several economic studies, such as Michalopoulos (2012) and Ashraf and Galor (2011).[8] It is an index based on soil and climate characteristics and is not particular to any one type of agriculture. "Ruggedness" is the measure of terrain ruggedness initially introduced by Nunn and Puga (2012).[9] Our measure of "malaria prevalence" was originally created by Kiszewski et al. (2004).[10] Altitude data are taken from the Consultative Group for International Agricultural Research's Shuttle Radar Topography Mission 30 dataset.[11] Means of precipitation, temperature, and suitabilities for specific crops are taken from the Food and Agriculture Organization's Global Agro-Ecological Zones data portal.[12] Similar suitability measures have been used by Alesina et al. (2013) and Alsan (2015). Correlations in rainfall are computed using the Matsuura and Willmott (2007) gridded series.[13] We join each market to the nearest point in these data and compute correlations in annual rainfall over the period

[7]If, as an alternative, we collapse Islam, Judaism, and Christianity into a single category, results are numerically indistinguishable because of the negligible share of Jews and Christians in the population. We omit these results for space.

[8]`https://nelson.wisc.edu/sage/data-and-models/atlas/maps.php?datasetid=19&includerelatedlinks=1&dataset=19`

[9]`http://diegopuga.org/data/rugged/tri.zip`

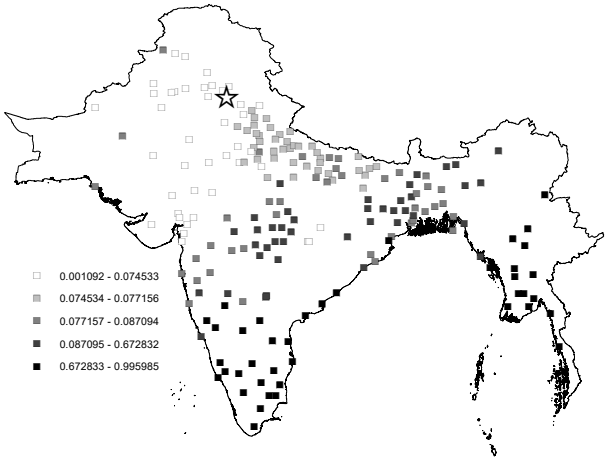
[10]We are grateful to Marcella Alsan for providing us with these data.

[11]`http://www.diva-gis.org/gdata`

[12]`http://www.fao.org/nr/gaez/en/`.

[13]`http://climate.geog.udel.edu/climate`

FIGURE 2. Ludhiana: Linguistic distances



1900-2000. Humidity data are taken from the Climatic Research Unit at the University of East Anglia.[14]

Like many studies that control for geographic confounders with historical outcome variables, we are compelled to use present-day raster data (e.g. Alsan (2015); Nunn and Puga (2012)). We expect that this will add measurement error to our right-hand-side variables, but that it is unlikely this measurement error will induce spurious correlation between linguistic distance and market integration. For the variables that require geographic data (that is, the coastal and river indicators, as well as those using raster data), we begin with a district map for modern India.[15] We compute the coastal and river indicators at this level, and compute other geographic variables by averaging over raster points within a district. If a market in our data shares the name of a modern-day district (or an updated name, as in the case of Benares and Varanasi), we have a unique match between the market and the modern district polygon. Otherwise, we match all districts that split from the erstwhile district that previously shared the name of the market to that market.

3.3. **Summary statistics.** Summary statistics are presented in Table 1. Some general patterns are apparent from this table. First, relative to a maximum number of observations of $\frac{206^2-206}{2} = 21,115$, we typically have fewer pairwise correlation coefficients. This is because not all products are traded in all markets. Second, while the degree of price integration is relatively high (> 0.8 for both wheat and rice), there is variation in price integration both across space and across markets. Some market pairs exhibit negative price correlations. Market integration is more limited for salt than for rice and wheat; the average price correlation for salt (< 0.35) is lower, and more than a quarter of these correlations are negative.

[14] `https://crudata.uea.ac.uk/cru/data/hrg/tmc/grid_10min_reh.dat.gz`
[15] In particular, we use the boundaries reported by `www.gadm.org`.

One possible explanation of this lower correlation is the limited number of inland production sites for salt; this limits arbitrage opportunities in response to shocks, causing low average salt price correlations across markets. Linguistic distances range from close to 0 (i.e., market pairs in which both markets are dominated by the same language) to 1 (i.e., market pairs in which the dominant languages spoken are unrelated).

4. Results

4.1. **Results by market.** Before presenting estimates of (1), we present preliminary descriptive evidence. For each market i in our data, we estimate:

$$\rho_{ij}^{p} = \beta_{i}^{p} LinguisticDistance_{ij} + x_{ij}^{p}{}'\gamma^{p} + \epsilon_{ij}^{p}. \tag{4}$$

In (4), ρ_{ij}^{p} and x_{ij}^{p} are defined as in (1). For each market i, we obtain a coefficient β_{i}^{p} that captures the degree to which its prices more closely track prices at other markets that are more linguistically similar, conditional on other measures of distance and dissimilarity.

To present these results, we order markets from those with the most negative estimates of β_{i}^{p} to those with the most positive estimates and present the point estimates and 95 percent confidence intervals in figures 3, 4, and 5. For each of the three major crops, the majority of coefficients is negative and significant. This demonstrates two points. First, our main results pooling together all market pairs are not driven by a small number of markets. Second, (1) yields estimates of β^{p} that capture a central tendency in the sample.

4.2. **Main results.** In Table 2, we present our main estimates of (1). Across the three major crops, linguistic distance predicts reduced market integration. This is statistically significant in all specifications save one: wheat with controls but without fixed effects. There are several ways to consider the magnitudes involved. First, taking the estimates from column (4), a one standard deviation increase in linguistic distance, conditional on controls and fixed effects, predicts a reduction in the price correlation between markets i and j by 0.121 standard deviations for wheat, 0.181 standard deviations for salt, and 0.088 standard deviations for rice.

It is striking that the coefficients and standardized magnitudes are largest for salt. Not only are salt markets less integrated in the data, in that they have lower mean correlation coefficients, there is also more dispersion in integration for salt, in that the standard deviation of the correlation coefficients across market pairs is larger. Salt was a differentiated good that could only be produced in a small number of locations (Donaldson, 2018). Further, in order to facilitate the taxation of salt, the British constructed an Inland Customs Line, which incorporated the Great Hedge of India, in order to prevent salt smuggling (Moxham, 2001).

TABLE 1. Summary statistics

	(1) Mean	(2) s.d.	(3) Min	(4) Max	(5) N
Correlation: Wheat	0.81	0.22	-1	1	15,652
Correlation: Salt	0.54	0.41	-0.78	1	20,909
Correlation: Rice	0.81	0.16	-0.25	1	20,909
Linguistic Distance	0.42	0.39	0.000061	1.00	21,115
Genetic Distance	0.0026	0.0016	1.8e-07	0.010	21,115
Ln Distance in KM	6.85	0.71	1.99	8.24	21,115
Same Province	0.11	0.32	0	1	21,115

FIGURE 3. Results by market: Wheat

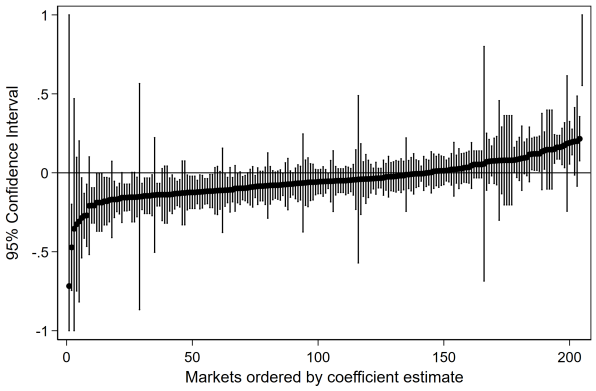
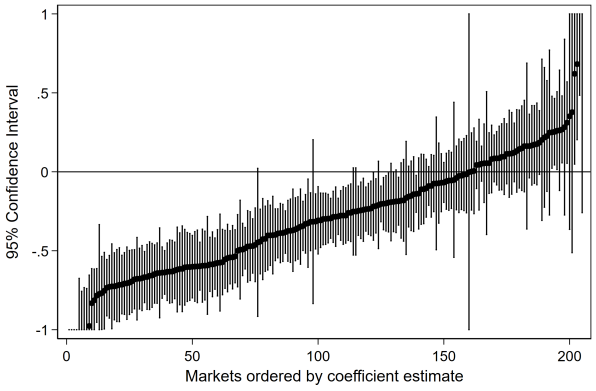
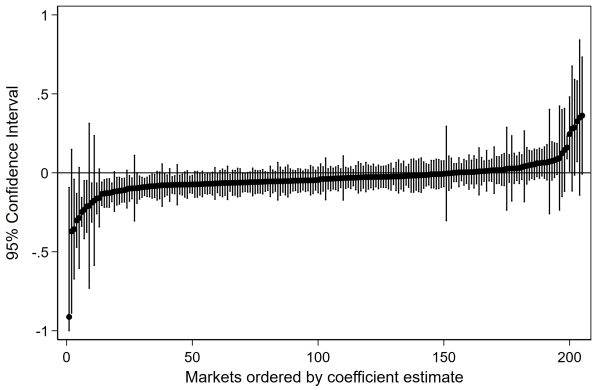


FIGURE 4. Results by market: Salt



An alternative approach to magnitudes is to divide $\hat{\beta}^p$ by the coefficient estimated on $\ln(Distance)$ in column (4). This suggests that moving one unit in linguistic distance (i.e. from a closely-related language to an unrelated one) predicts a reduction in the price correlation comparable to a distance change of 789 percent for wheat, 1,328 percent for salt, and 210 percent for rice. At the mean distance across pairs within our sample (1154 kilometers),

FIGURE 5. Results by market: Rice



this would correspond to distance increases of 9,101, 15,326, and 2,418 kilometers, respectively, all of which would be out of sample. These large numbers are driven in part by the small coefficients estimated on distance once additional controls are included.

In appendix table A4, we compare the pairwise correlations between our outcome variables and the measures of physical and linguistic distance. Both distance measures enter significantly and negatively on their own and, if both are put on the right hand side at once, both continue to enter negatively and significantly, while the coefficient on each is reduced slightly. Both have similar R-squared values when included as right hand side variables alone, and including both on the right hand side increases the R-squared.

5. MECHANISMS

In this section, we outline the mechanisms suggested in both the economic and historical literatures that provide plausible links between linguistic distance and market integration. We then assess these empirically to the extent our data allow.

5.1. **Mechanisms in the literature.** A recent economic literature has emphasized several possible channels that might link linguistic distance to market outcomes, and several of these mechanisms are reflected in observations made about colonial Indian markets in the secondary historical literature. One branch of this economic literature has focused on the importance of barriers to the transmission of the traits that are imparted across generations in driving dissimilarities in economic outcomes across populations (Spolaore and Wacziarg, 2009, 2018). Alternatively, differences in language may proxy for differences in tastes, which in turn shape prices and the volume of trade (Atkin, 2013, 2016). Where these taste-based differences lead to a thin local market for a given good, we might anticipate prices that do not track those in other South Asian markets. Similarly, if there are fixed costs of arbitrage

TABLE 2. Main results

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.257***	-0.210***	-0.023	-0.067**
	(0.035)	(0.036)	(0.025)	(0.030)
N	15,652	15,652	15,652	15,652
Rsq	0.139	0.762	0.580	0.806
		Correlation: Salt		
Linguistic Distance	-0.484***	-0.392***	-0.384***	-0.189***
	(0.061)	(0.072)	(0.051)	(0.044)
N	20,909	20,909	20,909	20,909
Rsq	0.216	0.708	0.566	0.791
		Correlation: Rice		
Linguistic Distance	-0.083***	-0.073***	-0.056***	-0.035***
	(0.017)	(0.010)	(0.018)	(0.010)
N	20,909	20,909	20,909	20,909
Rsq	0.045	0.834	0.282	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

between two markets, the limited size of the market for an unpopular product will reduce the returns to arbitrage.

Another branch of the economic literature suggests mechanisms by which language barriers may inhibit market integration by raising trade costs. For example, linguistic distance may affect the costs of acquiring information (Allen, 2014; Gomes, 2014). Alternatively, linguistic distance may act as a barrier to flows of people, who are likely to be put off by migration costs, the difficulty of establishing business connections, or by xenophobia (Bai and Kung, 2017; Falck et al., 2012; Iwanowsky, 2017; Lameli et al., 2015; Rauch and Trindade, 2002). These mechanisms would lead to missing or costly links in the network connecting any two markets.

This branch of the economics literature aligns most closely with descriptions of trade in the secondary literature on Indian history. Collins (1999) cites linguistic barriers as an explanation of the low migration rates in India and hence as a limiting factor on price integration. Several writers have highlighted the importance of trade networks that corresponded with linguistic divisions. In colonial India, trading networks were often caste or kinship networks (Bhattacharya, 1983; Kessinger, 1983). Markovits (2008, p.188-196) mentions several such

"middlemen minorities."[16] These groups, Divekar (1983) argues, contributed to the "unification of markets in India." They adopted new forms of business partnership and circulated information over wide regions. If the costs of one group maintaining a presence in a given market due to its linguistic dissimilarity are greater, this would be expected to increase transactions costs with other markets in which they are present.

Linguistic distance may also make it more difficult to acquire a language in which trade is conducted or to acquire common levels of education; Isphording and Otten (2014), Jain (2017), Laitin and Ramachandran (2016), and Shastry (2012) all find evidence that the costs of acquiring a new language – or education provided in that new language – are higher for those whose mother tongue is more dissimilar to the new language. Finally, linguistic distance may proxy for differences in preferences over public goods, redistribution, and the provision of infrastructure (Desmet et al., 2020, 2012, 2017). If these public goods and infrastructure investments affect trade costs, they may help explain our main result.

5.2. **Mechanisms: Evidence.**

5.2.1. *Genetic distance.* To evaluate whether linguistic distance operates as a proxy for a broader set of barriers to the transmission of information, technology, and culture, we compute a measure of the genetic distance between the markets in our data. We show that, while linguistic distance and genetic distance are correlated, neither one is a "sufficient statistic" that fully accounts for the coefficient on the other.

We obtain data on genetic distance from Pemberton et al. (2013). Similar to the data used by Spolaore and Wacziarg (2009), these data contain pairwise Weir and Cockerham (1984) F_{ST} coefficients based on differences in allele frequencies from microsatellites. While the raw data report coefficients based on 5795 individuals from 267 human populations, we restrict ourselves to the data on ethnic groups indigenous to South Asia. These are the Balochi, Brahui, Burusho, Hazara, Kalash, Makrani, Pathan, Sindhi, Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Marathi, Marwari, Miso, Oriya, Parsi, Punjabi, Tamil, and Telugu. While these groups cover the majority of the population in our sample, there are some major missing groups, of which Urdu is the largest.

Following Spolaore and Wacziarg (2009), given population shares of groups m and n in districts i and j of s_{mi} and s_{nj} with genetic distance F_{ST}^{mn}, we compute genetic distance between districts as:

[16]His list includes the Marwaris, Gujaratis, Parsis, Sindhis, Chettiars, Khatris, Aroras, Multanis, Bhatias, Khojas, Lohanas, Bohras, Memons, Banias, Pathans, Vanis, Shravaks, Agarwals, Maheshwaris, Oswals, Khandelwals, and Porwals. Roy (2014) similarly discusses the role of Marwaris, Banias, Parsis and Khojas. Divekar (1983) adds to this the Afghans, Voras, Lingayat Banjigs, Komtis, and Vanjaris. Kumar (1983) and McAlpin (1974), similarly, highlight the role of the Banjaras.

$$GD_{ij} = \sum_m \sum_n (s_{mi} \times s_{nj} \times F_{ST}^{mn}). \tag{5}$$

Note that we re-scale s_{1i} and s_{2j} as fractions of the population matched to the genetic data, rather than as fractions of the full district population. We present a map of genetic distances from Ludhiana in appendix Figure A1. This has many similarities to Figure 2. Other regions of South Asia that are proximate to the Punjab are more genetically similar, though it is clear that South Indian groups in Dravidian-speaking regions are more genetically dissimilar, conditional on physical distance. The apparent proximity with Burma is overstated due to the lack of coverage of major Burmese populations in the genetic data.

Our aim is to assess whether linguistic distance proxies for broader (and possibly deeper) barriers to the diffusion of information, culture, and technology. We re-estimate (1), first with genetic distance as an outcome, and second with genetic distance as an additional control. We report results in Table 3. Linguistic and genetic distance are correlated, even conditional on our baseline fixed effects and controls.[17] Genetic distance itself predicts less market integration and diminishes the coefficient on linguistic distance, but does not fully eliminate it in any specifications where linguistic distance was significant in Table 2. With fixed effects and controls, the change in coefficient on linguistic distance is slight when compared with Table 2. These results imply that, while linguistic distance may indeed proxy for other differences across populations, its relationship with market integration cannot be fully accounted for by the additional transactions costs imposed by barriers to the diffusion of beliefs, traditions, and practices stemming from ancestral distance.

5.2.2. *Coarse and fine distinctions.* We show that it is the highest-level distinctions in our data, such as those between Indo-European and Dravidian languages, that drive our results. This is, however, a crude proxy, and we cannot rule out the possibility that languages here proxy for past patterns of migration and state formation that themselves shaped markets and trade routes.

Recall that, in our baseline analyses, we computed the distance between any two languages i and j as:

$$d_{ij} = 1 - \Big(\frac{SharedBranches}{15}\Big)^{\delta}$$

While this follows the convention in the literature, it does not allow us to distinguish whether coarser distinctions (e.g., those between Indo-European and Dravidian languages) or lesser divisions (e.g, those between Bengali and Punjabi) drive our results. We replace

[17]In the sample of pairwise comparisons between the 24 ethnic groups in Pemberton et al. (2013), avoiding duplicates and self-comparisons by keeping only ij pairs where $i < j$, the correlation between genetic and linguistic distance is positive but small, with $\rho = 0.1216$.

TABLE 3. Genetic Distance

	(1)	(2)	(3)	(4)
		Genetic Distance X 100		
Linguistic Distance	0.046***	0.105***	0.041**	0.027**
	(0.014)	(0.012)	(0.020)	(0.013)
N	21,115	21,115	21,115	21,115
Rsq	0.012	0.857	0.360	0.895
		Correlation: Wheat		
Linguistic Distance	-0.253***	-0.159***	-0.021	-0.062**
	(0.036)	(0.035)	(0.025)	(0.030)
Genetic Distance X 100	-0.063*	-0.283***	-0.036	-0.058**
	(0.036)	(0.050)	(0.025)	(0.026)
N	15,652	15,652	15,652	15,652
Rsq	0.142	0.769	0.580	0.806
		Correlation: Salt		
Linguistic Distance	-0.465***	-0.367***	-0.371***	-0.194***
	(0.064)	(0.079)	(0.052)	(0.043)
Genetic Distance X 100	-0.415***	-0.234**	-0.287***	0.195**
	(0.126)	(0.096)	(0.100)	(0.081)
N	20,909	20,909	20,909	20,909
Rsq	0.242	0.710	0.574	0.792
		Correlation: Rice		
Linguistic Distance	-0.076***	-0.057***	-0.051***	-0.034***
	(0.019)	(0.012)	(0.020)	(0.010)
Genetic Distance X 100	-0.167***	-0.154***	-0.113*	-0.034*
	(0.064)	(0.030)	(0.064)	(0.018)
N	20,909	20,909	20,909	20,909
Rsq	0.074	0.838	0.291	0.869
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

d_{ij} with a dummy for having $\leq N$ shared branches, for $N = \{1, ..., 15\}$. We re-estimate (1), and present our results in figures 6, 7, and 8. These correspond to column (4) with fixed effects and controls. In all three figures, it is clear that coarser distinctions matter more than finer ones. Indeed, we show in Appendix Table A5 that limiting our sample only to district pairs in which the dominant language in both districts is Indo-European leads to coefficient estimates on linguistic distance that, while still negative, are generally insignificant and less robust across specifications. That is: our results are driven by coarser language distinctions, particularly those that separate major language families.

Consider a language such as Gujarati (Indo-European, Indo-Iranian, Indo-Aryan, Intermediate Divisions, Gujarati, Gujarati). It has no branches in common with a Dravidian language such as Tamil. It shares one branch with languages such as Yiddish that are Indo-European but not Indo-Iranian. It shares two branches with languages such as Balochi that are Indo-Iranian but not Indo-Aryan. It shares three branches with an Indo-Aryan language such as Hindi that is classified under "Western Hindi" rather than "Intermediate Divisions."

FIGURE 6. Results by level: Wheat

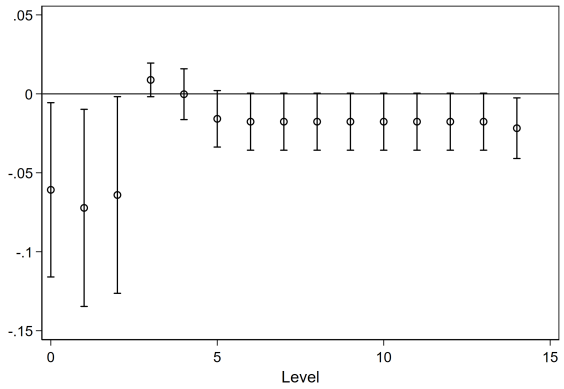


FIGURE 7. Results by level: Salt

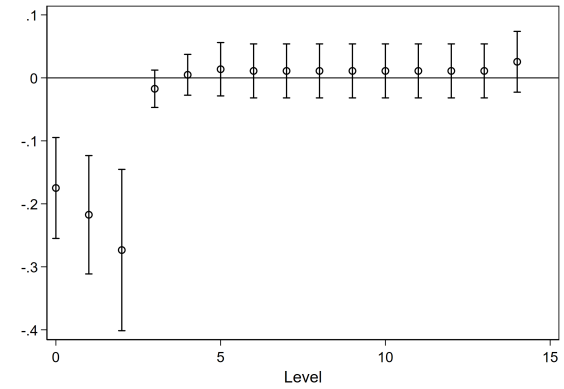
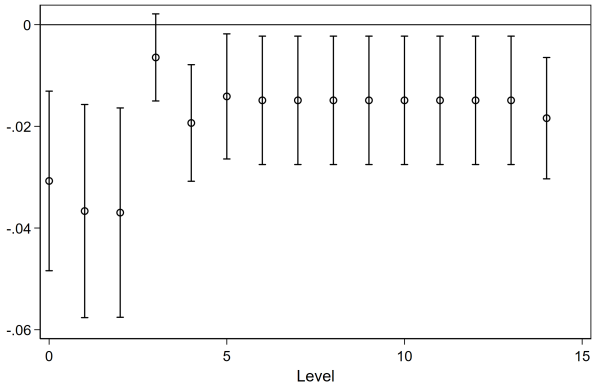


FIGURE 8. Results by level: Rice



It shares four branches with a language such as Nepali that is within these "Intermediate Divisions," but is not within the Gujarati sub-class. It shares five branches with other

Gujarati languages (such as Jandavra). In all three figures, language divisions with two common branches or fewer yield visibly greater differences than finer distinctions. These results suggest that our main results derive from divisions on the scale of Gujarati-Tamil, Gujarati-Yiddish, and Gujarati-Balochi, rather than from finer distinctions as those between Gujarati and Hindi, Nepali, or Jandavra. These coarser distinctions are those that have been shown before to correlate with conflict, redistribution and public goods provision – suggesting they are correlated with deeper differences in preferences – as opposed to finer distinctions that inhibit coordination and integration (Desmet et al., 2012). This is suggestive evidence that our results are driven not simply by ease of communication, but also by more fundamental differences in preferences.

5.2.3. *Missing markets.* To test whether missing markets, due, for example, to differences in tastes drive the correlation between linguistic distance and market integration, we evaluate whether linguistic distance predicts whether two given markets report a certain good's price in the same year, and whether markets that are more linguistically distant from their neighbors experience more volatile prices. When we look at the situation for major crops, we find little evidence of missing markets increasing with linguistic distance. Only limited evidence suggests that prices are more variable at markets that are more linguistically different from those around them.

We take two approaches. First, we test whether linguistic distance predicts how frequently prices are available for two markets in the same year. Taking N_{ij}^p as the number of common price observations at markets i and j for product p, we estimate (1), except that we now take N_{ij}^p as the dependent variable, and no longer control for minimum year, maximum year or the number of common observations. Results are presented in Table 4. There is only weak evidence of missing markets correlating with genetic distance; while we find a negative correlation between linguistic distance and N_{ij}^p for wheat, no such correlation is available for salt or rice. We find similar failures of linguistic distance to predict N_{ij}^p when using lesser crops from the data such as barley and maize, although we do not report these here. One explanation of the different result for wheat is the greater variability of the outcome variable: the standard deviation of the number of common years for wheat is 22.6, versus 8.8 for salt and 9.7 for rice. That is, as wheat is reported less often in many markets, there is more variation to be explained.

As a second approach, we evaluate whether markets that are more linguistically distant than those within a set radius experience prices that are more volatile. Our logic here is that linguistic distance from neighbours may lead to more volatile prices because of reduced trade and arbitrage. For each market i, we keep the other markets within 500 kilometers and take the average of their linguistic distance from i (denoted $\bar{Linguistic Distance}_{ij}$) as well as the average of the controls (denoted $\bar{x}_{ij}^p$). We estimate:

TABLE 4. Missing markets: Number of common years

	(1)	(2)	(3)	(4)
		Observations: Wheat		
Linguistic Distance	-37.518***	-15.483***	-36.672***	-13.412***
	(2.450)	(2.325)	(3.045)	(2.183)
N	21,115	21,115	21,115	21,115
Rsq	0.429	0.928	0.562	0.936
		Observations: Salt		
Linguistic Distance	-1.304	0.004	-3.279*	-0.017
	(1.278)	(0.071)	(1.868)	(0.157)
N	21,115	21,115	21,115	21,115
Rsq	0.003	0.954	0.212	0.954
		Observations: Rice		
Linguistic Distance	-1.441	0.011	-3.126	-0.097
	(1.316)	(0.085)	(1.938)	(0.165)
N	21,115	21,115	21,115	21,115
Rsq	0.003	0.954	0.205	0.955
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

$$CV_i^p = \beta^p \bar{Linguistic}Distance_{ij} + \bar{x}_{ij}^{p\prime}\gamma^p + \epsilon_i^p. \tag{6}$$

In (6), CV_i^p is the coefficient of variation of the price of product p at market i. We estimate (6) by OLS and report robust standard errors. Results are presented in Table 5. While we find evidence that wheat prices are more volatile at markets that are more linguistically distant from others in their neighborhood, we find no similar evidence for rice or salt. The differences by crop here are somewhat puzzling, as it is rice prices that are most volatile in our data, as measured by the coefficient of variation.

5.2.4. *Trading communities.* To evaluate whether the presence of trading networks sharing a common tongue drives our results (as might be the case if, for example, small communities of traders have lower costs of establishing themselves in regions where the dominant language resembles their own), we correlate linguistic distance with the common presence of communities such as the Marwaris or Parsis. We find little evidence that the co-presence of these communities correlates with linguistic distance.

We focus on one group that has received particular attention in the literature: the Marwaris. By 1920, between 200,000 and 400,000 Marwaris, most of them working as traders, lived outside of the Rajputana Agency (Markovits, 2008). These traders drew on capital and

TABLE 5. Missing markets: Volatility

	(1)	(2)	(3)
	CV: Whe	*CV: Sal*	*CV: Ric*
Linguistic Distance	0.127***	0.030	-0.113
	(0.049)	(0.049)	(0.279)
N	178	205	205
Rsq	0.528	0.400	0.121

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Robust standard errors in parentheses. All regressions are OLS and include a constant. Controls are averages of minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. .

personnel from throughout the subcontinent. They gained dominant positions in regional trade, importing, exporting and moneylending. These communities held assets jointly in patrilineal extended families, sharing information and personnel (Roy, 2014).

For each pair of markets i and j, we estimate the absolute difference in Marwari share, or $AD_{ij}^{Marwari} = |s_i^{Marwari} - s_j^{Marwari}|$. We then estimate (1) with $AD_{ij}^{Marwari}$ as both an outcome and as a control. That is: we test whether linguistic distance predicts the co-location of Marwaris across district pairs, and the degree to which the co-presence of this trading community can account for the conditional correlation between lingusitic distance and market integration. Results are presented in Table 6. There is little evidence of linguistic distance driving differences in the presence of this trading community, and little evidence that it explains price integration.

Results are similar if we perform the same exercise for the other communities listed in section 5.1, though we do not report these for space. While we cannot observe all these communities in our data, several are recorded in the census either as linguistic or religious groups. In particular, we are able to observe the Parsis, Afghanis, Gujaratis, Khatris, Memons, Multanis, and Sindhis. We also observe the Vanis, but they are not present in the markets in our data. Since the English could also be potentially thought of as another migrant mercantile community, we also consider their presence. Results are again similar, and again not reported, using the English. Our results are particularly unlikely to be explained by the spread of the English language: less than one tenth of one percent of the population in the 1901 census is recorded as "English" by language.

Alternatively, if we replace the absolute difference in the population share of a minority group with the maximum for a market pair, results are very similar. Because a group is often present in one market and not another, the maximum across a pair is highly correlated with the absolute difference in shares. Similarly, we find little correlation between linguistic distance and the minimum presence of a trading community across a market pair, and our

results are not generally sensitive to controlling for this minimum. Again, we omit these results for space.

5.2.5. *Literacy.* In a related test for the costs of information, we examine whether linguistic distance correlates with differences in literacy rates. While linguistically distant markets have more dissimilar literacy rates, this does not diminish the correlation of linguistic distance with market integration.

For data on literacy, we use the 1921 Census of India. These data report literacy at the district level, and we match each market to the district that contains it. As with the presence of trading communities, for each community, we take this difference as both an outcome and as a control. We present results in Table 7. More linguistically distant markets have more dissimilar literacy rates, but this does little to predict price correlations, or to explain away their correlation with linguistic distance.

5.2.6. *Infrastructure.* Finally, we examine whether linguistic distance proxies for shared preferences over public goods, in particular, those that facilitate trade. We show that more linguistically distant markets spend less time both connected to the railway network, but, nonetheless, this does not fully account for our main result.

Following a procedure similar to Donaldson (2018), we use the 1934 edition of *History of Indian Railways Constructed and In Progress* to identify the year each market became connected to the colonial railway. This source divides the Indian railway system into segments (e.g. "Karimganj to Badarpur") with a date of opening (in this example, 4-12-96) and length in miles (in this example, 12.00). We use these data to code the first date at which the district containing each market was connected to the Indian Railway system. For each market pair ij, we can then identify the number of years up to 1921 that both markets were connected to the railway system. We then estimate (1) with this variable as both an outcome and as a control. We present results in Table 8. More linguistically distant markets spend more time both connected to the railroad; however this does little to predict price correlations or explain away their correlation with linguistic distance. One possible contributing factor to these results is the nature of the Indian railways, which were often built to track pre-existing trade routes (Andrabi and Kuehlwein, 2010).

6. Robustness

6.1. **Selection on unobservables.** In this section, we demonstrate the robustness of our results to selection on unobservables. We present a number of additional exercises in the online appendix.

To demonstrate robustness to selection on unobservables, we use the approach of Altonji et al. (2005) as implemented by Bellows and Miguel (2009) and Nunn and Wantchekon

TABLE 6. Trading communities

	(1)	(2)	(3)	(4)
	Absolute difference in Marwaris share			
Linguistic Distance	-0.025**	0.001	0.055**	-0.001
	(0.012)	(0.001)	(0.024)	(0.001)
N	21,115	21,115	21,115	21,115
Rsq	0.004	0.984	0.263	0.984
	Correlation: Wheat			
Linguistic Distance	-0.255***	-0.210***	-0.023	-0.067**
	(0.035)	(0.036)	(0.025)	(0.030)
Abs. diff. in Marwaris share	0.066***	0.021*	-0.003	0.030*
	(0.014)	(0.011)	(0.016)	(0.017)
N	15,652	15,652	15,652	15,652
Rsq	0.142	0.762	0.580	0.806
	Correlation: Salt			
Linguistic Distance	-0.498***	-0.391***	-0.360***	-0.189***
	(0.060)	(0.072)	(0.051)	(0.044)
Abs. diff. in Marwaris share	-0.571***	-0.274*	-0.425***	-0.174
	(0.068)	(0.152)	(0.083)	(0.149)
N	20,909	20,909	20,909	20,909
Rsq	0.260	0.709	0.584	0.791
	Correlation: Rice			
Linguistic Distance	-0.085***	-0.073***	-0.054***	-0.035***
	(0.017)	(0.010)	(0.017)	(0.010)
Abs. diff. in Marwaris share	-0.074	-0.040***	-0.028	0.015
	(0.065)	(0.012)	(0.073)	(0.013)
N	20,909	20,909	20,909	20,909
Rsq	0.050	0.834	0.283	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

(2011). We estimate (1) with either a limited set of controls or with a full set of controls, and compute:

$$AET = \frac{\beta^{FullControls}}{\beta^{RestrictedControls} - \beta^{FullControls}} \tag{7}$$

We report results where the restricted set of controls is either empty or contains only ln($Distance$). Larger values of this statistic imply that the selection on unobservables would need to have a larger effect on β relative to that of observables in order to be consistent with a true β of 0. Results are presented in Table 9. The coefficient estimates for wheat are sensitive to controls regardless of what is in the base set of controls, but are not as sensitive to the addition of fixed effects. Results for salt and rice appear sensitive to adding fixed effects and controls together, but this is driven by ln($Distance$). Once this is included as a baseline control, AET is negative (i.e., controls push β away from zero) or greater than one.

TABLE 7. Literacy Rate

	(1)	(2)	(3)	(4)
		Difference in Literacy 1921		
Linguistic Distance	10.432***	6.920***	6.826***	4.691***
	(1.505)	(1.869)	(1.086)	(1.264)
N	20,503	20,503	20,503	20,503
Rsq	0.193	0.808	0.504	0.837
		Correlation: Wheat		
Linguistic Distance	-0.247***	-0.206***	-0.018	-0.067**
	(0.034)	(0.035)	(0.025)	(0.030)
Difference in Literacy 1921	-0.001	-0.001	-0.000	0.000
	(0.001)	(0.001)	(0.000)	(0.001)
N	15,125	15,125	15,125	15,125
Rsq	0.139	0.761	0.579	0.805
		Correlation: Salt		
Linguistic Distance	-0.339***	-0.323***	-0.327***	-0.164***
	(0.052)	(0.054)	(0.047)	(0.040)
Difference in Literacy 1921	-0.016***	-0.012***	-0.008***	-0.006***
	(0.002)	(0.003)	(0.002)	(0.002)
N	20,300	20,300	20,300	20,300
Rsq	0.343	0.732	0.589	0.800
		Correlation: Rice		
Linguistic Distance	-0.019	-0.064***	-0.017	-0.032***
	(0.025)	(0.008)	(0.025)	(0.010)
Difference in Literacy 1921	-0.006***	-0.001***	-0.006**	-0.001
	(0.002)	(0.000)	(0.002)	(0.001)
N	20,300	20,300	20,300	20,300
Rsq	0.155	0.836	0.347	0.869
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

That is, we find that the estimate of β is sensitive to controls for wheat, while for salt and rice, the estimate of β is no longer sensitive to controls once $\ln(Distance)$ has been included.

7. Conclusion

In this paper, we have shown that markets in colonial South Asia that were more linguistically distant from each other displayed less market integration, conditional on many other measures, including distance, literacy gaps, transportation links, and measures of dissimilarity. This finding holds across multiple products and markets, and survives several sensitivity checks. Genetic distance and lack of railway connections may help explain these results, but on their own, these factors do not explain the lack of market integration. There is less evidence for missing markets and presence of trading communities as mechanisms. The results show that cultural and linguistic barriers are salient to the functioning of markets, and that their importance is not limited to political economy or post-colonial, modern economies. Furthermore, the contribution of these cultural factors that enhance or impede market integration is substantial relative to other factors such as physical distance. More

TABLE 8. Railway connections

	(1)	(2)	(3)	(4)
		Years Both Connected to Railroad		
Linguistic Distance	-4.388**	-0.852*	-4.349*	-0.236
	(2.138)	(0.485)	(2.406)	(0.452)
N	21,115	21,115	21,115	21,115
Rsq	0.009	0.850	0.170	0.853
		Correlation: Wheat		
Linguistic Distance	-0.258***	-0.210***	-0.026	-0.067**
	(0.035)	(0.036)	(0.025)	(0.030)
Years Both Connected to Railroad	-0.000	0.000	-0.000	-0.000
	(0.001)	(0.000)	(0.000)	(0.000)
N	15,652	15,652	15,652	15,652
Rsq	0.140	0.762	0.580	0.806
		Correlation: Salt		
Linguistic Distance	-0.473***	-0.391***	-0.381***	-0.189***
	(0.060)	(0.072)	(0.051)	(0.044)
Years Both Connected to Railroad	0.002***	0.001***	0.001	0.000
	(0.001)	(0.000)	(0.001)	(0.000)
N	20,909	20,909	20,909	20,909
Rsq	0.227	0.709	0.567	0.791
		Correlation: Rice		
Linguistic Distance	-0.081***	-0.073***	-0.055***	-0.035***
	(0.017)	(0.010)	(0.018)	(0.010)
Years Both Connected to Railroad	0.001	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
N	20,909	20,909	20,909	20,909
Rsq	0.048	0.834	0.282	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE 9. Altonji-Elder-Taber Statistics

		Correlation: Wheat	
Baseline: No Controls	4.476	0.0977	0.351
Baseline: ln(distance)	2.437	0.141	0.562
		Correlation: Salt	
Baseline: No Controls	4.245	3.849	0.640
Baseline: ln(distance)	2.289	-9.983	1.205
		Correlation: Rice	
Baseline: No Controls	7.219	2.051	0.714
Baseline: ln(distance)	1.495	-2.525	-39.48
Fixed Effects	Yes	No	Yes
Controls	No	Yes	Yes

linguistically-similar markets are more likely to have been connected earlier via transport infrastructure (the colonial railway system), but this connection alone does not explain away the coefficient. These results indicate the importance and persistence of cultural differences in market integration, trade, and price volatility. Testing whether markets with greater gains

from trade learn the languages necessary for trade over time, and whether newer information and communication technologies reduce the importance of linguistic distance, remain important questions for future work.

References

Adams, J. and West, R. C. (1979). Money, prices, and economic development in India, 1861–1895. *The Journal of Economic History*, 39(01):55–68.

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the Origins of Gender Roles: Women and the Plough. *The Quarterly Journal of Economics*, 128(2):469–530.

Allen, R. C. (2007). India in the great divergence. *The new comparative economic history: Essays in honor of Jeffrey G. Williamson*, pages 9–32.

Allen, T. (2014). Information frictions in trade. *Econometrica*, 82(6):2041–2083.

Alsan, M. (2015). The effect of the tsetse fly on African development. *The American Economic Review*, 105(1):382–410.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy*, 113(1):151–184.

Anderson, J. E. and Van Wincoop, E. (2004). Trade costs. *Journal of Economic Literature*, 42(3):691–751.

Andrabi, T. and Kuehlwein, M. (2010). Railways and price convergence in British India. *The Journal of Economic History*, 70(02):351–377.

Asher, R. E. (2008). Language in historical context. *Language in South Asia*, pages 31–48.

Ashraf, Q. and Galor, O. (2011). Dynamics and stagnation in the Malthusian epoch. *The American Economic Review*, 101(5):2003–2041.

Ashraf, Q. and Galor, O. (2013). Genetic diversity and the origins of cultural fragmentation. *The American Economic Review*, 103(3):528–533.

Atkin, D. (2013). Trade, tastes, and nutrition in India. *The American Economic Review*, 103(5):1629–1663.

Atkin, D. (2016). The caloric costs of culture: Evidence from Indian migrants. *The American Economic Review*, 106(4):1144–1181.

Bai, Y. and Kung, J. (2017). Does Genetic Distance Have a Barrier Effect on Technology Diffusion? Evidence from Historical China. *Working Paper: Hong Kong University of Science and Technology*.

Bellows, J. and Miguel, E. (2009). War and local collective action in Sierra Leone. *Journal of Public Economics*, 93(11):1144–1157.

Bhattacharya, S. (1983). Regional Economy (1757–1857): Eastern India. *The Cambridge Economic History of India*, 2:270–95.

Burgess, R. and Donaldson, D. (2010). Can openness mitigate the effects of weather shocks? Evidence from India's famine era. *American Economic Review*, 100(2):449–53.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Chandavarkar, A. G. (1983). Money and credit, 1858–1947. *The Cambridge Economic History of India*, 2:762–803.

Chaudhary, L. (2009). Determinants of primary schooling in British India. *The Journal of Economic History*, 69(1):269–302.

Chaudhary, L. and Garg, M. (2015). Does history matter? Colonial education investments in India. *The Economic History Review*, 68(3):937–961.

Chaudhary, L., Musacchio, A., Nafziger, S., and Yan, S. (2012). Big BRICs, weak foundations: The beginning of public elementary education in Brazil, Russia, India, and China. *Explorations in Economic History*, 49(2):221–240.

Collins, W. J. (1999). Labor mobility, market integration, and wage convergence in late 19th century India. *Explorations in Economic History*, 36(3):246–277.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, 92(1):1–45.

Derbyshire, I. D. (1987). Economic Change and the Railways in North India, 1860–1914. *Modern Asian Studies*, 21(03):521–545.

Desmet, K., Gomes, J. F., and Ortuño-Ortín, I. (2020). The geography of linguistic diversity and the provision of public goods. *Journal of Development Economics*, 143:102384.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2012). The political economy of linguistic cleavages. *Journal of Development Economics*, 97(2):322–338.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2017). Culture, ethnicity and diversity. *American Economic Review*, 107(9):2479–2513.

Dickens, A. (2018). Ethnolinguistic Favouritism in African Politics. *American Economic Journal: Applied Economics*, 10(3):370–402.

Divekar, V. (1983). Regional Economy (1757–1857): Western India. *The Cambridge Economic History of India*, 2:332–51.

Donaldson, D. (2018). Railroads of the Raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5):899–934.

Dyen, I., Kruskal, J. B., and Black, P. (1992). An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical society*, 82(5):iii–132.

Egger, P. H. and Lassmann, A. (2012). The language effect in international trade: A meta-analysis. *Economics Letters*, 116(2):221–224.

Emeneau, M. B. (1956). India as a lingustic area. *Language*, 32(1):3–16.

Esteban, J., Mayoral, L., and Ray, D. (2012). Ethnicity and conflict: An empirical study. *The American Economic Review*, 102(4):1310–1342.

Estevadeordal, A., Frantz, B., Taylor, A. M., et al. (2003). The Rise and Fall of World Trade, 1870–1939. *The Quarterly Journal of Economics*, 118(2):359–407.

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

Federico, G. (2011). When did European markets integrate? *European Review of Economic History*, 15(1):93–126.

Frawley, W. J. (2003). *International encyclopedia of linguistics*, volume 4. Oxford University Press.

Gamkrelidze, T. V. and Ivanov, V. V. (1990). The early history of Indo-European languages. *Scientific American*, 262(3):110–117.

Giuliano, P., Spilimbergo, A., and Tonon, G. (2014). Genetic distance, transportation costs, and trade. *Journal of Economic Geography*, 14(1):179–198.

Gomes, J. F. (2014). The health costs of ethnic distance: evidence from Sub-Saharan Africa. *ISER Working Paper Series 2014-33*.

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

Gupta, B. (2014). Discrimination or social networks? Industrial investment in colonial India. *The Journal of Economic History*, 74(1):141–168.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207.

Hurd, J. (1975). Railways and the Expansion of Markets in India, 1861–1921. *Explorations in Economic History*, 12(3):263–288.

Hutchinson, W. K. (2005). "Linguistic distance" as a determinant of bilateral trade. *Southern Economic Journal*, 72(1):1–15.

Isphording, I. E. and Otten, S. (2014). Linguistic barriers in the destination language acquisition of immigrants. *Journal of Economic Behavior & Organization*, 105:30–50.

Iwanowsky, M. (2017). The Role of Ethnic Networks in Africa: Evidence from Cross-Country Trade. *Working Paper*.

Jacks, D. S., Meissner, C. M., and Novy, D. (2008). Trade Costs, 1870–2000. *The American Economic Review*, 98(2):529–534.

Jain, T. (2017). Common tongue: The impact of language on educational outcomes. *Journal of Economic History*, 77(2):473–510.

Jia, R. (2014). Weather shocks, sweet potatoes and peasant revolts in historical China. *The Economic Journal*, 124(575):92–118.

Kessinger, T. G. (1983). Regional Economy (1757–1857): North India. *The Cambridge Economic History of India*, 2:242–270.

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S. E., and Sachs, J. (2004). A global index representing the stability of malaria transmission. *The American journal of tropical medicine and hygiene*, 70(5):486–498.

Krishnamurti, B. (2003). *The Dravidian languages*. Cambridge University Press.

Kumar, D. (1983). Regional Economy (1757–1857): South India. *The Cambridge Economic History of India*, 2:352–375.

Laitin, D. and Ramachandran, R. (2016). Language Policy and Human Development. *American Political Science Review*, 110(3):457–480.

Lameli, A., Nitsch, V., Südekum, J., and Wolf, N. (2015). Same same but different: Dialects and trade. *German Economic Review*, 16(3):290–306.

Laval, G., Patin, E., and Rueda, V. (2016). Achieving the American Dream: Cultural Distance, Cultural Diversity and Economic Performance. *Oxford Economic and Social History Working Paper 140*.

Markovits, C. (2008). *Merchants, traders, entrepreneurs: Indian business in the colonial era*. Springer.

Matsuura, K. and Willmott, C. (2007). Terrestrial Air Temperature and Precipitation: 1900-2006 Gridded Monthly Time Series, Version 1.01. *University of Delaware*.

McAlpin, M. (1983). Price Movements and Economic Activity (1860-1947). *The Cambridge Economic History of India*, 2:878–904.

McAlpin, M. B. (1974). Railroads, Prices, and Peasant Rationality: India 1860–1900. *The Journal of Economic History*, 34(03):662–684.

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

Michalopoulos, S. (2012). The origins of ethnolinguistic diversity. *The American Economic Review*, 102(4):1508–1539.

Montaut, A. (2005). Colonial Language Classification, Post-colonial Language Movements and the Grassroot Multilingualism Ethos in India. *Mushirul Hasan & Asim Roy. Living Together Separately. Cultural India in History and Politics*, pages 75–116.

Moxham, R. (2001). *The great hedge of India*. Constable.

Nunn, N. and Puga, D. (2012). Ruggedness: The blessing of bad geography in Africa. *Review of Economics and Statistics*, 94(1):20–36.

Nunn, N. and Wantchekon, L. (2011). The slave trade and the origins of mistrust in Africa. *The American Economic Review*, 101(7):3221–3252.

O'Rourke, K. H. and Williamson, J. G. (2002). When did globalisation begin? *European Review of Economic History*, 6(1):23–50.

Özak, Ö. (2010). The Voyage of Homo-Economicus: Some Economic Measures of Distance. *Working Paper: Department of Economics, Southern Methodist University*.

Özak, Ö. (2018). Distance to the technological frontier and economic development. *Journal of Economic Growth*, 23(2):175–221.

Pandit, P. B. (1977). *Language in a plural society*. New Delhi: Dev Raj Chanana Memorial Committee.

Pascali, L. (2017). The wind of change: Maritime technology, trade and economic development. *American Economic Review*, 107(9):2821–2854.

Pemberton, T. J., DeGiorgio, M., and Rosenberg, N. A. (2013). Population structure in a comprehensive genomic data set on human microsatellite variation. *G3: Genes, Genomes, Genetics*, 7(2):g3–113.

Persaud, A. (2019). Escaping local risk by entering indentureship: Evidence from nineteenth-century indian migration. *The Journal of Economic History*, 79(2):447–476.

Persson, K. G. (1999). *Grain markets in Europe, 1500–1900: Integration and deregulation*, volume 7. Cambridge University Press.

Ramankutty, N., Foley, J. A., Norman, J., and McSweeney, K. (2002). The global distribution of cultivable lands: current patterns and sensitivity to possible climate change. *Global Ecology and Biogeography*, 11(5):377–392.

Rauch, J. E. and Trindade, V. (2002). Ethnic Chinese networks in international trade. *Review of Economics and Statistics*, 84(1):116–130.

Renfrew, C. (1989). The origins of Indo-European languages. *Scientific American*, 261(4):106–115.

Richards, J. F. (1995). *The Mughal Empire*, volume 5. Cambridge University Press.

Roy, T. (2012). *India in the world economy: from antiquity to the present*. Cambridge University Press.

Roy, T. (2014). Trading Firms in Colonial India. *Business History Review*, 88(1):9–42.

Shastry, G. K. (2012). Human capital response to globalization education and information technology in India. *Journal of Human Resources*, 47(2):287–330.

Shiue, C. H. and Keller, W. (2007). Markets in China and Europe on the Eve of the Industrial Revolution. *The American Economic Review*, 97(4):1189–1216.

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

Spolaore, E. and Wacziarg, R. (2016). Fertility and modernity. *UCLA CCPR Population Working Papers PWP-CCPR-2016-016*.

Spolaore, E. and Wacziarg, R. (2018). Ancestry and development: New evidence. *Journal of Applied Econometrics*, 33(5):748–762.

Studer, R. (2008). India and the great divergence: Assessing the efficiency of grain markets in eighteenth-and nineteenth-century India. *Journal of Economic History*, 68(02):393–437.

Waldinger, M. (2014). The economic effects of long-term climate change: Evidence from the little ice age. *Working Paper: London School of Economics*.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.

Wichmann, S., Müller, A., Velupillai, V., Brown, C. H., Holman, E. W., Brown, P., Sauppe, S., Belyaev, O., Urban, M., Molochieva, Z., et al. (2016). The asjp database (version 17). *URL: http://email. eva. mpg. de/˜ wichmann/ASJPHomePage. htm*.

Online Appendix: Not for publication

Appendix A. Additional robustness

Here, we present the results of additional robustness exercises not discussed in the text.

A.0.1. *Cost distance.* While in our baseline we control for the natural logarithm of pairwise distance in kilometers, we can show that our results survive controlling for an alternative cost distance measure constructed by Özak (2010, 2018). Using data on the maximum speeds that dismounted infantry can sustain in given conditions based on climate, topography, and terrain, Özak computes the time needed to cross any given grid cell. The cost distance between any two markets, then, is simply the number of weeks needed along the quickest routes between them. Results appear in Table A6 and are almost unchanged. This should not be surprising: the correlation coefficient between this distance measure and our baseline distance measure (log kilometers) is 0.8961.

A.0.2. *Other crops.* Although we have focused our analysis on the crops whose prices are reported most in the data (wheat, salt and rice), we are able to show similar results for a wide range of other crops. These data are again taken from *Wages and Prices in India*. We present estimates of (1) for these other prices and wages in tables A7, A8, A9, and A10. Several other prices show patterns similar to our main results. Where the conditional correlation between market integration and linguistic distance is insignificant, this is often for products whose pairwise price correlations we can compute for a much smaller set of market pairs than our main results.

A.1. **Sample.** In Table A11, we restrict our sample to modern India, in order to assuage concerns that the results are driven by comparisons between broad, administratively distinct, culturally dissimilar, and geographically distant regions, particularly in Burma. In Table A12, we remove any negative price correlations from the sample. In Table A13, we remove outliers by discarding the top and bottom 5 percent of observations by values of ρ_{ij}^{p}. In Table A14, we instead remove outliers by discarding the top and bottom 5 percent of observations by values of linguistic distance. Table A15, we show that market pairs with correlations computed from sparse data do not drive the results by only keeping pairs with at least ten observations in common.

In Table A16, we discard all markets with city populations above 75,000 in order to demonstrate that results are not driven by observations with unusual linguistic diversity and markets that may work differently than elsewhere. In Table A17, we drop coastal markets. These too might be unusually diverse in language and well integrated with other markets both domestic and foreign. In Table A18, we drop Gangetic markets, which are overwhelmingly Hindi-speaking and likely to be well integrated with each other. Tables A19 and A20 report results using only price observations from before or after 1891 (the midpoint in the sample) to compute ρ_{ij}^{p}. Across these sample restriction exercises, results remain similar to the baseline.

In figures A2, A3, and A4, we show that our results (corresponding to column (4) in Table 2) when we restrict our results to markets within a maximum cutoff distance from each other. For cutoffs of 1500 km and greater for wheat, 1000 km and greater for salt, and 750 km and greater for rice, results are similar in magnitude and significance to our baseline.

While readers may be concerned that our results are driven by linguistically similar markets facing correlated shocks, we note that our baseline analysis controls for the correlation in rainfall between two markets. As a further check, we drop all market pairs within 500 kilometers of each other in Table A21. Results are similar to the baseline except that the results with the correlation in wheat prices as an outcome have become insignificant in one column.

A.2. **Measures of linguistic distance and market integration.** In Table A22 we replace our baseline measure of market integration with the natural logarithm of (one plus) the correlation coefficient. Similarly, in Table A23 we replace our main measure with centiles of the correlation coefficient. In Table A24 we replace our baseline measure of linguistic distance with an alternative in which $\delta = 0.5$. In Table A25, we instead use the pairwise distance between the largest language in each district to compute linguistic distance. In Table A26, similarly, we use a dummy for whether the largest language differs. These exercises give results similar to those in Table 2.

Our baseline measure of linguistic distance follows the literature (e.g. Esteban et al. (2012)) in taking a nonlinear transformation of the number of branches shared by two languages. The results in figures 6, 7, and 8, in which we replace this with a dummy for having fewer than a given number of branches, is an alternative nonlinear transformation. Other nonlinear transformations are not as predictive of market integration. In Table A27, we include the square of linguistic distance as an additional right-hand-side variable. This adds noise to the estimation, often making the linear term insignificant while not itself being statistically significant. In Tables A28 and A29, we show that results obtained when taking the log of linguistic distance, or both the correlation coefficient and linguistic distance, are somewhat similar to our baseline results, but generally do not survive the inclusion of both controls and fixed effects. The R-squared values corresponding to the specification with fixed effects and controls are larger in our baseline than in the log-log specification: the relevant values are 0.81 and 0.70 for wheat, 0.61 and 0.45 for salt, and 0.87 and 0.80 for rice.

We report two alternative measures of linguistic distance, computed from the Wichmann et al. (2016) Automated Similarity Judgment Program Database. The first is an alternative cladistic measure that replaces the classification trees from Ethnologue with the classification trees from Glottolog. We use the same procedure as in section 3.2.2 to compute these distances. However, of the 257 unique ISO codes we match to languages in the 1901 census,

only 158 are present in the ASJP data. Like our genetic distance calculations in (5), then, we scale population shares by the share actually matched to the ASJP data.

The second alternative is a lexicostatistical measure similar to that in Dickens (2018). For 100 standard words (e.g. blood, bone) in each language, the ASJP reports the word in a standardized phonetic orthography. For any pair of languages, we compute the average Levenshtein distance between words that have the same meaning, and the average Levenshtein distance between words that have different meanings. The ratio of the two is a measure of linguistic distances across languages, corrected for any accidental similarity of sounds across words with different meanings. Because this ratio can be greater than one, we divide this by its maximum to rescale it between zero and one. We then use these language distances when computing linguistic distances between districts, again rescaling population shares by the share actually matched to the ASJP data.

Results are presented in tables A30 and A31. Though these have some similarities to our baseline measures, they are not as robust, being statistically insignificant in a larger number of specifications. Given the incomplete set of languages and the incomplete word lists in the data (the average entry in the ASJP data reports only 37 words), it is likely that this is due in part to measurement error of the right-hand-side variable.

A.3. **Standard errors.** Tables A32 and A33 present alternative approaches to standard errors. Rather than clustering by market i and market j, we report two-way clustering by either the largest language in each district or by the province in which each district falls. To account for possible correlation over space in the error term, we report Conley (1999) standard errors in Table A34, allowing dependence at distances up to five decimal degrees.

A.4. **Convergence.** Because it is possible that the gradual erosion of a large price gap across two markets could produce a negative correlation in the prices recorded in the two markets, we show that our results survive controlling for the mean absolute log price difference between any two markets. Results are presented in Table A35 and the results are little different from our main results.

A.5. **Additional checks.** We show in Table A36 that there is a significant coefficient on the interaction between linguistic and physical distance in our main equation only in one of the twelve reported specifications (fixed effects and controls for rice). For this exercise, we convert log physical distance into a standardized $N(0, 1)$ variable. We recognize that linguistic distance may simply be a marker of other differences across populations, such as the degree of shared history; thus, we show in Table A37, the results that we obtain when we control for whether both markets were part of the Mughal empire. In particular, using the maps in Richards (1995), we consider the extent of the empire in 1605, at the death of Akbar, and in 1707, at its maximum extent. Results are similar to our baseline. Results

for rice are the lone exception; these results are insignificant in two specifications. We show in Table A38 that results are similar if religion from the 1901 census is used to compute religious distance.

Appendix B. Additional Figures

FIGURE A1. Ludhiana: Genetic distances

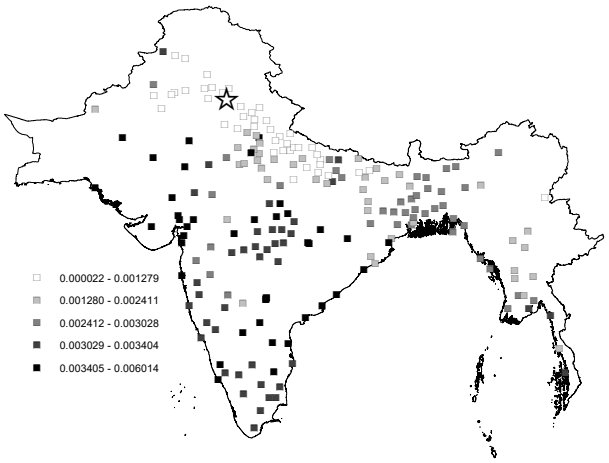


FIGURE A2. Distance cutoffs: Wheat

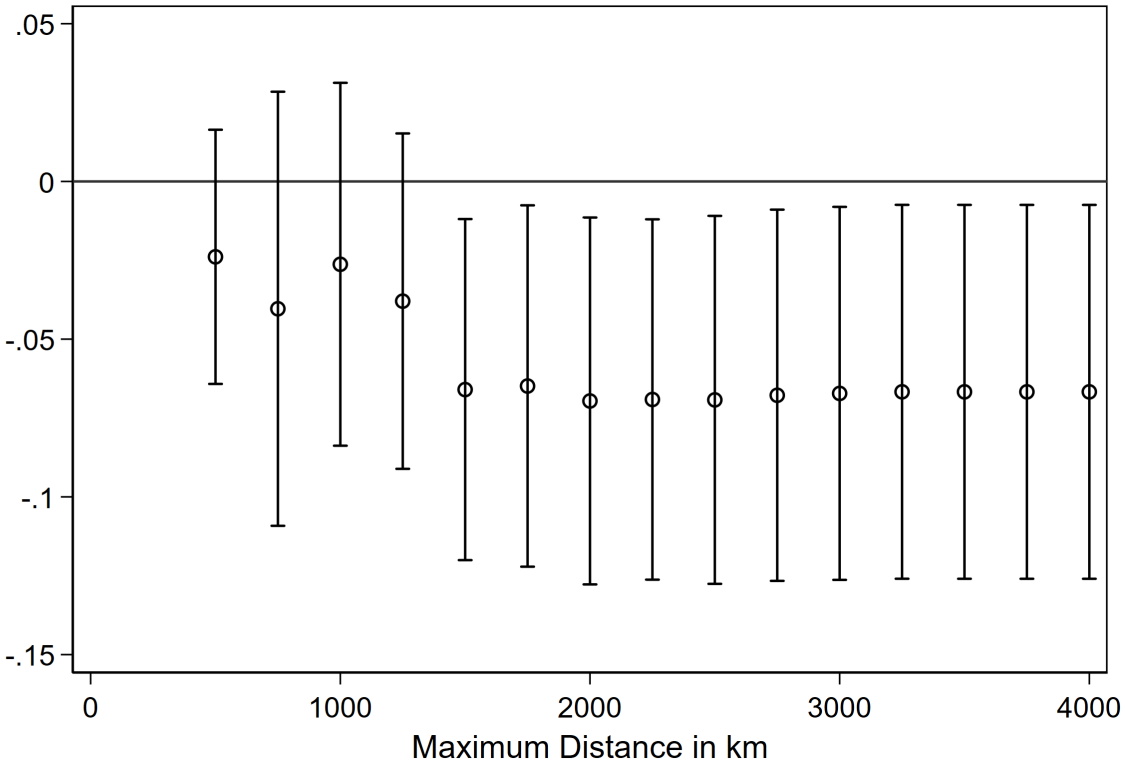


FIGURE A3. Distance cutoffs: Salt

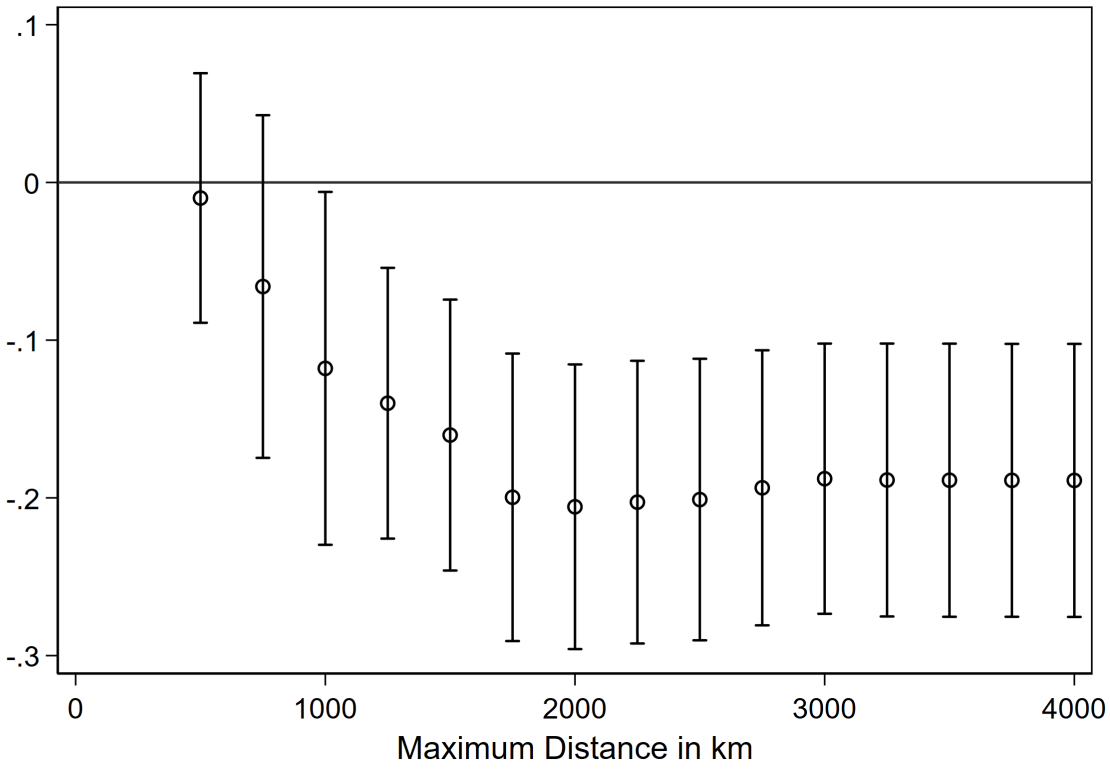
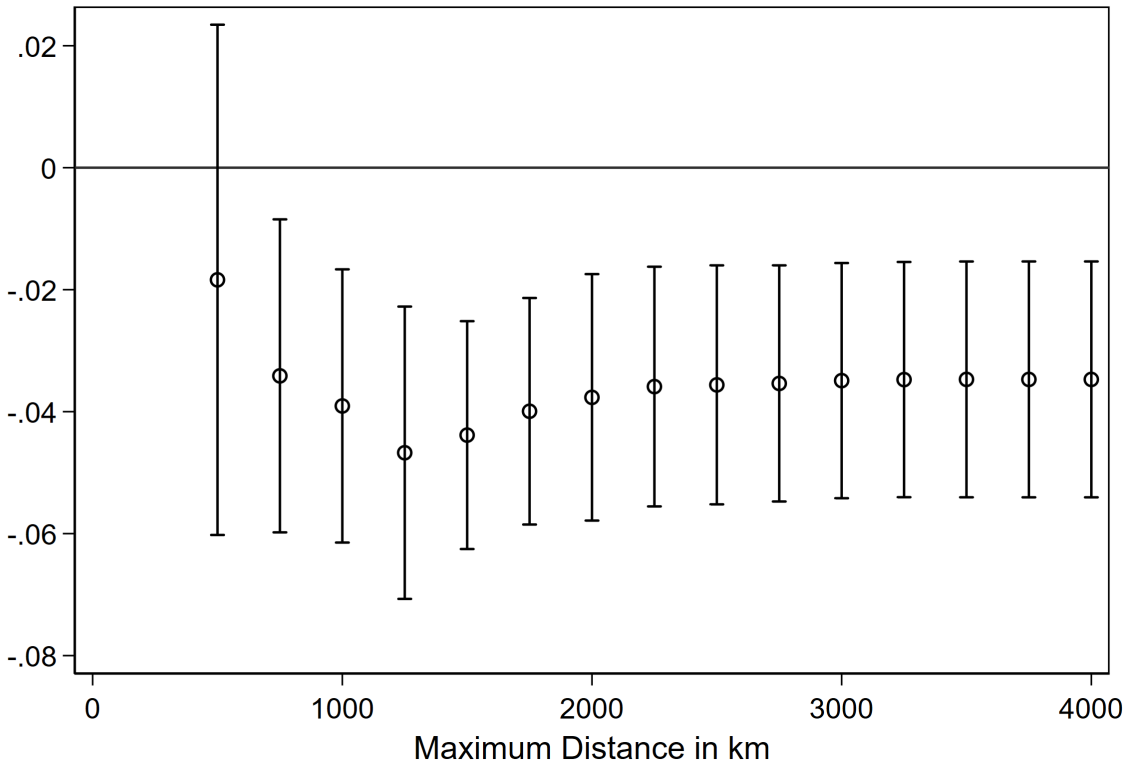


FIGURE A4. Distance cutoffs: Rice



Appendix C. Additional Tables

Table A1. Correlation coefficients: Part 1

	Corr.: Salt	Corr.: Wheat	Corr.: Rice	Linguistic Distance	Ln Distance in KM	Same Province	Both Coastal	Same River	Rainfall Corr.
Correlation: Salt	1	.248	.286	-.46	-.42	.277	-.05	.161	.275
Correlation: Wheat	.248	1	.165	-.37	-.36	.146	-.12	.083	.351
Correlation: Rice	.286	.165	1	-.21	-.28	.226	-.06	.117	.185
Linguistic Distance (d=0.05)	-.46	-.37	-.21	1	.510	-.25	.106	-.17	-.36
Ln Distance in KM	-.42	-.36	-.28	.510	1	-.57	.026	-.29	-.72
Same Province	.277	.146	.226	-.25	-.57	1	.059	.301	.457
Both Coastal	-.05	-.12	-.06	.106	.026	.059	1	-.02	-.02
Same River	.161	.083	.117	-.17	-.29	.301	-.02	1	.215
Rainfall Correlation	.275	.351	.185	-.36	-.72	.457	-.02	.215	1
Difference in Land Quality	-.05	-.03	-.02	-.08	.289	-.16	-.04	-.09	-.20
Difference in Ruggedness	-.31	-.17	-.09	.364	.271	-.12	.083	-.07	-.18
Difference in Malaria	-.39	-.29	-.20	.290	.307	-.06	.144	-.01	-.27
Difference in Humidity	-.18	-.23	-.16	.169	.455	-.28	-.08	-.13	-.37
Difference in Altitude	.027	-.10	-.04	.110	.184	-.15	-.08	-.06	-.12
Difference in Banana Suitability	-.07	-.29	-.12	.189	.273	-.08	.076	-.04	-.34
Difference in Chickpea Suitability	-.20	-.10	-.04	.096	.240	-.21	-.13	-.12	-.17
Difference in Cocoa Suitability	-.22	-.44	-.05	.322	.290	-.03	.170	-.01	-.26
Difference in Cotton Suitability	-.07	.020	.029	-.13	.158	-.08	-.07	-.03	-.05
Difference in Groundnut Suitability	-.13	-.09	-.05	-.01	.222	-.08	.031	-.08	-.14
Difference in Dry Rice Suitability	-.18	-.23	-.06	.267	.478	-.26	-.09	-.09	-.31
Difference in Oil Palm Suitability	-.10	-.42	-.01	.190	.202	-.00	.171	-.01	-.21
Difference in Onion Suitability	-.10	.037	-.07	-.04	.204	-.09	.001	-.08	-.05
Difference in Precipitation	-.40	-.29	-.21	.227	.423	-.17	.081	-.11	-.36
Difference in Slope	-.35	-.16	-.09	.403	.284	-.12	.107	-.09	-.17
Difference in Soybean Suitability	-.12	-.00	-.03	-.10	.183	-.12	-.05	-.07	-.07
Difference in Sugar Suitability	-.12	-.32	-.17	.258	.452	-.23	.005	-.06	-.46
Difference in Tea Suitability	-.01	-.29	-.14	.102	.202	-.07	.046	-.04	-.32
Difference in Wetland Rice Suitability	-.28	-.22	-.19	.192	.512	-.26	-.02	-.13	-.40
Difference in White Potato Suitability	-.17	-.04	-.03	.022	.251	-.16	-.12	-.07	-.11
Difference in Wheat Suitability	-.20	-.08	-.03	.063	.301	-.21	-.15	-.08	-.16
Difference in Tomato Suitability	-.11	.069	-.09	-.14	.167	-.10	-.04	-.05	-.06
Difference in Temperature	-.00	-.04	-.03	.151	.261	-.14	-.03	-.04	-.13
Latitude Difference	-.19	-.25	-.02	.531	.605	-.29	-.04	-.16	-.31
Longitude Difference	-.44	-.32	-.35	.286	.672	-.31	.079	-.13	-.54
Religious Distance	-.44	-.18	-.23	.311	.397	-.16	.044	-.10	-.30

	D. in Land Quality	D. in Ruggedness	D. in Malaria	D. in Humidity	D. in Altitude	D. in Banana Suit.	D. in Chickpea Suit.	D. in Cocoa Suit.	D. in Cotton Suit.
Correlation: Salt	-.05	-.31	-.39	-.18	.027	-.07	-.20	-.22	-.07
Correlation: Wheat	-.03	-.17	-.29	-.23	-.10	-.29	-.10	-.44	.020
Correlation: Rice	-.02	-.09	-.20	-.16	-.04	-.12	-.04	-.05	.029
Linguistic Distance (d=0.05)	-.08	.364	.290	.169	.110	.189	.096	.322	-.13
Ln Distance in KM	.289	.271	.307	.455	.184	.273	.240	.290	.158
Same Province	-.16	-.12	-.06	-.28	-.15	-.08	-.21	-.03	-.08
Both Coastal	-.04	.083	.144	-.08	-.08	.076	-.13	.170	-.07
Same River	-.09	-.07	-.01	-.13	-.06	-.04	-.12	-.01	-.03
Rainfall Correlation	-.20	-.18	-.27	-.37	-.12	-.34	-.17	-.26	-.05
Difference in Land Quality	1	-.05	.007	.265	.002	.009	.009	-.02	.469
Difference in Ruggedness	-.05	1	.089	.097	.352	.239	.075	.278	-.04
Difference in Malaria	.007	.089	1	.228	-.01	.377	-.01	.510	-.09
Difference in Humidity	.265	.097	.228	1	.157	.395	.019	.282	.206
Difference in Altitude	.002	.352	-.01	.157	1	.015	.021	.043	.016
Difference in Banana Suitability	.009	.239	.377	.395	.015	1	-.00	.535	-.02
Difference in Chickpea Suitability	.009	.075	-.01	.019	.021	-.00	1	-.01	.097
Difference in Cocoa Suitability	-.02	.278	.510	.282	.043	.535	-.01	1	-.08
Difference in Cotton Suitability	.469	-.04	-.09	.206	.016	-.02	.097	-.08	1
Difference in Groundnut Suitability	.458	.020	.072	.287	.060	.234	.054	.142	.747
Difference in Dry Rice Suitability	.193	.176	.232	.252	.060	.159	.144	.302	.091
Difference in Oil Palm Suitability	-.01	.184	.354	.218	-.03	.543	-.02	.847	-.04
Difference in Onion Suitability	.511	.031	-.12	.208	.136	-.12	.045	-.11	.668
Difference in Precipitation	.231	.287	.510	.498	.022	.528	.021	.422	.151
Difference in Slope	-.05	.933	.124	.113	.261	.238	.059	.328	-.06
Difference in Soybean Suitability	.504	-.02	-.02	.283	.165	-.01	.025	-.00	.786
Difference in Sugar Suitability	.090	.169	.360	.677	.047	.705	-.00	.432	-.03
Difference in Tea Suitability	.023	.157	.330	.361	.032	.867	-.02	.400	-.05
Difference in Wetland Rice Suitability	.475	.129	.289	.667	.032	.347	.028	.297	.255
Difference in White Potato Suitability	.066	.176	-.05	-.04	.092	.017	.600	-.05	.189
Difference in Wheat Suitability	.050	.148	-.05	-.03	.048	.017	.648	-.04	.205
Difference in Tomato Suitability	.534	.048	-.11	.240	.093	-.02	.118	-.12	.732
Difference in Temperature	.109	.244	-.06	.069	.309	.090	.196	-.08	.194
Latitude Difference	.217	.241	-.00	.144	.107	.061	.254	.222	.151
Longitude Difference	.236	.199	.525	.447	.085	.348	.039	.280	.113
Religious Distance	.249	.208	.515	.169	-.02	.258	.009	.280	.253

Table A2. Correlation coefficients: Part 2

	D. in Groundnut Suit.	D. in Dry Rice Suit.	D. in Oil Palm Suit.	D. in Onion Suit.	D. in Precipitation	D. in Slope	D. in Soybean Suit.	D. in Sugar Suit.	D. in Tea Suit.
Correlation: Salt	-.13	-.18	-.10	-.10	-.40	-.35	-.12	-.12	-.01
Correlation: Wheat	-.09	-.23	-.42	.037	-.29	-.16	-.00	-.32	-.29
Correlation: Rice	-.05	-.06	-.01	-.07	-.21	-.09	-.03	-.17	-.14
Linguistic Distance (d=0.05)	-.01	.267	.190	-.04	.227	.403	-.10	.258	.102
Ln Distance in KM	.222	.478	.202	.204	.423	.284	.183	.452	.202
Same Province	-.08	-.26	-.00	-.09	-.17	-.12	-.12	-.23	-.07
Both Coastal	.031	-.09	.171	.001	.081	.107	-.05	.005	.046
Same River	-.08	-.09	-.01	-.08	-.11	-.09	-.07	-.06	-.04
Rainfall Correlation	-.14	-.31	-.21	-.05	-.36	-.17	-.07	-.46	-.32
Difference in Land Quality	.458	.193	-.01	.511	.231	-.05	.504	.090	.023
Difference in Ruggedness	.020	.176	.184	.031	.287	.933	-.02	.169	.157
Difference in Malaria	.072	.232	.354	-.12	.510	.124	-.02	.360	.330
Difference in Humidity	.287	.252	.218	.208	.498	.113	.283	.677	.361
Difference in Altitude	.060	.060	-.03	.136	.022	.261	.165	.047	.032
Difference in Banana Suitability	.234	.159	.543	-.12	.528	.238	-.01	.705	.867
Difference in Chickpea Suitability	.054	.144	-.02	.045	.021	.059	.025	-.00	-.02
Difference in Cocoa Suitability	.142	.302	.847	-.11	.422	.328	-.00	.432	.400
Difference in Cotton Suitability	.747	.091	-.04	.668	.151	-.06	.786	-.03	-.05
Difference in Groundnut Suitability	1	.056	.183	.736	.314	.014	.826	.143	.213
Difference in Dry Rice Suitability	.056	1	.190	.033	.261	.204	.035	.306	.110
Difference in Oil Palm Suitability	.183	.190	1	-.09	.268	.197	.031	.344	.312
Difference in Onion Suitability	.736	.033	-.09	1	.146	-.00	.719	-.00	-.13
Difference in Precipitation	.314	.261	.268	.146	1	.358	.208	.545	.509
Difference in Slope	.014	.204	.197	-.00	.358	1	-.04	.185	.172
Difference in Soybean Suitability	.826	.035	.031	.719	.208	-.04	1	.013	-.02
Difference in Sugar Suitability	.143	.306	.344	-.00	.545	.185	.013	1	.648
Difference in Tea Suitability	.213	.110	.312	-.13	.509	.172	-.02	.648	1
Difference in Wetland Rice Suitability	.351	.308	.184	.375	.653	.160	.345	.637	.304
Difference in White Potato Suitability	-.00	.163	-.05	.060	-.03	.122	-.02	-.03	-.03
Difference in Wheat Suitability	.005	.316	-.04	.043	-.02	.108	-.01	-.02	-.01
Difference in Tomato Suitability	.654	.028	-.11	.822	.183	.011	.734	-.01	-.01
Difference in Temperature	.067	.127	-.08	.120	.000	.160	.040	.015	.067
Latitude Difference	.109	.543	.190	.145	.098	.237	.058	.091	-.04
Longitude Difference	.207	.205	.172	.133	.567	.238	.171	.514	.320
Religious Distance	.299	.175	.169	.126	.421	.262	.273	.240	.230

	D. in Wetland Rice Suit.	D. in White Potato Suit.	D. in Wheat Suit.	D. in Tomato Suit.	D. in Temperature	Latitude D.	Longitude D.	Religious Distance
Correlation: Salt	-.28	-.17	-.20	-.11	-.00	-.19	-.44	-.44
Correlation: Wheat	-.22	-.04	-.08	.069	-.04	-.25	-.32	-.18
Correlation: Rice	-.19	-.03	-.03	-.09	-.03	-.02	-.35	-.23
Linguistic Distance (d=0.05)	.192	.022	.063	-.14	.151	.531	.286	.311
Ln Distance in KM	.512	.251	.301	.167	.261	.605	.672	.397
Same Province	-.26	-.16	-.21	-.10	-.14	-.29	-.31	-.16
Both Coastal	-.02	-.12	-.15	-.04	-.03	-.04	.079	.044
Same River	-.13	-.07	-.08	-.05	-.04	-.16	-.13	-.10
Rainfall Correlation	-.40	-.11	-.16	-.06	-.13	-.31	-.54	-.30
Difference in Land Quality	.475	.066	.050	.534	.109	.217	.236	.249
Difference in Ruggedness	.129	.176	.148	.048	.244	.241	.199	.208
Difference in Malaria	.289	-.05	-.05	-.11	-.06	-.00	.525	.515
Difference in Humidity	.667	-.04	-.03	.240	.069	.144	.447	.169
Difference in Altitude	.032	.092	.048	.093	.309	.107	.085	-.02
Difference in Banana Suitability	.347	.017	.017	-.02	.090	.061	.348	.258
Difference in Chickpea Suitability	.028	.600	.648	.118	.196	.254	.039	.009
Difference in Cocoa Suitability	.297	-.05	-.04	-.12	-.08	.222	.280	.280
Difference in Cotton Suitability	.255	.189	.205	.732	.194	.151	.113	.253
Difference in Groundnut Suitability	.351	-.00	.005	.654	.067	.109	.207	.299
Difference in Dry Rice Suitability	.308	.163	.316	.028	.127	.543	.205	.175
Difference in Oil Palm Suitability	.184	-.05	-.04	-.11	-.08	.190	.172	.169
Difference in Onion Suitability	.375	.060	.043	.822	.120	.145	.133	.126
Difference in Precipitation	.653	-.03	-.02	.183	.000	.098	.567	.421
Difference in Slope	.160	.122	.108	.011	.160	.237	.238	.262
Difference in Soybean Suitability	.345	-.02	-.01	.734	.040	.058	.171	.273
Difference in Sugar Suitability	.637	-.03	-.02	-.01	.015	.091	.514	.240
Difference in Tea Suitability	.304	-.03	-.01	-.01	.067	-.04	.320	.230
Difference in Wetland Rice Suitability	1	-.00	-.01	.391	.016	.168	.586	.328
Difference in White Potato Suitability	-.00	1	.935	.206	.526	.364	.026	.118
Difference in Wheat Suitability	-.01	.935	1	.201	.484	.439	.015	.080
Difference in Tomato Suitability	.391	.206	.201	1	.280	.080	.163	.180
Difference in Temperature	.016	.526	.484	.280	1	.387	.025	.096
Latitude Difference	.168	.364	.439	.080	.387	1	.024	.112
Longitude Difference	.586	.026	.015	.163	.025	.024	1	.534
Religious Distance	.328	.118	.080	.180	.096	.112	.534	1

TABLE A3. Main results: All coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
			Correlation: Wheat				*Correlation: Salt*				*Correlation: Rice*	
Linguistic Distance	-0.257***	-0.210***	-0.023	-0.067**	-0.484***	-0.392***	-0.384***	-0.189***	-0.083***	-0.073***	-0.056***	-0.035***
	(0.035)	(0.036)	(0.025)	(0.030)	(0.061)	(0.072)	(0.051)	(0.044)	(0.017)	(0.010)	(0.018)	(0.010)
Ln Distance in KM			-0.001	-0.008			-0.065**	-0.014			-0.041***	-0.017***
			(0.013)	(0.010)			(0.026)	(0.024)			(0.012)	(0.005)
Same Province			-0.011	0.018*			0.118***	0.105***			0.050***	0.031***
			(0.013)	(0.010)			(0.027)	(0.018)			(0.017)	(0.006)
Both Coastal			-0.014	0.003			0.052	0.080**			-0.040	-0.006
			(0.022)	(0.016)			(0.033)	(0.038)			(0.032)	(0.008)
Same River			0.033**	0.017*			0.019	-0.005			0.014	0.007
			(0.016)	(0.009)			(0.018)	(0.012)			(0.010)	(0.005)
Rainfall Correlation			-0.003	-0.020			-0.110**	0.050			-0.112***	-0.015
			(0.025)	(0.018)			(0.052)	(0.037)			(0.039)	(0.010)
D Land Quality			0.018	-0.009			0.087*	0.019			0.053**	-0.004
			(0.026)	(0.017)			(0.046)	(0.029)			(0.023)	(0.008)
D Ruggedness			0.000	-0.000			-0.000**	-0.000**			-0.000**	-0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Malaria			0.005*	0.003			-0.019***	-0.020***			-0.000	0.001
			(0.003)	(0.002)			(0.006)	(0.006)			(0.002)	(0.001)
D Humidity			0.002***	0.002***			-0.006***	0.001			-0.001	-0.001*
			(0.001)	(0.001)			(0.002)	(0.001)			(0.001)	(0.000)
D Altitude			-0.000***	-0.000			0.000***	0.000***			0.000	-0.000**
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Banana Suit			0.000***	0.000			0.000	-0.000			0.000	0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Chickpea Suit			-0.000	0.000			0.000	0.000			0.000	0.000***
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Cocoa Suit			-0.000*	0.000			-0.000	-0.000			0.000*	0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Cotton Suit			-0.000	0.000			0.001	-0.000			0.001***	0.000
			(0.000)	(0.000)			(0.001)	(0.000)			(0.000)	(0.000)
D Groundnut Suit			0.000	0.000**			-0.000	-0.000			0.000	0.000*
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Dry Rice Suit			-0.000	0.000**			0.000	0.000			-0.000	0.000**
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Oil Palm Suit			-0.000	-0.000**			0.000**	0.000			-0.000	-0.000***
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Onion Suit			-0.000	-0.000**			0.000	0.000			-0.000**	-0.000*
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Precipitation			-0.000	-0.000			-0.000***	-0.000*			-0.000	-0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Slope			-0.000	-0.000			0.000	0.000			0.000***	0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Soybean Suit			-0.000	-0.000			-0.000	0.000			0.000	0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Sugar Suit			-0.000	0.000			0.000***	0.000*			0.000***	0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Tea Suit			-0.000***	-0.000*			0.001***	0.000*			-0.000**	-0.000**
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Wetland Rice Suit			0.000	0.000			-0.000	-0.000			0.000	0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D White Potato Suit			0.000	-0.000**			0.000	-0.000***			0.000	-0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Wheat Suit			-0.000	-0.000**			-0.000***	-0.000			-0.000	-0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Tomato Suit			0.000	0.000**			-0.000*	0.000***			-0.000**	-0.000
			(0.000)	(0.000)			(0.000)	(0.000)			(0.000)	(0.000)
D Temperature			-0.003	0.005			0.045***	0.016**			0.005	0.004***
			(0.005)	(0.003)			(0.010)	(0.007)			(0.004)	(0.001)
Latitude Difference			-0.001	-0.006***			0.016***	0.008*			0.004**	0.001*
			(0.002)	(0.002)			(0.006)	(0.004)			(0.002)	(0.001)
Longitude Difference			-0.009***	-0.005***			-0.007***	-0.004***			-0.005***	-0.003***
			(0.002)	(0.001)			(0.003)	(0.002)			(0.001)	(0.001)
Religious Distance			-0.051**	-0.009			-0.196***	-0.238***			-0.055**	-0.006
			(0.023)	(0.020)			(0.071)	(0.055)			(0.026)	(0.012)
N	15,652	15,652	15,652	15,652	20,909	20,909	20,909	20,909	20,909	20,909	20,909	20,909
R-squared	0.139	0.762	0.580	0.806	0.216	0.708	0.566	0.791	0.045	0.834	0.282	0.868
FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Controls	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A4. Comparing linguistic and physical distance

	(1)	(2)	(3)
		Correlation: Wheat	
Linguistic Distance	-0.257***		-0.185***
	(0.035)		(0.037)
Ln Distance in KM		-0.114***	-0.080***
		(0.010)	(0.008)
N	15,652	15,652	15,652
Rsq	0.139	0.134	0.195
		Correlation: Salt	
Linguistic Distance	-0.484***		-0.346***
	(0.061)		(0.067)
Ln Distance in KM		-0.250***	-0.152***
		(0.022)	(0.021)
N	20,909	20,909	20,909
Rsq	0.216	0.184	0.266
		Correlation: Rice	
Linguistic Distance	-0.083***		-0.034*
	(0.017)		(0.019)
Ln Distance in KM		-0.064***	-0.054***
		(0.006)	(0.006)
N	20,909	20,909	20,909
Rsq	0.045	0.084	0.089
Fixed Effects	No	No	No
Controls	No	No	No

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A5. Restrict market pairs to districts where the major language is Indo-European

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.168**	-0.751***	-0.005	-0.050
	(0.084)	(0.141)	(0.035)	(0.068)
N	12,364	12,364	12,364	12,364
Rsq	0.011	0.773	0.536	0.810
		Correlation: Salt		
Linguistic Distance	-0.127	-1.057***	-0.032	0.133
	(0.136)	(0.196)	(0.079)	(0.163)
N	12,719	12,719	12,719	12,719
Rsq	0.002	0.753	0.431	0.808
		Correlation: Rice		
Linguistic Distance	-0.210*	-0.599***	0.004	-0.067
	(0.111)	(0.107)	(0.049)	(0.042)
N	12,719	12,719	12,719	12,719
Rsq	0.023	0.848	0.245	0.892
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A6. Control for cost distance

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.257***	-0.210***	-0.023	-0.067**
	(0.035)	(0.036)	(0.025)	(0.031)
N	15,652	15,652	15,652	15,652
Rsq	0.139	0.762	0.580	0.806
		Correlation: Salt		
Linguistic Distance	-0.484***	-0.392***	-0.383***	-0.190***
	(0.061)	(0.072)	(0.052)	(0.045)
N	20,909	20,909	20,909	20,909
Rsq	0.216	0.708	0.566	0.792
		Correlation: Rice		
Linguistic Distance	-0.083***	-0.073***	-0.055***	-0.032***
	(0.017)	(0.010)	(0.018)	(0.010)
N	20,909	20,909	20,909	20,909
Rsq	0.045	0.834	0.282	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A7. Other crops

	(1)	(2)	(3)	(4)
		Correlation: Arhar Dal		
Linguistic Distance	-0.159**	-0.098***	-0.123	-0.053***
	(0.077)	(0.016)	(0.081)	(0.016)
N	11,628	11,628	11,628	11,628
Rsq	0.077	0.920	0.410	0.928
		Correlation: Bajra		
Linguistic Distance	-0.113***	-0.152***	-0.057**	-0.078***
	(0.023)	(0.020)	(0.026)	(0.020)
N	6,097	6,097	6,097	6,097
Rsq	0.079	0.838	0.585	0.890
		Correlation: Barley		
Linguistic Distance	-0.237*	-0.357**	-0.216**	-0.092
	(0.133)	(0.175)	(0.085)	(0.099)
N	5,465	5,465	5,465	5,465
Rsq	0.022	0.784	0.688	0.841
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A8. Other crops

	(1)	(2)	(3)	(4)
		Correlation: Gram		
Linguistic Distance	-0.204***	-0.102***	-0.149***	-0.053**
	(0.034)	(0.014)	(0.022)	(0.022)
N	16,470	16,470	16,470	16,470
Rsq	0.223	0.816	0.672	0.868
		Correlation: Jawar		
Linguistic Distance	-0.184***	-0.155***	-0.036*	-0.075***
	(0.045)	(0.014)	(0.020)	(0.014)
N	8,001	8,001	8,001	8,001
Rsq	0.194	0.800	0.652	0.841
		Correlation: Kangni		
Linguistic Distance	-0.520	-0.004	-0.799*	0.218
	(0.714)	(0.337)	(0.469)	(0.283)
N	1,275	1,275	1,275	1,275
Rsq	0.003	0.594	0.340	0.645
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A9. Other crops

	(1)	(2)	(3)	(4)
		Correlation: Maize		
Linguistic Distance	-0.503***	-0.285***	0.009	-0.003
	(0.049)	(0.059)	(0.079)	(0.052)
N	2,850	2,850	2,850	2,850
Rsq	0.433	0.919	0.609	0.944
		Correlation: Marua		
Linguistic Distance	-0.054	-0.139***	0.034	0.002
	(0.043)	(0.030)	(0.028)	(0.025)
N	1,275	1,275	1,275	1,275
Rsq	0.008	0.796	0.671	0.857
		Correlation: Bulrush Millet		
Linguistic Distance	-0.295***	-0.462***	0.084*	0.016
	(0.054)	(0.046)	(0.049)	(0.047)
N	855	855	855	855
Rsq	0.160	0.559	0.586	0.730
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A10. Other crops

	(1)	(2)	(3)	(4)
		Correlation: Great Millet		
Linguistic Distance	-0.115*	-0.343***	0.231***	0.118
	(0.059)	(0.053)	(0.070)	(0.079)
N	1,228	1,228	1,228	1,228
Rsq	0.018	0.576	0.570	0.706
		Correlation: Lesser Millet		
Linguistic Distance	-0.520***	-0.533***	-0.264***	-0.225***
	(0.125)	(0.103)	(0.102)	(0.085)
N	253	253	253	253
Rsq	0.213	0.686	0.592	0.826
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A11. Restrict sample to present-day India

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.268***	-0.217***	-0.044*	-0.074**
	(0.033)	(0.038)	(0.026)	(0.032)
N	10,854	10,854	10,854	10,854
Rsq	0.203	0.792	0.553	0.853
		Correlation: Salt		
Linguistic Distance	-0.178***	-0.223***	-0.145***	-0.074**
	(0.041)	(0.037)	(0.046)	(0.036)
N	13,040	13,040	13,040	13,040
Rsq	0.055	0.585	0.454	0.729
		Correlation: Rice		
Linguistic Distance	-0.010	-0.053***	-0.000	-0.012**
	(0.014)	(0.006)	(0.019)	(0.006)
N	13,040	13,040	13,040	13,040
Rsq	0.001	0.877	0.241	0.908
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A12. No negative correlations

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.243***	-0.207***	-0.028	-0.066**
	(0.031)	(0.035)	(0.024)	(0.029)
N	15,479	15,479	15,479	15,479
Rsq	0.160	0.770	0.592	0.825
		Correlation: Salt		
Linguistic Distance	-0.269***	-0.255***	-0.255***	-0.118***
	(0.033)	(0.030)	(0.040)	(0.031)
N	18,211	18,211	18,211	18,211
Rsq	0.148	0.586	0.382	0.696
		Correlation: Rice		
Linguistic Distance	-0.089***	-0.073***	-0.061***	-0.035***
	(0.017)	(0.010)	(0.018)	(0.010)
N	20,768	20,768	20,768	20,768
Rsq	0.063	0.799	0.338	0.842
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A13. Remove outliers by price correlation

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.191***	-0.178***	-0.020	-0.042
	(0.024)	(0.028)	(0.021)	(0.026)
N	14,243	14,243	14,243	14,243
Rsq	0.161	0.718	0.633	0.799
		Correlation: Salt		
Linguistic Distance	-0.362***	-0.310***	-0.370***	-0.167***
	(0.048)	(0.055)	(0.045)	(0.040)
N	19,027	19,027	19,027	19,027
Rsq	0.161	0.647	0.482	0.741
		Correlation: Rice		
Linguistic Distance	-0.077***	-0.070***	-0.059***	-0.036***
	(0.014)	(0.010)	(0.015)	(0.009)
N	19,027	19,027	19,027	19,027
Rsq	0.086	0.765	0.373	0.823
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A14. Remove outliers by linguistic distance

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.230***	-0.204***	-0.035	-0.066**
	(0.038)	(0.035)	(0.025)	(0.030)
N	14,586	14,586	14,586	14,586
Rsq	0.108	0.763	0.577	0.809
		Correlation: Salt		
Linguistic Distance	-0.417***	-0.370***	-0.377***	-0.201***
	(0.065)	(0.072)	(0.054)	(0.048)
N	19,015	19,015	19,015	19,015
Rsq	0.161	0.703	0.527	0.785
		Correlation: Rice		
Linguistic Distance	-0.072***	-0.077***	-0.055***	-0.036***
	(0.019)	(0.011)	(0.019)	(0.010)
N	19,015	19,015	19,015	19,015
Rsq	0.030	0.836	0.267	0.872
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A15. Remove market pairs with fewer than 10 common observations

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.261***	-0.210***	-0.021	-0.070**
	(0.035)	(0.036)	(0.024)	(0.030)
N	15,494	15,494	15,494	15,494
Rsq	0.155	0.787	0.592	0.834
		Correlation: Salt		
Linguistic Distance	-0.484***	-0.392***	-0.384***	-0.189***
	(0.061)	(0.072)	(0.051)	(0.044)
N	20,907	20,907	20,907	20,907
Rsq	0.216	0.709	0.566	0.791
		Correlation: Rice		
Linguistic Distance	-0.083***	-0.073***	-0.056***	-0.035***
	(0.017)	(0.010)	(0.018)	(0.010)
N	20,907	20,907	20,907	20,907
Rsq	0.045	0.836	0.283	0.870
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A16. Drop cities above 75,000

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.265***	-0.219***	-0.013	-0.081**
	(0.036)	(0.040)	(0.028)	(0.035)
N	10,929	10,929	10,929	10,929
Rsq	0.138	0.758	0.568	0.801
		Correlation: Salt		
Linguistic Distance	-0.493***	-0.398***	-0.383***	-0.203***
	(0.066)	(0.078)	(0.055)	(0.045)
N	15,051	15,051	15,051	15,051
Rsq	0.219	0.712	0.560	0.789
		Correlation: Rice		
Linguistic Distance	-0.094***	-0.076***	-0.068***	-0.042***
	(0.017)	(0.011)	(0.018)	(0.010)
N	15,051	15,051	15,051	15,051
Rsq	0.085	0.782	0.318	0.833
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A17. Drop coastal

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.238***	-0.216***	-0.037	-0.074***
	(0.037)	(0.035)	(0.028)	(0.027)
N	11,895	11,895	11,895	11,895
Rsq	0.154	0.779	0.509	0.830
		Correlation: Salt		
Linguistic Distance	-0.431***	-0.370***	-0.381***	-0.228***
	(0.069)	(0.077)	(0.070)	(0.055)
N	14,195	14,195	14,195	14,195
Rsq	0.181	0.740	0.505	0.797
		Correlation: Rice		
Linguistic Distance	-0.110***	-0.088***	-0.073***	-0.052***
	(0.023)	(0.014)	(0.025)	(0.015)
N	14,195	14,195	14,195	14,195
Rsq	0.091	0.816	0.372	0.848
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A18. Drop Gangetic

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.250***	-0.171***	0.001	-0.035
	(0.035)	(0.036)	(0.029)	(0.027)
N	10,362	10,362	10,362	10,362
Rsq	0.148	0.789	0.578	0.834
		Correlation: Salt		
Linguistic Distance	-0.445***	-0.372***	-0.327***	-0.167***
	(0.063)	(0.074)	(0.055)	(0.045)
N	14,705	14,705	14,705	14,705
Rsq	0.178	0.651	0.548	0.756
		Correlation: Rice		
Linguistic Distance	-0.078***	-0.075***	-0.044**	-0.031***
	(0.017)	(0.010)	(0.019)	(0.009)
N	14,705	14,705	14,705	14,705
Rsq	0.036	0.841	0.263	0.871
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A19. Prices before 1891

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.236***	-0.264***	-0.090	-0.032
	(0.049)	(0.043)	(0.068)	(0.051)
N	15,165	15,165	15,165	15,165
Rsq	0.075	0.567	0.329	0.654
		Correlation: Salt		
Linguistic Distance	-0.490***	-0.672***	-0.392***	-0.261***
	(0.081)	(0.090)	(0.080)	(0.082)
N	19,701	19,701	19,701	19,701
Rsq	0.112	0.430	0.352	0.597
		Correlation: Rice		
Linguistic Distance	-0.158***	-0.229***	-0.077**	-0.067*
	(0.024)	(0.028)	(0.032)	(0.038)
N	19,697	19,697	19,697	19,697
Rsq	0.049	0.401	0.258	0.504
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A20. Prices after 1891

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.081***	-0.148***	-0.058***	-0.039**
	(0.015)	(0.025)	(0.014)	(0.020)
N	13,690	13,690	13,690	13,690
Rsq	0.037	0.733	0.622	0.799
		Correlation: Salt		
Linguistic Distance	-0.344***	-0.195***	-0.213***	-0.091***
	(0.047)	(0.060)	(0.036)	(0.023)
N	20,908	20,908	20,908	20,908
Rsq	0.200	0.789	0.613	0.863
		Correlation: Rice		
Linguistic Distance	-0.079***	-0.070***	-0.066***	-0.038***
	(0.017)	(0.013)	(0.016)	(0.009)
N	20,909	20,909	20,909	20,909
Rsq	0.039	0.879	0.261	0.902
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A21. Drop pairs within 500km

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.238***	-0.151***	-0.021	-0.042
	(0.036)	(0.037)	(0.028)	(0.032)
N	12,681	12,681	12,681	12,681
Rsq	0.125	0.771	0.576	0.807
		Correlation: Salt		
Linguistic Distance	-0.454***	-0.255***	-0.404***	-0.112**
	(0.065)	(0.064)	(0.052)	(0.047)
N	17,552	17,552	17,552	17,552
Rsq	0.189	0.732	0.561	0.801
		Correlation: Rice		
Linguistic Distance	-0.063***	-0.044***	-0.051***	-0.022**
	(0.018)	(0.010)	(0.020)	(0.010)
N	17,552	17,552	17,552	17,552
Rsq	0.026	0.845	0.271	0.867
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A22. Log $1 + \rho$ as outcome

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.158***	-0.127***	-0.012	-0.046**
	(0.024)	(0.024)	(0.016)	(0.020)
N	15,648	15,648	15,648	15,648
Rsq	0.100	0.638	0.462	0.668
		Correlation: Salt		
Linguistic Distance	-0.393***	-0.310***	-0.295***	-0.138***
	(0.057)	(0.069)	(0.043)	(0.037)
N	20,909	20,909	20,909	20,909
Rsq	0.189	0.706	0.559	0.780
		Correlation: Rice		
Linguistic Distance	-0.046***	-0.041***	-0.031***	-0.020***
	(0.010)	(0.006)	(0.011)	(0.006)
N	20,909	20,909	20,909	20,909
Rsq	0.033	0.844	0.244	0.871
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A23. Centiles of ρ as outcome

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-40.570***	-32.937***	-6.507	-2.753
	(3.538)	(4.127)	(4.277)	(3.325)
N	15,652	15,652	15,652	15,652
Rsq	0.197	0.790	0.663	0.885
		Correlation: Salt		
Linguistic Distance	-35.701***	-31.079***	-27.586***	-13.313***
	(3.554)	(3.882)	(3.822)	(3.025)
N	20,909	20,909	20,909	20,909
Rsq	0.237	0.664	0.532	0.773
		Correlation: Rice		
Linguistic Distance	-23.418***	-20.859***	-13.190***	-6.610***
	(3.138)	(1.876)	(3.631)	(1.879)
N	20,909	20,909	20,909	20,909
Rsq	0.102	0.750	0.400	0.827
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A24. $\delta = 0.5$

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.333***	-0.189***	-0.030	-0.037**
	(0.039)	(0.025)	(0.022)	(0.019)
N	15,652	15,652	15,652	15,652
Rsq	0.133	0.764	0.580	0.805
		Correlation: Salt		
Linguistic Distance	-0.723***	-0.515***	-0.435***	-0.137***
	(0.072)	(0.075)	(0.056)	(0.042)
N	20,909	20,909	20,909	20,909
Rsq	0.222	0.713	0.545	0.788
		Correlation: Rice		
Linguistic Distance	-0.148***	-0.116***	-0.087***	-0.042***
	(0.020)	(0.012)	(0.020)	(0.012)
N	20,909	20,909	20,909	20,909
Rsq	0.065	0.840	0.283	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A25. Measure distance using largest language

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Distance by largest language	-0.206***	-0.141***	-0.038*	-0.047**
	(0.035)	(0.027)	(0.020)	(0.020)
N	15,652	15,652	15,652	15,652
Rsq	0.128	0.759	0.581	0.806
		Correlation: Salt		
Distance by largest language	-0.415***	-0.303***	-0.302***	-0.135***
	(0.054)	(0.061)	(0.046)	(0.040)
N	20,909	20,909	20,909	20,909
Rsq	0.210	0.704	0.560	0.790
		Correlation: Rice		
Distance by largest language	-0.064***	-0.055***	-0.045***	-0.023***
	(0.014)	(0.009)	(0.014)	(0.008)
N	20,909	20,909	20,909	20,909
Rsq	0.035	0.833	0.281	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A26. Measure distance as dummy for different largest language

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Different Language	-0.123***	-0.071***	-0.017***	-0.011*
	(0.015)	(0.010)	(0.006)	(0.006)
N	15,652	15,652	15,652	15,652
Rsq	0.033	0.758	0.580	0.805
		Correlation: Salt		
Different Language	-0.332***	-0.206***	-0.028	0.010
	(0.033)	(0.037)	(0.022)	(0.018)
N	20,909	20,909	20,909	20,909
Rsq	0.057	0.688	0.514	0.787
		Correlation: Rice		
Different Language	-0.104***	-0.057***	-0.031***	-0.013***
	(0.009)	(0.008)	(0.010)	(0.004)
N	20,909	20,909	20,909	20,909
Rsq	0.039	0.834	0.276	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A27. Linguistic distance squared

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.191	-0.400***	-0.190**	-0.013
	(0.164)	(0.130)	(0.079)	(0.078)
Squared	-0.068	0.190	0.184**	-0.053
	(0.160)	(0.123)	(0.090)	(0.071)
N	15,652	15,652	15,652	15,652
Rsq	0.140	0.762	0.582	0.806
		Correlation: Salt		
Linguistic Distance	0.182	-0.024	0.144	0.327
	(0.268)	(0.255)	(0.187)	(0.204)
Squared	-0.649**	-0.350	-0.524***	-0.485**
	(0.263)	(0.260)	(0.190)	(0.192)
N	20,909	20,909	20,909	20,909
Rsq	0.227	0.709	0.572	0.792
		Correlation: Rice		
Linguistic Distance	-0.139	-0.074	-0.005	0.023
	(0.118)	(0.057)	(0.067)	(0.032)
Squared	0.054	0.000	-0.051	-0.055*
	(0.115)	(0.057)	(0.066)	(0.031)
N	20,909	20,909	20,909	20,909
Rsq	0.045	0.834	0.282	0.869
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A28. Log linguistic distance variable

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
ln Distance	-0.047***	-0.018***	-0.007**	-0.000
	(0.006)	(0.004)	(0.003)	(0.002)
N	15,652	15,652	15,652	15,652
Rsq	0.118	0.756	0.581	0.805
		Correlation: Salt		
ln Distance	-0.102***	-0.072***	-0.047***	-0.014***
	(0.012)	(0.011)	(0.008)	(0.005)
N	20,909	20,909	20,909	20,909
Rsq	0.161	0.699	0.530	0.788
		Correlation: Rice		
ln Distance	-0.022***	-0.012***	-0.009***	-0.000
	(0.003)	(0.002)	(0.003)	(0.001)
N	20,909	20,909	20,909	20,909
Rsq	0.055	0.831	0.279	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A29. Log-log specification

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
ln Distance	-0.074***	-0.028***	-0.012	-0.003
	(0.011)	(0.007)	(0.007)	(0.004)
N	15,479	15,479	15,479	15,479
Rsq	0.061	0.668	0.338	0.697
		Correlation: Salt		
ln Distance	-0.110***	-0.088***	-0.073***	-0.011
	(0.014)	(0.013)	(0.013)	(0.010)
N	18,211	18,211	18,211	18,211
Rsq	0.060	0.459	0.246	0.543
		Correlation: Rice		
ln Distance	-0.030***	-0.015***	-0.013***	-0.000
	(0.005)	(0.003)	(0.004)	(0.002)
N	20,768	20,768	20,768	20,768
Rsq	0.025	0.783	0.164	0.802
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A30. Cladistic Distance from Glottolog

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Glottolog Distance	-0.235***	-0.083***	-0.076**	-0.018
	(0.033)	(0.019)	(0.035)	(0.012)
N	15,652	15,652	15,652	15,652
Rsq	0.177	0.755	0.589	0.805
		Correlation: Salt		
Glottolog Distance	-0.322***	-0.317***	-0.075**	-0.126***
	(0.058)	(0.062)	(0.032)	(0.029)
N	20,909	20,909	20,909	20,909
Rsq	0.095	0.700	0.516	0.789
		Correlation: Rice		
Glottolog Distance	-0.072***	-0.076***	-0.009	-0.015**
	(0.016)	(0.009)	(0.011)	(0.006)
N	20,909	20,909	20,909	20,909
Rsq	0.033	0.837	0.274	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A31. Lexicostatistical Distance from ASJP

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Lexicostatistical Distance	-0.488***	-0.187***	-0.117**	-0.016
	(0.057)	(0.022)	(0.051)	(0.017)
N	15,652	15,652	15,652	15,652
Rsq	0.125	0.759	0.583	0.805
		Correlation: Salt		
Lexicostatistical Distance	-0.891***	-0.685***	-0.154**	-0.159***
	(0.116)	(0.108)	(0.068)	(0.059)
N	20,909	20,909	20,909	20,909
Rsq	0.111	0.710	0.515	0.788
		Correlation: Rice		
Lexicostatistical Distance	-0.172***	-0.159***	0.035	-0.045***
	(0.035)	(0.017)	(0.026)	(0.012)
N	20,909	20,909	20,909	20,909
Rsq	0.029	0.840	0.275	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A32. Cluster by largest ethnic group

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.257***	-0.210***	-0.023	-0.067**
	(0.041)	(0.038)	(0.035)	(0.031)
N	15,652	15,652	15,652	15,652
Rsq	0.139	0.762	0.580	0.806
		Correlation: Salt		
Linguistic Distance	-0.484***	-0.392***	-0.384***	-0.189**
	(0.128)	(0.149)	(0.076)	(0.074)
N	20,909	20,909	20,909	20,909
Rsq	0.216	0.708	0.566	0.791
		Correlation: Rice		
Linguistic Distance	-0.083***	-0.073***	-0.056***	-0.035**
	(0.032)	(0.018)	(0.019)	(0.016)
N	20,909	20,909	20,909	20,909
Rsq	0.045	0.834	0.282	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by largest ethnic groups in market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. . Fixed effects are for market i and j.

TABLE A33. Cluster by province

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.257***	-0.210***	-0.023*	-0.067**
	(0.046)	(0.043)	(0.012)	(0.032)
N	15,652	15,652	15,652	15,652
Rsq	0.139	0.762	0.580	0.806
		Correlation: Salt		
Linguistic Distance	-0.484***	-0.392**	-0.384***	-0.189**
	(0.173)	(0.178)	(0.084)	(0.094)
N	20,909	20,909	20,909	20,909
Rsq	0.216	0.708	0.566	0.791
		Correlation: Rice		
Linguistic Distance	-0.083**	-0.073***	-0.056***	-0.035*
	(0.038)	(0.023)	(0.016)	(0.018)
N	20,909	20,909	20,909	20,909
Rsq	0.045	0.834	0.282	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by provinces of market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A34. Conley Standard Errors

Crop	Column	Coefficient	Standard Error	$p < 0.05$
Wheat	1	-.2567	(.0276)	*
	2	-.1611	(.0368)	*
	3	-.0228	(.0260)	
	4	-.0667	(.0313)	*
Salt	1	-.4840	(.0521)	*
	2	-.3917	(.0943)	*
	3	-.3842	(.0504)	*
	4	-.1889	(.0579)	*
Rice	1	-.0833	(.0116)	*
	2	-.0731	(.0114)	*
	3	-.0560	(.0115)	*
	4	-.0347	(.0090)	*

This table reports results analogous to those in Table 2, but with Conley standard errors accounting for spatial correlation in the error term at distances up to five decimal degrees. The "Crop" column indicates which crop's correlation coefficient is being used as an outcome variable. "Column" indicates the corresponding column in Table 2. "Coefficient" is the corresponding coefficient estimate. "Standard error" is the corresponding standard error. Coefficients that are statistically significant at the 5% level are indicated with an asterisk.

TABLE A35. Control for mean absolute log difference

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.084***	-0.112***	-0.004	-0.047*
	(0.032)	(0.030)	(0.025)	(0.027)
N	15,652	15,652	15,652	15,652
Rsq	0.287	0.783	0.594	0.814
		Correlation: Salt		
Linguistic Distance	-0.347***	-0.254***	-0.301***	-0.130***
	(0.059)	(0.049)	(0.048)	(0.035)
N	20,909	20,909	20,909	20,909
Rsq	0.485	0.804	0.663	0.837
		Correlation: Rice		
Linguistic Distance	-0.086***	-0.057***	-0.072***	-0.036***
	(0.015)	(0.008)	(0.016)	(0.010)
N	20,909	20,909	20,909	20,909
Rsq	0.222	0.859	0.378	0.873
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A36. Interact linguistic and physical distance

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.182***	-0.114***	-0.031	-0.068**
	(0.039)	(0.035)	(0.024)	(0.032)
Normalized ln Distance	-0.056***	-0.033***	-0.000	-0.006
	(0.006)	(0.005)	(0.009)	(0.007)
Interaction	-0.006	-0.028	0.032	0.003
	(0.032)	(0.023)	(0.022)	(0.021)
N	15,652	15,652	15,652	15,652
Rsq	0.195	0.777	0.581	0.806
		Correlation: Salt		
Linguistic Distance	-0.273***	-0.214***	-0.357***	-0.182***
	(0.068)	(0.067)	(0.049)	(0.045)
Normalized ln Distance	-0.067***	-0.073***	-0.043**	-0.009
	(0.018)	(0.013)	(0.019)	(0.018)
Interaction	-0.157***	-0.048	-0.110***	-0.018
	(0.040)	(0.031)	(0.037)	(0.025)
N	20,909	20,909	20,909	20,909
Rsq	0.279	0.729	0.570	0.791
		Correlation: Rice		
Linguistic Distance	-0.037*	-0.025**	-0.058***	-0.044***
	(0.020)	(0.013)	(0.019)	(0.012)
Normalized ln Distance	-0.040***	-0.031***	-0.029***	-0.014***
	(0.005)	(0.004)	(0.008)	(0.004)
Interaction	0.007	0.008	0.007	0.028**
	(0.013)	(0.010)	(0.013)	(0.011)
N	20,909	20,909	20,909	20,909
Rsq	0.090	0.852	0.282	0.870
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A37. Mughal History

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.225***	-0.196***	-0.014	-0.070**
	(0.035)	(0.038)	(0.028)	(0.031)
Both Mughal 1605	0.003	0.034	0.032***	0.018
	(0.018)	(0.021)	(0.012)	(0.015)
Both Mughal 1707	0.060	-0.084	-0.024	-0.113*
	(0.046)	(0.074)	(0.036)	(0.065)
N	15,652	15,652	15,652	15,652
Rsq	0.146	0.762	0.581	0.806
		Correlation: Salt		
Linguistic Distance	-0.340***	-0.261***	-0.212***	-0.119***
	(0.067)	(0.063)	(0.048)	(0.044)
Both Mughal 1605	-0.053	0.083	0.089***	0.111**
	(0.050)	(0.070)	(0.032)	(0.052)
Both Mughal 1707	0.325***	0.407***	0.237***	0.172*
	(0.060)	(0.134)	(0.048)	(0.089)
N	20,909	20,909	20,909	20,909
Rsq	0.290	0.719	0.592	0.794
		Correlation: Rice		
Linguistic Distance	-0.001	-0.064***	0.003	-0.030***
	(0.021)	(0.010)	(0.019)	(0.009)
Both Mughal 1605	0.030	-0.009	0.035**	0.006
	(0.022)	(0.010)	(0.017)	(0.010)
Both Mughal 1707	0.090***	0.070***	0.073***	0.015
	(0.020)	(0.019)	(0.022)	(0.018)
N	20,909	20,909	20,909	20,909
Rsq	0.097	0.836	0.301	0.869
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A38. Use religious distance from 1901 census

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.257***	-0.210***	-0.023	-0.067**
	(0.035)	(0.036)	(0.025)	(0.030)
N	15,652	15,652	15,652	15,652
Rsq	0.139	0.762	0.579	0.806
		Correlation: Salt		
Linguistic Distance	-0.484***	-0.392***	-0.387***	-0.180***
	(0.061)	(0.072)	(0.050)	(0.043)
N	20,909	20,909	20,909	20,909
Rsq	0.216	0.708	0.562	0.791
		Correlation: Rice		
Linguistic Distance	-0.083***	-0.073***	-0.053***	-0.035***
	(0.017)	(0.010)	(0.019)	(0.010)
N	20,909	20,909	20,909	20,909
Rsq	0.045	0.834	0.282	0.868
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, ln(*distance*) in km, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.