Dialects, Cultural Identity, and Economic Exchange

Oliver Falck[+], Stephan Heblich[*], Alfred Lameli[‡], Jens Südekum[†]

October 2009

Preliminary and incomplete version

[+] Ifo Institute for Economic Research, Poschingerstr. 5, D-81679 Munich (Germany), Phone: +49 89 9224 1370, Fax: +49 89 9224 1460, Email: falck@ifo.de, CESifo and Max Planck Institute of Economics.

[*] Max Planck Institute of Economics, Entrepreneurship, Growth, and Public Policy Group, Kahlaischestr. 10, D-07745 Jena (Germany), Phone: +49 3641 686 733, Fax: +49 3641 686 710, Email: heblich@econ.mpg.de.

[‡] Research Centre *Deutscher Sprachatlas*, Hermann-Jacobsohn-Weg 3, D-35032 Marburg (Germany), Phone: +49 6421 28 22 482, Fax: +49 6421 28 28 936, Email: lameli@staff.uni-marburg.de.

[†] Mercator School of Management, University of Duisburg-Essen, Lotharstraße 65, D-47057 Duisburg (Germany), Phone: +49 203 379 2357, Fax: +49 203 379-1786, Email: jens.suedekum@uni-due.de

Dialects, Cultural Identity, and Economic Exchange

Abstract

Recent empirical research finds that there are intangible cultural borders that impede economic exchange across countries. In this paper, we investigate whether time-persistent cultural borders also exist at a finer geographical level, namely, across regions of the same country. To distinguish regional cultures, we utilize, for the first time in the economics literature, detailed linguistic micro-data about phonological and grammatical features of German dialects. These data are taken from a unique survey that was conducted between 1879 and 1888 by the linguist Georg Wenker in about 45,000 schools across the German Empire. Matching this information to the 439 German NUTS3 regions, we construct a dialect similarity matrix and analyze current pair-wise gross migration flows in a gravity analysis. Our central finding is that current regional migration is significantly positively affected by the linguistic similarity of dialects that were prevalent in the source and destination areas in the late 19th century. This finding, which is robust in a variety of specifications, suggests that cultural identity at the local level has long-lasting effects and that cultural identities formed more than a century ago continue to influence economic behavior today.

JEL Codes: R11; R23; Z10

Keywords: Language; Local Dialects; Culture; Migration; Germany

Acknowledgments: We thank Gilles Duranton, Mario Larch, Jost Nickel†, Klaus Schmidt, Matthew Turner for insightful comments and suggestions. All errors and shortcomings are solely our responsibility.

1. Introduction

Cultural similarities and a common language are essential for building trust and engaging in economic exchange. When individuals share some common background and are able to coordinate their behavior by using the same set of symbols and vocal expressions, they can more easily develop relationships and conduct transactions than would be the case in the absence of such commonalities (Lazear 1999).[1] This link between language, culture, and economic activity is confirmed by recent empirical research. For example, Guiso *et al.* (2009) show that common cultural and linguistic roots enhance trust between countries, which in turn boosts trade and investment. Conversely, there are intangible borders between culturally distant nations that impede economic exchange.

In this paper, we investigate whether time-persistent cultural borders also exist at a finer geographical level, namely, across regions of the same country. To distinguish regional cultures, we utilize, for the first time in the economics literature, detailed linguistic micro-data about the intra-national variation of phonological and grammatical attributes within the same language (German). We then study the effects of historic dialect similarities on current cross-regional migration flows in a gravity analysis.

Nations are by no means monolithic linguistically. Often, one language can have hundreds of dialects, all substantially different (Chambers & Trudgill 1998). These dialects reflect the everyday experience of individuals living in different regions of the country and strongly shape their cultural identity. For example, people often do not communicate in (and sometimes are not even familiar with) the codified standard language of English, but are intimately familiar with and conversant in their particular varieties of it; someone from Boston, say, sounds very different than someone from Texas, for example, but both are speaking English and if they speak to each other, will have a good guess as to where the other is from. This phenomenon is also true of German, Italian, and many other languages, that is, it is fairly easy for natives to guess a person's regional provenance during a

[1] There is an extensive literature in behavioral and experimental economics as well as in sociology and related fields showing that individuals exchange and cooperate more the more they trust each other. See, among others, Chwe (1999), Coleman (1988), Glaeser *et al.* (2002), Knack and Keefer (1997), Sobel (1985), and Watson (1999).

conversation. As we shall argue below, this vast variation in just one language is likely to reflect long-term and highly persistent cultural differences across the local populations.

We use data on German dialects that are taken from a unique language survey that was conducted between 1879 and 1888. By the order of the just established German Empire, the linguist Georg Wenker collected detailed data about the language characteristics of pupils from about 45,000 schools across the Empire. To this day, the Wenker survey is the most complete documentation ever of a nation's language and has defined standards in the linguistics discipline.[2] Based on these data, we construct a dialect similarity matrix between 439 German regions, the current NUTS3 districts (*Landkreise)*. The characterization of each district's dialect is based on 293 phonological and grammatical features, which may thought of as the "micro-foundations" of language. We then analyze pair-wise gross migration flows across German districts over the period 2000–2006. Our central finding is that the current migration between German regions is significantly positively affected by the similarity of dialects prevalent in the source and destination areas in the late 19th century. This result remains robust in a variety of specifications and holds even after controlling for idiosyncratic regional effects, physical distance, and travel time, as well as a host of political and geological regional differences. It implies that an individual who decides to migrate today—all else equal—will choose a destinations with historic dialect characteristics that are similar to those of his or her source region.

What does this finding imply? In this paper, we argue that the local dialects as recorded in the 19th century were shaped by past (i.e., pre-19th century) interactions, including prior mass migration waves, ancient routes, religious and political divisions, and so forth. Almost like a genome, language acts as a sort of memory that stores such information, a point made by anthropologists, including Cavalli-Sforza (2000), who stresses the close resemblance between linguistic and genetic evolution. Local dialect data can thus explain much more than simply phonetic and grammatical variations, as these variations are imprints from the past. We show that the observed linguistic patterns can, in fact, often be traced to cultural and religious congruencies as well as to unique historical events, and

[2] See Lameli (2008) for a detailed introduction to the Wenker survey from a linguistic perspective.

they certainly capture more than just geographical distances. In other words, local languages can be interpreted as a comprehensive measure for the more general concept of local *cultural identity*, and with our linguistic micro-data we are able to account for these local cultural identities to an unusual degree.

German regions with more similar dialects in the late 19th century should therefore be regarded as culturally more closely connected at that time. Our findings then suggest that cultural similarity at the regional level has long-lasting effects and still influences economic behavior (such as individual migration decisions) today. These results are consistent with other research that finds positive effects of cultural similarity, common language, and trust on international trade and other country-pair specific economic and political outcome variables (see Alesina & La Ferrara 2005; Barro & McCleary 2003; Giuliano 2007; Guiso *et al*. 2006, 2009; Melitz 2008; Rauch 1999; Rauch and Trinidade 2002; Tabellini 2007, 2008).[3] Our analysis adds an important dimension to this literature by showing that intangible borders also exist on a much finer geographical scale.

The findings of our study are also related to a few recent contributions that consider the economic effects of *genetic* differences across countries. Spolaore and Wacziarg (2009) find a positive relationship with differences in current income, as populations more closely genetically related are more apt to learn from each other. Desmet *et al.* (2009) show that countries with more distant gene profiles also exhibit stronger cultural differences, which is in line with Guiso *et al.* (2009), who consider both linguistic and somatic determinants of cross-country trust. These papers thus emphasize the relationship between genetic and cultural characteristics, and show that groups that are more closely related *genetically* tend to have closer economic contacts. We obtain a consistent result for *linguistically* (culturally) related groups, even on a finely disaggregated geographical level.

[3] Other research on the economic effects of language similarities focuses more on domestic versus foreign languages. Lazear (1999) develops a model of a multi-cultural society where minorities may or may not assimilate to the official majority tongue. Alesina and La Ferrara (2005) survey the literature on the effects of diversity of foreign languages and ethnicities on the economic performance of the host country. Melitz (2008) studies the effects of common language on international trade flows in a gravity analysis, distinguishing between different modes of communication, and Rauch (1999) considers cultural and language networks in international trade. Our focus in this paper is on the regional variation of the *same* language in the form of dialects. Historically, at the time of data collection, German was the only prevalent language in the Empire and knowledge of foreign languages was extremely limited.

The remainder of this paper is organized as follows. In section 2 we describe our linguistic data and discuss in greater detail the meaning of local dialects, especially in the historical context of our study. Section 3 sets out a simple gravity model for current migration flows that serves as the underlying framework for the empirical analysis. Section 4 presents our estimation results. Section 5 concludes.

2. Background and data

2.1. Genetic, cultural, and linguistic evolution

Anthropologists emphasize the similarities between genetic, cultural, and linguistic evolution. Consider, as an extreme thought experiment, a number of initially identical autarkic populations that are separated and have absolutely no contact with each other. The genetic profile of each population evolves over time as a result of mutation, natural selection, and genetic drift, and the DNA profiles of any two groups are likely to drift apart due to the random elements of evolution. As forcefully argued in Cavalli-Sforza (2000), the same phenomenon is likely to occur for cultures and languages. Isolated populations, even if initially identical, develop idiosyncratic habits and expressions. After the passage of a certain amount of time, it would be difficult for members of the different groups to understand each other if they met. In fact, linguistic evolution would occur much faster and more drastically than genetic evolution, i.e., language differences across groups would become visible earlier and be clearer than DNA differences in this hypothetical scenario where there is no contact between groups. Now add migration to the picture. Cross-border contact of the (now differentiated) populations through migration is the major force behind diffusion and convergence of characteristics. The more often two populations interact, the more diffusion occurs and the more similar these groups will once more become. Linguistic and cultural diffusion (adaption of words, habits, etc.) would be faster and more intensive than genetic diffusion, but would still occur very slowly. Even with very intensive contact, the existing differences between two populations would not disappear any time soon, if at all.

In short, both genes and languages are the product of evolution, the two are likely to be correlated and persistent over time, but linguistic variation is likely to be the more pronounced.[4] In the context of our study, this implies that the local dialects from the late 19th century, just as the DNA profiles of the local populations at that time, were shaped during centuries of previous interaction and mass migrations. Examples of this will be given in Section 2.3. There are no comprehensive data on the DNA profiles of local German populations from the 19th century (nor for any other time), but even if such data were available, they would not necessarily be preferable, precisely because genetic variation is much smaller and may require millenniums before becoming visible. Linguistic variation, on the other hand, may reveal differences across local populations relatively quickly and clearly.

It should also be noted that cultural evolution is not restricted to language, but occurs in many other domains. A key sociological concept in this regard is *cultural identity*. A social group's cultural identity depends on its distinctiveness from other groups, be it differences in technology, art, or social practices like religion, traditions, habits, laws, etc. Once this distinctness is established, delimitation of the group from the rest of society leads to the emergence of norms within the group that have an impact on individual members' behavior (Bernhard *et al.* 2006). Depending on the level of inclusiveness, these borders can enclose groups as large as nations or even continents. A general insight from the sociological literature regarding group size, however, is that the smaller the group, the stronger the degree of identification (Simon 1992). In our context this means that people may more strongly identify themselves as Bavarians, for example, than as Germans; other examples include the Basques in Spain, the Quebecois in Canada, and the Cajuns in the United States.

Even if the particular definition of a group's identity might vary according to specific contexts, there are still clear borders of distinction (Brewer 1991). Language is probably the strongest marker of cultural identity, has the added advantage of being an overt one, and, moreover, is measurable using

[4] This correlation between genes and languages has in fact already been noted by Charles Darwin himself in his seminal book *Origin of Species*: "*If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the languages now spoken around the world; and if all extinct languages, and all intermediate and slowly changing dialect, were to be included, such an arrangement would be the only possible one*" (cited in Cavalli-Sforza 2000:167). Studies dedicated to this correlation include Barbujani *et al.* (1996), Dupanloup de Ceuninck *et al.* (2000), and Manni *et al.* (2006). A detailed overview is provided by Manni (in press).

linguistic techniques. More specifically, a stranger's religious faith, if any, is usually not immediately obvious, whereas his or her speech and inflections are; people can disguise their true norms and values, but not their accent, which is formed during early childhood and is enormously difficult to suppress. Summing up, we cannot think of any better variable to comprehensively capture *cultural identity* at the regional level than the local dialect.

2.2. Historical background and the measurement of linguistic characteristics

Historical background: In the centuries following Charlemagne, France, Spain, England, and Habsburg Austria developed into states where power was wielded by a centralized sovereign. In contrast, the Holy Roman Empire became increasingly fragmented because the emperor had to buy the loyalty of kings, princes, and dukes within the empire by granting territorial and governance concessions. When, in 1648, the Treaty of Westphalia finally ended the Thirty Years' War and, by association, the Holy Roman Empire, what we know as Germany today was comprised of hundreds of sovereign kingdoms, principalities, and dukedoms. This political fragmentation continued until the German Empire (*Deutsches Reich*) was established in the second half of the 19th century. Therefore, when Georg Wenker conducted his language survey shortly after the Empire was established, each of these independent territories had been in existence for several centuries.

Between 1879 and 1888, Wenker asked teachers and pupils in about 45,000 schools from all over the German Empire to translate 40 characteristic sentences into their local dialect. These sentences were especially designed so as to reveal specific dialect characteristics. At that time, these dialects were, for most people, their everyday language, whereas a standardized national language had not yet become prevalent. Wenker's surviving material contains millions of phonological and grammatical observations, which were integrated into a linguistic atlas of the German Empire (*Sprachatlas des Deutschen Reichs*).[5] To this day, there has not been an equally comprehensive inquiry into the dialect structure of a language.

[5] The *Sprachatlas* was developed between 1889 and 1923 using Wenker's original data. See Lameli (2008) for further details, and the website http://www.diwa.info for several maps.

Linguistic characterization of local dialects: To make these data usable, some simplification was necessary. First, we needed to reduce the number of linguistic attributes to be considered. The selection of variables was based on phenomena known to be especially important for the structuring of the German language according to the linguistics literature. The selected variables match, to a high degree, those characteristics that Ferdinand Wrede (Wrede *et al.* 1927–1956) used in his still relevant structuring map of German dialects. We selected a total of 61 linguistic attributes having to do with the pronunciation of consonants and vowels, as well as with grammar. For each school we use the teacher's phonetic protocols on the pupils' language characteristics (see Figure A1 in the Appendix for an example). We then represent the dialect of each school in the form of binary codes over the 61 chosen language characteristics. For example, the German word for *pound* is, depending on the dialect, pronounced as "pfund," "pund," or "fund." These variants are transferred into a binary coding of the type: "Pfund" = {1 0 0}; "Pund" = {0 1 0}; "Fund" = {0 0 1}. Altogether, the 61 language characteristics give us K=293 binary variables that represent a specific dialect.

Second, we needed to reduce the number of locations in order to merge the historic dialect data with current migration data. Specifically, our choice was predetermined by the current 439 German NUTS3 districts (*Landkreise*) for which data on cross-regional migration flows are available. Notice that this NUTS3 classification system of current German regions is basically inconsistent with the independent territories out of which the German Empire was built, and also with the Empire's regional classification system in use when Wenker collected the data. This poses the problem of matching the more than 45,000 observations for the phonetic protocols of the single schools with the *R*=439 current German districts.

We proceed as follows: for each district $r = 1, \ldots, R$ we select one representative school that is located within the current boundaries of that district. We then measure the dialect that was spoken in the area of that district at the end of the 19th century by the phonetic protocol of the respective school. For this approach to be consistent and reliable, we need clear criteria for selecting the schools and checking their representativeness.

First, we aimed at selecting rural schools for each district, as it is well known in linguistics that dialects are more pronounced in rural areas than in urban areas (Chambers & Trudgill 1998). Second, we found that some questionnaires had been answered by teachers, but that most questionnaires had been answered by pupils. To hold the social group uniform, we decided to consider only schools where the pupils answered the language survey. In the next step, we considered the 61 individual language attributes separately. Given our initial choice of 439 specific schools, we projected the geographical structure of each of the 61 variables separately using geographic information system (GIS) software. We then compared this map to the historic language maps of the *Sprachatlas*, which do not follow the NUTS3 classification system, referring to the same particular language characteristic. When we discovered inconsistencies between the recent and the historic maps for single areas, we took this as evidence that our initial choice of schools for these areas was not representative and so then used a different school, proceeding in this fashion until we achieved the maximum possible consistency between our projected maps for the single language characteristics and the maps provided in the *Sprachatlas*. To illustrate this procedure more formally consider one particular NUTS3 district r whose dialect is represented by the phonetic protocol of some school s_1 that is located in the area of r. Using the language characteristic $i=1,\ldots61$ (for example, "*pfund*" versus *"pund"* versus "*fund"*), we compare the realization in the chosen school s_1, denoted s_{1i},with what should be observed in the area now comprising region *r*, given the linguistic information provided in the *Sprachatlas*, which was also compiled from the Wenker data. That information from the *Sprachatlas* about characteristic *i* is denoted by a_i. When s_{1i} and a_i correspond, our selection of school s_1 appears to be representative for district r with respect to characteristic *i*. When s_{1i} and a_i do not correspond, we used another school, s_2, as the representative of region *r* and repeated the above procedure. Since the database contains some 45,000 schools, there is a sufficient number of observations available for each single district so that we can select schools that match the dialect structures a_i very closely with respect to virtually all language characteristics.

Dialect similarity matrix: We end up with a selection of *R* schools, one for each German district. The dialect of region *r* is then represented by a vector $\mathbf{r} = \{r_1, r_2, ..., r_K\}$ of length K=293, where each vector element is a binary variable [0,1]. Using these data, we then construct a dialect similarity matrix across the *R* regions. This is done by taking two German districts *i* and *j* whose dialects are represented by $\mathbf{i} = \{i_1, i_2, ..., i_K\}$ and $\mathbf{j} = \{j_1, j_2, ..., j_K\}$, respectively. We then use a simple count similarity measure to quantify the overlap of the two dialects, namely, $\ell_{ij} = \mathbf{i} \cdot \mathbf{j}$, where $\ell_{ii} = K$ and $0 \leq \ell_{ij} \leq K$ for $i \neq j$. The resulting similarity matrix has dimension $R \times R$ with generic element ℓ_{ij}.[6]

2.3. A first look at dialect similarities

What does dialect similarity ℓ_{ij} capture? As discussed in Section 2.1, language evolves very slowly. In this subsection we illustrate with some examples that the observed linguistic patterns across German regions are the result of past interactions and can often be traced to other dimensions of cultural identity as well as to unique historical events.

Religion: The Reformation of the 16th century resulted in distinct Protestant and Catholic areas in Germany.[7] Figure 1 shows the geography of religious orientation in the south-west of Germany (Baden-Württemberg) in 1547 and in 1820.

FIGURES 1 - 3 HERE

With the sole exception of today's borders with Switzerland and France, the observed spatial patterns are virtually identical. This is chiefly due to social practice. In earlier times it was uncommon, if not completely unheard of, to marry across religious borders; Protestants marry Protestants, Catholics marry Catholics. Other factors that might stabilize such patterns are geography (the Black

[6] We also calculated different similarity indices, such as those proposed in the linguistics literature by Jaccard (1901) and Tanimoto (1957). Our main results in Section 4 are not sensitive to the specific choice of the similarity index. We therefore focus on the simple count similarity measure in the following analysis.

[7] See also Becker and Woessmann (2009) on the transmission of values after the Reformation.

Forest), or national and administrative borders (in this case, the border of the archbishopric Freiburg). Considering that the picture shown in Figure 1 covers nearly 300 years, it becomes obvious the extent to which cultural (here, religious) patterns remain stable over time.

Now consider the linguistic pattern of the same area. Figure 2 shows the regional similarities to the dialect spoken in the *Waldshut* district. The reference point *Waldshut* is the colorless area, which is located in the extreme south-west of Germany. Warm colors in Figure 2 indicate a high, and cold colors a low, degree of dialect similarity. The specific shape of the dialect similarity pattern is strikingly similar to the religious geography shown in Figure 1. This similarity becomes even more obvious in Figure 3 where we superimpose the religious orientation map on the dialect similarity map: the two virtually coincide. This example clearly shows that dialect similarities can match religious borders, which fits nicely with the evolutionary perspective on culture and language as discussed above. Catholic localities are in closer contact with other Catholic localities, Protestants more in contact with Protestants. Hence, over time patterns of cultural and linguistic similarities can co-evolve.

Mass migrations: As language is the result of evolutionary processes, it necessarily also reflects the influences of migration waves. To illustrate this point, let us consider the example of the *Goslar* district. Linguists view the Harz region near Goslar as a *language enclave* in the sense that the dialect spoken in Goslar is not similar to the dialects spoken in neighboring districts but resembles a dialect spoken about 300 kilometers away in some Saxon districts in the *Erzgebirge* (see also Wiesinger 1983).

FIGURE 4 HERE

This initially seems very peculiar, but can be explained, as can so much, by history. It turns out that the revival of silver mining in the Goslar area between 1520 and 1620 motivated starving miners from Saxony to move to Goslar. What is so striking about this example is that evidence of this 16th-century relationship between the two regions is still visible in dialect data from the late 19th century.

Figure 4 maps the dialect similarity between Goslar (white) and all other districts. The map clearly reveals an accumulation of warm colors (indicating high similarity) in the Erzgebirge south of Saxony. An important aspect of pre-modern migration is that it was nearly always a social or *mass* phenomenon, and thus much different from current migration behavior, which is strongly based in *individual* economic motives. With very few exceptions, these mass migrations in Germany ended during the 18th century (Wiesinger 1983). Therefore, at the time Wenker conducted his language survey (1879–1888), roughly one and a half centuries had elapsed without a major perturbation of regional culture or dialect.[8] The local cultures and dialects had thus quite some time to develop and "harden." During the 20th century, the nature of migration changed. There were no more mass migrations motivated by hunger or economic devastation; migration became an individual phenomenon. This implies that the dialect patterns recorded by Wenker were not subject to major disturbance after he collected them, with the exception of World War II and German Reunification, which are discussed below. Regional dialects do, of course, constantly evolve and the increase of economic exchange and individual migration during the 20th century certainly played a role in this process. However, as argued in Section 2.1., evolution progresses slowly and local cultures and dialects are, even today, still far from perfectly assimilated. Our empirical results, reported in Section 4, support this view.

Geographic distance: Geographic distance also plays a role, and a more complicated one than is intuitive, in dialect similarity, as illustrated by the Waldshut example. The districts directly adjacent to Waldshut tended to have similar dialects, as can be seen in Figure 2, yet we also find districts relatively close to Waldshut that are less similar than districts that are farther away. This counterintuitive configuration suggests that our dialect data contain information that goes beyond

[8] The last incident known to us that can be classified, albeit rather broadly, as a *mass migration* occurred between 1749 and 1832. Initially, a rather small community of people from the Palatinate decided to immigrate to America, but ended up as settlers in a region near the city of Kleve. The reason for migrating was hunger caused by a poor harvest and is, thus, bound to geographic and climatic factors. Once settled in that area, other families from the Palatinate followed. However, this is the last migration of this kind, and because of the small size of the involved communities (only three very small villages), it does not have a huge influence on our data.

what can be explained by mere physical distance, a point made quite plain by the Goslar example (Figure 4), where there is virtually no relationship between geographical distance and dialect similarity. The linguistic measure thus seems to capture far more than just geographical distance. It could, however, still reflect the existence of old trading routes, which, by taking advantage of rivers, natural passages, and forts, historically led to more contact between regions. In other words, dialect similarity may be correlated with the effects of ancient transportation networks, although the Goslar example suggests that this is unlikely to capture the entire story.

Historic borders: At the time Wenker collected the data, the German Empire had just been created out of formerly independent kingdoms, principalities, and dukedoms. These independent territories were not new themselves, of course, having been in existence for at least one and a half centuries prior to their incorporation into the empire, and they were most certainly of profound influence on cultural identity and dialects. In fact, the dialectology of the 19th century was quite aware of the congruencies between the areal distribution of historic territories and language (see Haag 1898; Aubin *et al.* 1926; and, more recently, Barbour & Stevenson 1990). This suggests that the borders of the historic German territories, which do not correspond to the borders of the current regional classification system, were still influencing which dialects are spoken in the German NUTS3 districts. One possible reason for such a persistence of historic border effects may be that both medieval and early modern territories tended to encourage internal traffic, and discourage, or at least not improve the means for, travel external to their borders. Hence, communication and all that it leads to, such as trade, tended to be rather territory specific (Bach 1950:81) and exchange between territories somewhat hindered. From an evolutionary perspective, it is clear that such limitations led to a higher degree of dialect similarities among current NUTS3 regions that used to belong to the same historic territory.

The examples discussed in this subsection suggest that the geography of dialect similarity is far from random, but instead reflects region-pair-specific congruencies such as common religious and historical political borders, distance, and possibly the influences of ancient transportation networks, as well as the long-lasting impact of unique historical events and previous migration waves. All these influences have left long-lasting imprints on the linguistic structure of local dialects. These dialects can therefore be understood as a comprehensive measure for the *cultural identity* of a region that has been shaped during centuries of interaction. In the empirical analysis we investigate whether and, if so, to what extent these historical dialect similarities continue to affect bilateral economic exchange in the form of cross-regional migration. In doing so, we aim at identifying the magnitude of intangible cultural borders that may impede economic exchange at the regional level. Before turning to the empirical analysis, however, we first develop a simple theoretical model of cross-regional migration flows that will serve as the underlying framework for the analysis.

3. A Gravity Model of Regional Migration

In today's economy, migration is an individual economic decision: single workers (or families) choose a location to maximize utility. There is a great deal of literature analyzing the determinants of such individual migration decisions. For example, it is well known that people tend to move toward areas that offer good job prospects, high wages, low unemployment rates, etc. A salient feature of regional migration data, however, is the presence of two-way gross migration flows that are substantially larger than net flows (see Dahl 2002). That is, there is not only migration from economically poor to rich regions, but also the other way around.[9] This suggests that the location decisions of individuals are also guided by other than strictly economic variables, and that individuals are heterogeneous in their perceptions of different regions. Regional cultural differences are likely to play a major role in location decisions. A second important fact about regional migration flows is that they are costly and that the overall migration costs are distance-dependent. This fact is explicitly acknowledged in the gravity literature on migration, which has found much larger flows over short than over long distances.

[9] This fact is not easily reconciled with standard models of regional labor mobility (e.g., Krugman 1991) that predict only one-way migration flows.

In this section we develop a simple and highly stylized gravity model of gross migration flows. Individuals are heterogeneous and face distance-dependent mobility costs should they decide to move. We not only include standard mobility costs (for moving furniture, finding accommodation, etc.), which may be approximated by physical distances, we also incorporate, in the spirit of Sjastaad (1962), non-pecuniary (psychic) costs of migration at the region-pair level, which capture the costs of adaption to a new cultural environment. These costs are more substantial when source and destination areas exhibit huge cultural differences and we will measure these current cultural mobility costs by the historic dialect similarity index ℓ_{ij} (see below).

3.1. Basic setup

Consider a country that consists of *i=1,…,R* regions and a huge mass of individuals (indexed by *h*) with heterogeneous tastes for the different regions. For individual *h*, indirect utility in region *i* is given by

$$V_i^h = u_i + \varepsilon_i^h \tag{1}$$

The variable u_i stands for the "economic" level of well-being in region *i*. This includes the local wage level, unemployment rate, price level, etc. This economic level of well-being is the same for all individuals in a region, and may even include regional amenities to which all individuals assign the same value. For our purposes it suffices to think of u_i as being exogenously given. That is, we abstract from market interactions and assume for the sake of simplicity that the regional levels of economic well-being do not respond to the location decisions of the workers.[10] The term ε_i^h in Equation (1) is an idiosyncratic term for individual *h* and region *i* capturing his or her perception of the attributes and characteristics associated with that particular region.

The model specified in Equation (1) is a "random utility model," which makes use of discrete choice theory as pioneered by McFadden (1974). As shown in Anderson *et al.* (1992:ch. 3), individual taste

[10] This economic level of well-being could be endogenized. In models of the new economic geography, for example, economic well-being u_i is typically a function of the size of the population that resides in region i via the effects on wages, price levels, consumption variety, etc. See Murata (2003) and Tabuchi and Thisse (2002) for such models that allow for heterogeneous locational tastes and treat u_i as endogenous. Our aim in this paper is to estimate cultural costs of regional migration. For this purpose it is sufficient to focus on a simple location model without market interactions which yields a standard gravity equation for gross migration flows.

heterogeneity can be modeled such that the actual matching value between a worker and region is the realization of a random variable. We follow this modeling strategy and assume that ε_i^h in Equation (1) is a random variable that is distributed i.i.d. across individuals and regions. Furthermore, we adopt the standard parameterization of a double exponential distribution

$$F(x) = \Pr\left(\varepsilon_i^k \leq x\right) = \exp\left[-\exp\left(-\frac{x}{\beta} - \gamma\right)\right] \tag{2}$$

where γ (≈0.5572) is the Euler constant and β>0 is a parameter. This distribution has mean zero and variance $\left(\pi^2/6\right)\cdot\beta^2 \approx 1.6449\cdot\beta^2$. The term β, which is positively associated with the variance of the distribution, is referred to as the "degree of taste heterogeneity." It is a well-established result that under this parameterization, the choice probability of some individual *h* to live in region *i* can be calculated as follows (see Anderson *et al.* 1992):

$$\mathrm{P}_i = \Pr\left[V_i^h > \max_{j\neq i}\left\{V_j^h\right\}\right] = \frac{\exp\left[u_i/\beta\right]}{\sum_{j=1}^{R}\exp\left[u_j/\beta\right]} \tag{3}$$

The larger β, the more heterogeneous are the individual attachments to the regions. It can be shown that if β → 0, people will make location decisions based only on the economic levels of well-being u_i. We are then back to a model having homogeneous individuals. On the other hand, if β extends to infinity, people choose among the *R* regions with equal probability *(1/R)*. In this case, location tastes are extremely heterogeneous and the economic levels of well-being have no effect on location decisions.

3.2. Mobility costs and the gravity equation

Notice that this model has the realistic property of two-way gross migration flows. To see this, it is useful to embed the above model in a two-period framework. Individuals are distributed in some way across regions, and the random variables ε_i^h are drawn in the first period. Individuals then choose the location they most prefer during the second period. Depending on the realizations of the random variables, this may involve migration to an area with a lower level of economic well-being than in the current source region, as well as parallel gross flows of individuals from *i* to *j* and from *j* to *i.* It is straightforward to include mobility costs and to derive a gravity equation for gross migration flows from this simple two-period version of the random utility model. Specifically, an individual *h* will migrate from the initial location *i* to some other region *j* in the second period if the overall utility from living in *j*, net of the region-pair specific mobility costs c_{ij}, exceeds the (net of mobility costs) utility level of all other locations *s* including the current location *i*. Formally, a move from *i* to *j* takes place if

$$V_j^h - c_{ij} > \max_{s \neq j} \left\{ V_s^h - c_{is} \right\} \text{ with } c_{ii} = 0 \text{ and } c_{is} \geq 0 \text{ if } s \neq i \qquad (4)$$

Making use of the previously mentioned parameterization and the results from discrete choice theory, we can calculate the following probability for an individual to migrate from *i* to *j*:

$$P_{ij} = \frac{\exp\left[\left(u_j - c_{ij}\right)/\beta\right]}{\sum_{s=1}^{R} \exp\left[\left(u_s - c_{is}\right)/\beta\right]} \qquad (5)$$

Aggregating across individuals it is easy to see that the total gross migration flow from region *i* to *j* is given by $M_{ij} = P_{ij} \cdot L_i$, where L_i is the current population size in the source region *i*. Using Equation (5) and taking logs we obtain

$$\log\left(\frac{M_{ij}}{L_i}\right) = \frac{u_j - c_{ij}}{\beta} - \log\left[\sum_{s=1}^{R} \exp\left[\left(u_s - c_{is}\right)/\beta\right]\right] \qquad (6)$$

Notice that the second term on the right-hand side of Equation (6) varies only at the level of the source region, while the term $\left(u_j/\beta\right)$ varies only at the level of the destination region. The mobility costs c_{ij} are region-pair specific. We assume the following specification:

$$c_{ij} = a_1 \cdot \log\left[d_{ij}\right] + a_2 \cdot \log\left[\ell_{ij}\right] \qquad (7)$$

where d_{ij} is physical distance and ℓ_{ij} is cultural distance between regions *i* and *j*. Taking Equation (7) into account, we can rewrite the gravity Equation (6) in stochastic form to arrive at our final estimation equation:

$$\log\left(\frac{M_{ij}}{L_i}\right) = D_i + D_j + \alpha_1 \cdot \log\left(d_{ij}\right) + \alpha_2 \cdot \log\left(\ell_{ij}\right) + e_{ij} \tag{8}$$

where $D_j = \left(u_j / \beta\right)$ and $D_i = -\log\left[\sum_{s=1}^{N} \exp\left[\left(u_s - c_{is}\right)/\beta\right]\right]$ are source and destination area fixed effects, and e_{ij} is a standard error term. Note that we have $\alpha_m = a_m / \beta$ for *m=1,2*. That is, we can identify the distance elasticity, referring to physical and cultural distance, up to the unobservable positive constant $(1/\beta)$, thus capturing taste heterogeneity. The main coefficients of interest are the physical distance parameter α_1 and, in particular, the parameter α_2, which measures the effects of cultural similarity on gross regional migration flows.

3.3. Discussion of identification and estimation issues

Migration versus trade flows: Let us briefly put this gravity model into perspective. Gravity equations are a standard tool for analyzing trade flows across countries or regions, but the conceptual idea of gravity was applied to migration flows even earlier.[11] There are two main reasons why we focus on migration rather than trade flows (or other cross-regional flows). The first issue is data availability. While there are accurate and highly disaggregated current regional migration data for Germany, there is no information at the regional level about commodity flows, goods or service trade, or financial flows (not even at the NUTS1 level). Second, while trade or financial flows would certainly be an interesting region-pair-specific outcome variable for studying the effects of intangible cultural borders, we believe that migration flows are at least equally well suited for this purpose. Individuals do not migrate very often during a lifetime, even at the regional level.[12] Hence, moving from one

[11] The earliest reference is Ravenstein (1885). Other important contributions include Schwarz (1973) and Greenwood (1975).

[12] Using Japanese data at the prefecture level from 1954–2005, Nakayima and Tabuchi (2008) report that individuals in Japan move on average only 2.3 times during their lifetime.

region to another is a substantial act, and cultural biases may influence such a decision even more strongly than, say, they would the decision to engage in goods trade with someone from a different region.

Current versus historic cultural differences: Migration decisions today are influenced by the perception of *current* cultural differences between regions. In the empirical analysis we measure these current cultural mobility costs by the historic dialect similarity index ℓ_{ij}. This approach makes two implicit assumptions: (1) that dialect differences are a good comprehensive measure for cultural differences and (2) that cultural differences across regions are highly persistent over time so that today's cultural differences are still captured by the historic differences across German regions in the 19th century.

The discussion in Section 2 suggests that both assumptions are reasonable. Regarding the first, we argued above that dialects comprehensively measure cultural identity at the local level (see Section 2.1). Regarding the second assumption, the linguistic diffusion that occurred between the late 19th century and today is unlikely to have nullified the local dialect differences as recorded in the Wenker survey roughly 120 years ago. Even though linguistic evolution progresses faster than genetic evolution, such a time period is still much too short to erase all regional differences given the enormous degree of inertia inherent in evolutionary processes. This is especially true because there have been no further mass migrations or other major perturbations of local cultures and dialects in Germany, as argued above, migration instead becoming more and more of an individual phenomenon during the 20th century. It is therefore not surprising that linguists have noted a close correspondence of current dialect characteristics across today's German regions with historic patterns that were recorded in the respective areas (see Bellmann 1985:213). This supports the view that local language patterns are very persistent over time.

There may be two exceptions to this second assumption, namely, the major perturbations that occurred in the aftermath of World War II and during German Reunification. After World War II, there was a huge inflow of people into Germany from East Prussia, Pomerania, Silesia, East

Brandenburg, and the Sudetenland. These people were not entirely free to choose their destinations, but were instead generally allocated across Germany by the administrations of the zones of occupation based on available housing and food supply (see Falck *et al.* 2009). There was also large-scale migration from Eastern to Western Germany after German Reunification in 1990 (see Redding & Sturm 2008). Both events may have caused cultural and linguistic perturbations that would not be captured by our historic dialect data. In the empirical analysis below we find, however, that controlling for these two extreme events hardly changes the impact of historic dialect similarity on actual migration flows. That is, even though both events might have influenced the cultural identity of German regions and migration behavior, there also continue to be independent effects from the long-term cultural borders across German regions that are measured by the historic dialect data.

In short, the historic dialect differences as measured in the 19th century seem to be a sensible proxy for current cultural differences across regions. Given this measure for cultural mobility costs ℓ_{ij}, the estimation Equation (8) is not plagued by the problem of reverse causality because of the time lag of about 120 years between the dialect survey and the current migration data.[13]

<u>Fixed effects and omitted regional variables</u>: Our gravity equation for migration flows includes fixed effects for both the source and the destination area (see Equation (8)). Such a specification is standard practice in the international trade literature (see Anderson & van Wincoop 2003; Feenstra 2004). The fixed effects capture all impact variables that vary only at the *regional* level in our cross-sectional analysis. These include contemporaneous influences on migration flows such as wages, unemployment rates, housing prices, etc., as well as unobservable regional features such as amenities or other “soft” location factors.[14] This fixed effects specification should also take into

[13] This problem is discussed, for example, in Guiso *et al.* (2009), who measure trust between countries by using recent Eurobarometer survey data. To address the problem that survey responses may be endogenous to the level of trade between countries, Guiso *et al.* (2009) instrument the level of trust by deeply lagged linguistic and somatic differences across national populations.

[14] This fixed effects specification also takes into account the problem of interdependent flows in a multi-region economy (Anderson & van Wincoop 2003). As shown in Feenstra (2004) in the context of trade flow analysis, the fixed effects approach allows for a consistent estimation of region-pair-specific impacts, which is the main aim of the empirical analysis in this paper.

account historic omitted variables at the regional level. To illustrate this latter point, think of some omitted factor that has led to persistent economic prosperity in some region (at least relative to other locations), both historically and today. The resulting persistent pull effects on migration into that region are, however, captured by the origin and destination area fixed effects in the estimation because the two fixed effects should level all actual differences in economic prosperity between the region of origin and the region of destination, regardless of whether these differences have their origin in history or are the result of current developments.

To support our assumption of "persistent fixed effects," we show that controlling for variables that reflect historic differences in economic prosperity across regions hardly change our results for the impact of historic dialect similarity on actual migration. In our empirical specification, we control for differences in historic economic prosperity, geological features, religious denomination, and for historic inner-German borders. These variables might have had an impact on the formation of a cultural identity and might still today shape regional prosperity. However, the effect of dialect similarity as a measure for regional cultural identity always maintains a robust positive impact on current cross-regional migration flows.

4. The Effect of Dialect Similarity on Regional Migration

4.1. Current migration data

To estimate Equation (8), we use data from the German Federal Statistical Office (Statistisches Bundesamt) on pair-wise gross migration flows for the R=439 districts (*Landkreise)* in Germany averaged over the period 2000–2006. In Germany, every person who changes his or her place of residence is legally required to register at the new residence within two weeks, and thus the migration flow data are of high accuracy.[15] Table 1 provides an overview of German regional migration data.

<< Table 1a - c about here >>

[15] As of 2007, variation in federal state registration law means that in Rhineland-Palatine, new residents must register immediately; in Berlin, Brandenburg, Bremen, Schleswig Holstein, and Saxony, newcomers are given two weeks to register; all other states require registration within one week.

In the aggregate, across all regional pairs, there has been some gross migration for more than 96% of all pairs of regions. Zero (gross migration) flows are found for only 4% of all pairs of districts and are therefore a relatively minor issue. Nevertheless, we deal with the zero flow problem below (see Section 4.5), as previous work in the gravity literature suggests that zero flow can pose a potentially severe estimation problem. Table 1 also indicates that migration flows are still relatively small in Germany. On average, there are only seven (nine) migrants per 100,000 German inhabitants (German working-age inhabitants) in the district of origin. Table 1b gives additional summary statistics for the pair-wise migration flows, as well as for pair-wise geographic distances and dialect similarities. Table 1c provides descriptive statistics for the other control variables we apply in the following specifications.

4.2. Baseline results

In our baseline approach, we estimate Equation (8) by simple ordinary least squares with origin and destination fixed effects. Table 2 presents the results. The left panel refers to gross migration flows of the entire regional populations, while the right panel presents the results when only considering the working-age population, that is, those aged 18–65.

<< Table 2 about here >>

We start by estimating the effects of physical and cultural distance on migration flows separately. As can be seen in specifications 1 and 5, the effect of geographic distance on gross migration flows is negative and highly statistically significant. Doubling the geographic distance between two regions, all else equal, drives down gross migration flows by roughly 140–150%. This result is similar for both the entire population and the working-age population. These numbers are somewhat lower than previous estimates for the distance elasticity of migration (see, e.g., Greenwood 1975), where researchers have sometimes found values larger than 2 in absolute terms. It should be kept in mind, however, that distance elasticity is deflated in our model by the unobservable degree of locational taste heterogeneity β, so that the overall magnitude that we identify with our approach appears reasonable.

The main result of Table 2, however, is the finding of a positive and highly statistically significant effect of dialect similarity on gross regional migration flows. When including only dialect similarity, as in specifications 2 and 4, we find a huge positive elasticity that even exceeds the physical distance elasticity in absolute terms. However, the examples discussed in Section 2.3 show that dialect similarity is correlated with physical distance. In fact, when controlling for both physical and cultural distance, as in specifications 3 and 7, we find that the coefficient α_2 drops substantially. However, we still find a positive and highly significant effect of dialect similarity on gross regional migration flows, somewhere in the range of 15–16%. In other words, even conditional on physical distance, there are intangible borders between German regions. Recalling that the coefficient α_2 is also scaled by the heterogeneity parameter β, the true elasticity of cultural similarity (a_2) is thus even higher than the coefficient α_2 reported in Table 2.

Specifications 1–3 and 5–7 follow the theoretical gravity equation derived from the simple model in Section 3. In that model, the number of migrants from region *i* to *j*, M_{ij}, is deflated only by the population in the district of origin, L_i (see the left-hand side of Equation (8)). The typical specification in the gravity literature, however, would be to deflate M_{ij} with the product of the populations in the district of origin, $L_{i,}$ and the district of destination, L_j (see, e.g., Greenwood 1975). When using this specification, as in columns 4 and 8 of Table 2, we obtain virtually identical results.

4.3. Alternative distance measures and borders in space

Travel time as an alternative distance measure: In the remainder of this paper we address the robustness of these main results. Our first check deals with the possibility that linear physical distance may be a poor proxy for the true pecuniary mobility costs. Recent research by Giuliano *et al.* (2006) even suggests that, in the context of the above-mentioned literature on how genetic similarities affect international trade flows, there may actually be no effects of genetic similarity once transportation costs across countries are properly controlled for.

To discover whether comparable issues play a role in our analysis of cross-regional migration flows, we need to more realistically control for pecuniary mobility costs (the analogue of transportation costs for goods). We thus consider travel time by car between any pair of regions (in minutes) as this measure may capture remoteness and accessibility better than linear physical distances. The results reported in Table 3 (column 1) show that the elasticity with respect to travel distance is in fact a bit larger than for geographic distance (about 170%).[16] This is intuitive as the linear geographic distance might not always match the shortest travel distance, e.g., because of natural barriers like rivers or mountains. However, when plugging in both geographic distance and travel distance (as in column 3), it turns out that geographic distance clearly dominates More important yet, even after controlling for geographical and travel distance, there is still a positive and significant effect of dialect similarity on cross-regional migration flows. The alternative control for pecuniary mobility costs, therefore, does not contradict our main conclusions about the prevalence of intangible borders to regional migration.

<< Table 3 about here >>

East and West Germany and the Federal States: In a next step, we consider a dummy that equals unity if either the district of origin or the district of destination is located in former West Germany while the other is located in former East Germany. As is well known, East Germany, which is still in transition after Reunification, suffers from labor market conditions that have resulted in the migration of a great many East Germans to the West. Furthermore, the very fact that the country was divided for so long may have resulted in persistent cultural differences between individuals born and raised in the different parts. These possible cultural differences may have an independent impact

[16] Table 3 concentrates on the total populations. The corresponding results for the working-age population are reported in Table A4 in the Appendix.

on non-pecuniary mobility costs, so that it is worth analyzing how the effects of dialect similarity are affected once we control for systematic differences between East and West Germany.

Following a similar rationale, we also consider a dummy that equals unity if the district of origin and destination are not located in the same Federal State (*Bundesland*). First, crossing state borders may increase pecuniary mobility costs. For example, German Federal States have different regulations and laws applicable to various occupational groups. It is, for instance, much more difficult for teachers or lawyers to change jobs across states than within a state. As for the non-pecuniary mobility costs, there may also be an independent effect of leaving one Federal State for another, and there may even be a correlation with recorded dialect similarity. As argued above, dialect similarity sometimes can reflect historic administrative borders, which, at least in some cases, may also be captured by the current borders of the Federal States.

As can be seen in columns 4–6 of Table 3, we do in fact find that there is systematically more migration between Eastern and Western Germany (which is driven by the still huge emigration from Eastern Germany), while Federal State borders significantly reduce gross migration flows. The effect of dialect similarity decreases somewhat when considering these additional controls, but there is no qualitative change in our conclusion. We still find significantly more gross migration between regions with more similar historic dialect characteristics.

Infrastructure and accessibility: Finally, we control for infrastructure indicators as published by the Federal Office for Building and Regional Planning. These were collected in 2004 and have to do with the availability of modern transportation systems in a region (cf. Maretzke 2005). The first indicator reports the accessibility of the nearest three national or international agglomerations in combined road and rail traffic in minutes; the second indicator measures the accessibility of European metropolis in combined road and air traffic in minutes. Based on these indicators, we calculate the absolute difference for all pairs of regions, which expresses differences in the distance-related migration costs between regions that are presumably not considered in Euclidian measures of geographic distance. It turns out that controlling for these additional factors again leads to a reduction of the effect of dialect similarity, but the effect is still significant (see column 7).[17]

4.4. Historic regional differences

As argued above, migration may be driven by persistent regional differences in economic prosperity and wealth. As long as these omitted factors are purely region-specific, they should be captured by the origin and destination fixed effects in the gravity equation so that our estimate for the dialect similarity elasticity is consistent. However, to add support to this approach, we also consider various types of historic region-pair-specific differences that have as yet been left out of the regression.

Geological regional differences*:* Salient candidates for controlling for historic differences in regional economic prosperity are indicators describing a region's suitability for agriculture, forestry, and mining, all of which were major sources of regional wealth before and during the Industrial Revolution. Along this line, Combes *et al.* (in press) argue that soil characteristics can be regarded as a major determinant of local labor demand in an agrarian society. Accordingly, differences in geologic indicators for the suitability of the soil for agriculture and forestry should provide a meaningful insight into the distribution of regional wealth before the heyday of industrialization.

To use current indicators of soil quality to determine a region's past agricultural productivity, we need to assume that soil characteristics have not changed during the past centuries, i.e., they are persistent over time. Following Combes *et al.* (in press), indicators such as soil mineralogy and the soil's dominant parent material were determined millions of years ago and are rather persistent, whereas other soil characteristics, such as its depth to rock or its carbon content, could be an outcome of human activity. Therefore, we use the presence of *minerals in the subsoil*, i.e., the intermediate layer between the topsoil and the bedrock, and the dominant parent material describing the underlying bedrock as indicators of regional agricultural conditions (cf. column 1 of

[17] We also tried a modified specification of Equations (7) and (8), where we included geographic distance and dialect similarity in levels instead of logs. Table A3 in the Appendix summarizes the results. Note, however, that the results can no longer be interpreted as a scaled elasticity. However, the results are qualitatively very similar when taking levels (and squared levels) instead of logs.

Table 4).[18] Both variables are scales of eight characteristics.[19] Additionally, we consider differences between regions' *slope* because this characteristic might well influence a region's agricultural productivity as it has an influence on the efficiency of agricultural production (cf. column 2 of Table 4). Slope is measured as the difference between the median maximum and minimum elevations in meters and thus has a natural interpretation.[20]

We expect these soil characteristics to capture in large part the most fundamental determinants of historic economic prosperity. Moreover, these variables also allow for some inferences as to a region's mineral wealth. However, as the simple existence of minerals does not necessarily imply their exploitation, we further consider the location of historic mining academies, believing this to be a good indicator for regional exploitation of mineral resources contributing to historic regional prosperity (cf. column 3 of Table 4).[21] Doing so leaves us with some measure of the relative importance of mining across regions. The difference between two regions' minimum distances to a mining academy then enters our estimations as a control for differences in the historic exploitation of mineral resources.

Historic borders: The discussion in Section 2.3 shows that dialect similarity often reveals a spatial pattern similar to that of historic religious borders and may also reflect historic political borders. We therefore control for these historical region-pair-specific differences in order to evaluate whether our previous effect for dialect similarity may reflect a spurious correlation with these historic borders.

In the first step, we control for differences in religious denominations in 1890, roughly the time at which our linguistic data were collected (cf. column 4 of Table 4). The measure is available in eight bins that divide each region's population share of Catholics into eight categories, where the share of

[18] We nevertheless tried a variety of other indicators related to climate and soil but as they did not affect our coefficients of interest, i.e., geographic distance and dialect similarity, we chose to concentrate on these two persistent indicators.

[19] Note that only five characteristics apply to Germany in the case of subsoil mineralogy.

[20] We are deeply indebted to Gilles Duranton for providing the data for these three indicators. For a more detailed description of the variables and their generation process, see Combes *et al.* (in press).

[21] More precisely, we consider 11 locations that are or were home to a mining academy within today's German borders and, additionally, one location in Bohemia (today the Czech Republic) that was part of the former German Empire (cf. Table A1). We calculate each district's minimum distance to the next mining academy, referring to the regions' centroids. We also consider two more mining academies in Silesia, which, however, did not appear to be in minimum distance to any district in today's Germany.

Catholics constantly increases over the categories. Based on these shares, we calculate the absolute difference for all pairs of regions *i* and *j*. The absolute value expresses that differences in religious denomination affect regional exchange in both directions equally and are thus region pair specific. Second, to control for historic political borders, we include a dummy that equals unity if a historic border was crossed (cf. column 5 of Table 4). Historic boarders surround 38 member states and 4 independent cities that were part of the German Confederation at its foundation in 1815. They reflect the environment of political fragmentation that prevailed until the German Empire was established in the second half of the 19th century.

Table 4 shows the estimation results when successively including the historic control variables and, additionally, the state and east/west dummies from the previous subsection.[22] The main result is that the magnitude of the scaled dialect similarity elasticity remains robust and is quite similar in magnitude over all estimations. The robustness of this central coefficient of interest confirms our assumption of "persistent fixed effects," i.e., region of origin and region of destination fixed effects are suitable for capturing actual differences in economic prosperity between the region of origin and the region of destination regardless of whether these differences have their origin in history or are the result of current developments.

Migration flows after WWII: In the last specification, column 6 of Table 4, we consider the immense migration flows of expellees after WWII (see also Section 3.3). These expellees were either German citizens or ethnic Germans who, before and/or during World War II, lived within the Eastern German borders as existed between 1917–1937 or in Austria-Hungary (§1, Federal Expellee Law, May 19, 1953). Late in World War II, these individuals were forced by the Red Army and, after World War II, by the *Potsdam Treaty*, to leave their homeland and settle within the new borders of Germany or Austria. Almost 12 million ethnic Germans fled or were expelled from their homes in East Prussia, Pomerania, Silesia, East Brandenburg, and the Sudetenland to find refuge in other German states.

[22] Table 4 refers to the entire population. The results for the working-age population are quite similar and are reported in Table A4 in the Appendix.

The distribution of *Heimatvertriebene* across the settlement states was to a considerable extent based on a central allocation formula (Edding, 1952; Grosser, 2001; 2006; Hoffmann, 2000) that was based on the availability of food and housing.[23]

The fact that families were sometimes separated in the allocation process shows very plainly that the expellees had no choice in where to settle, and this was thus not a "natural" migration flow that might have let to changes in dialects and thus influence our results (Bellmann & Göschel 1970:12 f.). We nevertheless control for differences in the regional population share of expellees. The data on expellees are taken from the West German population census conducted in 1950.[24] Therefore, we have no observations for Eastern Germany and the Saarland, which was French until the mid 1950s.[25]

<< Table 4 about here >>

Because of this reduced sample size, the specification in column 6 is not comparable to the other columns. Therefore, we run only a basic specification with geographic distance and dialect similarity and add the regional differences in the share of expellees as a control. Doing so does not affect our coefficients' robustness and they remain significant.

We conclude that current migration flows are robustly positively affected by similarities in regional cultural identity, as measured by dialect similarity. This is true even after controlling for other region-pair-specific variation in religious, political, and geological conditions.

4.5. Zero flows and heteroscedasticity

By taking logs of the left-hand side variable in our baseline specification, pairs of districts with zero migration flows are dealt with as missing values. In Table 5, we propose several ways of coping with zero flow pairs of districts. In columns 1 and 4, we reestimate our baseline specification; however, we add one migrant to all zero migration flows so that we do not use zero flows due to taking logs. We

[23] For details, see Falck *et al.* (2009).

[24] These data were published by the Minster for Expellees (*Bundesminister für Vertriebene*) in 1952.

[25] More precisely, the Saarland was annexed by France in 1947 but in a national referendum in 1955, the inhabitants opted to join the Federal Republic of Germany. In 1957, the Saarland was politically integrated and, finally, in 1959, it became economically integrated.

additionally control for this manipulation by means of a zero flow dummy that equals unity for all initial zero migration flows.

In columns 2 and 5, we employ a two-stage Heckman estimation procedure that uses a non-linear probit equation for selection into migration in the first stage, and then estimate Equation (8) in the second stage. We thus rely on the normality assumption for identification of our second-stage estimates. For international trade, Helpman *et al.* (2008) argue that the first stage reflects the entry decision, i.e., whether to export in a certain country, while the second stage reflects the marginal decision, i.e., how much to export to this country. Similar considerations may apply for the case of regional migration, as there may be both fixed and variable costs of moving across regions.

Second, the interpretation of the parameters of log-linearized models estimated by linear least squares methods can be misleading in the presence of heteroscedasticity. To overcome this problem, we estimate Equation (8) by means of a Poisson pseudo-maximum-likelihood estimator (PPML) with Eicker-White robust standard errors, as proposed by Santos-Silva and Tenreyro (2006). This estimator can be used even though the dependent variable—the level of the migration share instead of the log—is not an integer. Columns 3 and 6 present the results of the PPML estimator.

<< Table 5 about here >>

When adding one migrant to all zero migration flows and thereby treating them as positive migration flows, the scaled distance elasticity, as well as the scaled dialect elasticity, drop to about -1.13 and 0.07. By contrast, when applying the two-step Heckman selection model, the estimates are similar to the basic specification. Obviously, these results suggest that there is some additional information in the zero flows. In the PPML estimations, the parameters of interest again can be interpreted as an elasticity value. In this specification, the scaled geographic distance elasticity is somewhat larger than in the baseline specification and reaches a value of about –1.7. Also, the scaled dialect elasticity is somewhat larger and reaches a value of about 0.34. These results suggest that heteroscedasticity is an issue and that we have underestimated the effect of geographic distance and dialect similarity on migration in our baseline specification. This prediction is in line with our theoretical model, which explicitly considers preference heterogeneity of migrants. Again, there are no important differences between the results based on the entire population and the results based on the working-age population. All in all, this robustness check also shows that there are positive and significant effects of dialect similarity across German regions on current bilateral gross migration flows.

5. Conclusion

In this paper, we analyze the impact of cultural identity on regional economic exchange. We argue that regions develop a common cultural identity from past interactions, including those occurring via mass migrations and ancient travel routes and due to religious and political divisions, and that the resulting cultural similarities between two regions do not disappear quickly, if at all. To proxy cultural similarities, we utilize detailed linguistic micro-data on the intra-national variation of phonological and grammatical attributes within the same language, German.

We study the effects of historic dialect similarities on actual bilateral economic exchange across regions in a gravity analysis. Within the framework of a region of origin and region of destination fixed effects model, we find that cultural identity has a strongly significant and positive impact on regional migration beyond what geographic distance would suggest. This result is robust to the choice of distance measures, i.e., Eucledian distance or travel time; the inclusion of control variables that represent historic regional differences in prosperity, religion, and institutions; the control for major perturbations in the aftermath of World War II, i.e., mass migration of German expellees; and the division of Germany into the Federal Republic of Germany and the German Democratic Republic.

The inclusion of origin and destination fixed effects, combined with these robustness tests, makes us confident that the effect of historic dialect similarity on regional migration can plausibly be interpreted as a causal effect of *cultural identity* on migration. We certainly cannot capture a causal effect of language, in the sense of posing a question such as: “What are the effects of linguistic similarity across regions on current migration flows that do not reflect other cultural influences?” Such an endeavor would be doomed to failure, since language can never be detached from various other cultural influences. However, since we interpret dialects as a comprehensive measure for regional cultural identity, our empirical results may answer the question “What are the effects of cultural similarity across regions on current migration flows that do not reflect other obvious influences, such as religious, political, or geological-economic similarities across regions?” In this respect, we find a robust positive effect of cultural similarity that seems to be highly persistent over time. In other words, there are cultural borders across regions that impede economic exchange.

We close by discussing two directions for further research on the role of cultural identity in empirical economics. First, further research on cultural identity could contribute to the discussion on interregional knowledge flows. Job-hopping by highly qualified employees across regions, patent citations across regions, or interregional phone calls, for example, are only three of the many ways of analyzing interregional knowledge flows that might be affected by cultural identity. Second, cultural identity might be a less technically driven way of thinking about spatial dependence in econometrics. Against this background, our dialect similarity matrix could serve as a spatial weighting matrix in econometric analyses at the regional level.

References

Alesina, A., and E. La Ferrara (2005). Ethnic Diversity and Economic Performance. *Journal of Economic Literature*, 43(3), 762–800.

Anderson, S.P., A. de Palma, J.-F. Thisse (1992). Discrete Choice Theory of Product Differentiation. Cambridge, MA: MIT Press.

Anderson, J.E., and E. van Wincoop (2003). Gravity with Gravitas: A Solution to the Border Puzzle. *American Economic Review* 93(1), 170-192.

Aubin, H., T. Frings, and J. Müller (1926). *Kulturströmungen und Kulturprovinzen in den Rheinlanden. Geschichte, Sprache, Volkskunde*. Bonn: Röhrscheid.

Bach, A. (1950). *Deutsche Mundartforschung. Ihre Wege, Ergebnisse und Aufgaben.* 2nd edition. Heidelberg: Winter.

Barbujani, G., M. Stenico, L. Excoffier, and L. Nigro (1996). Mitochondrial DNA sequence variation across linguistic and geographic boundaries in Italy. *Human Biology* 68(2). 201–215.

Barbour, S., and P. Stevenson (1990). *Variation in German. A Critical Approach to German Sociolinguistics*. Cambridge: Cambridge University Press.

Barro, R. J., and R. M. McCleary (2003). Religion and Economic Growth. *American Sociological Review* 68(5), 760–781.

Becker, S. O., and L. Woessmann (2009). Was Weber Wrong? A Human Capital Theory of Protestant Economic History. *Quarterly Journal of Economics* 124(2), 531-596.

Bellmann, G. (1985). Substandard als Regionalsprache. In G.Stötzel, G. (ed.) *Germanistik – Forschungsstand und Perspektiven*. Part 1: Germanistische Sprachwissenschaft. Didaktik der Deutschen Sprache und Literatur. Berlin / New York: de Gruyter, 211–218.

Bellmann, G., and J. Göschel (1970). Tonbandaufnahme ostdeutscher Mundarten 1962–1965. Gesamtkatalog. Marburg: Elwert.

Bernhard, H., E. Fehr, and U. Fischbacher (2006). Third-Party Punishment Within and Across Groups: An Experimental Study in Papua New Guinea. *American Economic Review*, *Papers and Proceedings* 92(2), 217–221.

Brewer, M. B. (1991). The Social Self: On Being the Same and Different at the Same Time, *Personal and Social Psychology Bulletin* 17(5), 475–482.

Cavalli-Sforza, L.L. (2000): *Genes, peoples, and languages*. London: Penguin.

Chambers, J.K., and P.Trudgill (1998). *Dialectology.* 2nd edition. Cambridge: Cambridge University Press.

Chwe, M. (1999). Structure and Strategy in Collective Action. *American Journal of Sociology* 105(11), 128–156.

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology* 94(S1), S95–S120.

Combes, P.-P., G. Duranton, L. Gobillon, and S. Roux (in press). Estimating Agglomeration Economies with History, Geology, and Worker Effects. In E. Glaeser (ed.) *The Economics of Agglomeration*, Chicago, IL: University of Chicago Press.

Dahl, G. (2002). Mobility and the Return to Education: Testing a Roy Model with Multiple Markets. *Econometrica* 70(6), 2367-2420.

Desmet, K., M. Le Breton, I. Ortuno-Ortin, and S. Weber (2009). The Stability and Breakup of Nations: A Quantitative Analysis. Unpublished Manuscript, Universidad Carlos III Madrid.

Dupanloup de Ceuninck, I., S. Schneider, A. Langaney, and L. Excoffier (2000). Inferring the Impact of Linguistic Boundaries on Population Differentiation: Application to the Afro-Asiatic-Indo-European case. *European Journal of Human Genetics* 8(10). 750–756.

Edding, F. (1952). *Die Flüchtlinge als Belastung und Antrieb der Westdeutschen Wirtschaft*. Kieler Studien, 12, Kiel.

Falck, O., S. Heblich, and H. Patzelt (2009). Entrepreneurship Policy and Regional Entrepreneurship: German Expellees as a Natural Experiment. Unpublished Manuscript, CESifo Munich.

Feenstra, R. (2004). *Advanced International Trade*. Princeton: Princeton Univ. Press.

Giuliano, P. (2007). On the Determinants of Living Arrangements in Western Europe: Does Cultural Origin Matter. *Journal of the European Economic Association* 5(5), 927–952.

Giuliano, R., A. Spilimbergo, and G. Tonon (2006). Genetic, Cultural and Geographical Distances. IZA Working Paper 2229.

Glaeser, E. L., D. Laibson, and B. Sacerdote (2002). An Economic Approach to Social Capital. *Economic Journal* 112(483), 437–458.

Greenwood, M. J. (1975). Research on Internal Migration in the United States: A Survey. *Journal of Economic Literature* 13(2), 397-433.

Grosser, T. (2001). Die Aufnahme der Heimatvertriebenen in Württemberg-Baden und die regionalen Rahmenbedingungen ihrer Integration 1946-1956. In P. Heumos (ed.) *Heimat und Exil: Emigration und Rückwanderung, Vertreibung und Integration in der Geschichte der Tschechoslowakei*. München: Oldenbourg, 223-261.

Grosser, T. (2006). *Die Integration der Heimatvertriebenen in Württemberg-Baden (1945-1961)*. Stuttgart: Kohlhammer.

Guiso, L., P. Sapienza, and L. Zingales (2006). Does Culture Affect Economic Outcomes? *Journal of Economic Perspectives* 20(2), 23–48.

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

Haag, K. (1898). *Die Mundarten des oberen Neckar- und Donaulandes*. Reutlingen: Hutzler.

Helpman, E., M. Melitz, and Y. Rubinstein (2008). Estimating Trade Flows: Trading Partners and Trading Volumes. *Quarterly Journal of Economics* 123(2), 441-487.

Hoffmann, D. (2000). Binnenwanderung und Arbeitsmarkt: Beschäftigungspolitik unter dem Eindruck der Bevölkerungsverschiebung in Deutschland nach 1945. In D. Hoffmann, M. Krauss, and M. Schwartz (eds.) *Vertriebene in Deutschland: Interdisziplinäre Ergebnisse und Forschungsperspektiven*. München: Oldenbourg, 219-235.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* 37, 547–579.

Knack, S., and P. Keefer (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *Quarterly Journal of Economics* 112(4), 1251–1288.

Krugman, P. (1991). Increasing returns and economic geography. Journal of Political Economy 99(3), 483-499.

Lameli, A. (2008). Was Wenker noch zu sagen hatte...Die unbekannten Teile des 'Sprachatlas des deutschen Reichs'. *Zeitschrift für Dialektologie und Linguistik* 75(3), 255–281.

Lazear, E. P. (1999). Culture and Language. *Journal of Political Economy* 107(6), S95–S126.

Manni, F. (in press). Sprachraum and genetics. In: A. Lameli, R. Kehrein, and S. Rabanus (eds.) *Language and* Space. Vol 2: Language mapping. Berlin, New York: de Gruyter.

Manni, F., W.J. Heeringa, and J. Nerbonne (2006). To what Extent are Surnames Words? Comparing the Geographic Patterns of Surname and Dialect Variation in the Netherlands. *LLC Literary and Linguistic Computing* 21*(*Special issue: "Progress in Dialectometry: Toward Explanation"), 507–527.

Maretzke, S. (2005). *Aktualisierung des Infrastrukturindikators für die Neuabgrenzung der Fördergebiete der Gemeinschaftsaufgabe "Verbesserung der regionalen Wirtschaftsstruktur"*. Bonn: Bundesamtes für Bauwesen und Raumordnung.

McFadden, D. (1974). Conditional logit Analysis of Qualitative Choice Behavior. P.Zarembka (ed.) *Frontiers in Econometrics*. New York: Academic Press, 105– 142.

Melitz, J. (2008). Language and foreign trade. *European Economic Review* 52(4), 667–699.

Murata, Y. (2003). Product Diversity, Taste Heterogeneity, and Geographic Distribution of Economic Activities: Market vs. Non-Market Interactions. *Journal of Urban Economics* 53(1), 126-144.

Nakayima, K., and T. Tabuchi (2008). Estimationg Interregional Utility Differentials, Working Paper, University of Tokyo.

Rauch, J. (1999). Networks versus Markets in International Trade. *Journal of International Economics* 48(1), 7–35.

Rauch, J., and V. Trindade (2002). Ethnic Chinese Networks in International Trade. *Review of Economics and Statistics* 84(1), 116–130.

Ravenstein, E. (1885). The laws of migration. *Proceedings of the Royal Statistical Society* XLVII(2), 167—235.

Redding, S. J., and D. M. Sturm (2008). The Costs of Remoteness: Evidence from German Division and Reunification. *American Economic Review* 98(5), 1766-97.

Santos Silva, J. M. C., and S. Tenreyro (2006). The Log of Gravity. Review of Economics and Statistics 88(4), 641-658.

Schwartz, A. (1973). Interpreting the Effect of Distance on Migration. *Journal of Political Economy* 81(5), 1153-1169.

Simon, B. (1992). The Perception of Ingroup and Outgroup Homogeneity: Re-Introducing the Ingroup Context, in: W. Stroebe and M. Hewstone (eds.) *European Review of Social Psychology*, 3, Chichester: Wiley, 1–30.

Sjaastad, L.A. (1962). The Costs and Returns of Human Migration. *Journal of Political Economy* 70(5), 80-93.

Sobel, J. (1985). A Theory of Credibility. *Review of Economic Studies* 52(4), 557–573.

Spolaore, E. and R. Wacziarg (2009). The Diffusion of Development. *Quarterly Journal of Economics* 124(2), 469-529.

Tabellini, G. (2007). Culture and Institutions: Economic Development in the Regions of Europe, Working Paper.

Tabellini, G. (2008). The Scope of Cooperation: Values and Incentives. *Quarterly Journal of Economics* 123(3), 905–950.

Tabuchi,T., and J.-F. Thisse (2002). Taste Heterogeneity, Labor mobility and Economic Geography. *Journal of Development Economics* 69 (1), 155–177.

Tanimoto, T. T. (1957). IBM Internal Report 17th Nov. 1957.

Watson, J. (1999). Starting Small and Renegotiation. *Journal of Economic Theory* 85(1), 52–90.

Wiesinger, P. (1983). Deutsche Dialektgebiete außerhalb des deutschen Sprachgebiets: Mittel-, Südost- und Osteuropa. , in: W. Besch et al. (ed.) *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. Zweiter Halbbd. (Handbücher zur Sprach- und Kommunikationswissenschaft, 1.2.). Berlin/New York, 807–900.

Wrede, F., W. Mitzka, and B. Martin (1927–1956). Deutscher Sprachatlas. Auf Grund des von Georg Wenker begründeten Sprachatlas des Deutschen Reichs. Marburg: Elwert.

Table 1a: Descriptive Statistics of Zero Flows, Average 2000–2006

	Mean of $\frac{M_{ij}}{L_i}$ (in 10,000)	Mean of all positive $\frac{M_{ij}}{L_i}$ (in 10,000)	Share of district pairs with $\frac{M_{ij}}{L_i} > 0$
German Inhabitants, entire population	0.711	0.735	96.75%
German Inhabitants, working-age population (18–65)	0.884	0.921	96.04%

Notes: Means are calculated across 192,282 observations for migration flows from every region i to j (i ≠ j and i=j=439). The number of positive observations is 186,025 (184,667) for the entire population (working-age population).

Table 1b: Descriptive Statistics of Main Variables, Average 2000–2006

	Entire Population				Working-Age Population (18–65)			
	Mean	Std. Dev.	Max. .	Min.	Mean	Std. Dev.	Max. .	Min.
Migrants Over Population in Origin (log)	-11.01	1.36	-15.27	-3.10	-10.78	1.35	-14.87	-3.11
Migrants Over Population in Origin and Destination (log)	-22.85	1.21	-26.83	-15.12	-22.15	1.20	-26.02	-14.44
Geographic Distance (log)	5.58	0.63	0.07	6.74	5.58	0.63	0.07	6.74
Dialect Similarity (log)	3.30	0.29	1.95	4.08	3.30	0.29	1.95	4.08
Travel Distance (log)	5.46	0.53	2.17	6.53	5.46	0.53	2.17	6.53

Notes: Summary statistics are calculated across 192,282 observations for migration flows from every region i to j (i ≠ j and i=j=439.

Table 1c: Descriptive Statistics of Control Variables, Average 2000–2006

Variable	Mean	Std. Dev.	Min	Max
Different State Dummy	0.893	0.309	0	1
East/West Dummy	0.380	0.485	0	1
Historic Border Dummy	0.840	0.367	0	1
Δ Access to High-Speed Infrastructure	47.359	37.840	0	294.896
Δ Expellees (1950)	0	13.499	-36.150	36.150
Δ Mineralogy of the Subsoil	0	1.461	-5	5
Δ Mining	0	117.596	-345.095	345.095
Δ Parental Soil	0	2.994	-8	8
Δ Reachability Next European Agglomeration	41.324	30.319	0	211.676
Δ Reachability Next National Agglomeration	35.337	25.907	0	170.667
Δ Slope	0	2.221	-7.273	7.273
Δ Catholics (1890)	3.096	2.572	0	7

Notes: Summary statistics are calculated across 192,282 observations for migration flows from every region i to j (i ≠ j and i=j=439.

Table 2: FE-OLS Regressions

	Entire Population				Working-Age Population (18–65)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i L_j}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i L_j}\right)$
Geographic Distance	-1.493***		-1.452***	-1.452***	-1.481***	-	-1.440***	-1.440***
	(0.012)		(0.012)	(0.012)	(0.012)		(0.013)	(0.013)
Dialect Similarity		2.072***	0.157***	0.157***	-	2.059***	0.156***	0.156***
		(0.031)	(0.017)	(0.017)		(0.031)	(0.017)	(0.017)
F-Statistic	***	***	***	***	***	***	***	***
R^2	0.744	0.475	0.744	0.670	0.758	0.491	0.758	0.687
N	186,025	186,025	186,025	186,025	184,667	184,667	184,667	184,667

Notes: The table reports the results from OLS regressions of geographic distance and language similarity on the log of the number of German migrants from region i to j divided by the origin region's number of inhabitants (working-age inhabitants) in column (1)–(3) and (5)–(7) and is divided by the product of the population (working-age population) in the origin and destination region in columns (4) and (8). We use a fixed effects specification with fixed effects for both origin region i and target region j. Zero flows drop out in these specifications. Geographic Distance and Language Similarity are logs in all specifications. Robust standard errors are reported in parenthesis.

**** statistically significant at the 1% level; ** statistically significant at the 5% level; * statistically significant at the 10% level.*

Table 3: Regressions with Different Distance Measures (Entire Population)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$
Travel Distance	-1.773***	-1.718***	-0.250***	-	-	-	-
	(0.014)	(0.015)	(0.044)				
Geographic Distance	-	-	-1.253***	-1.488***	-1.213***	-1.251***	-1.275***
			(0.036)	(0.012)	(0.012)	(0.011)	(0.0123)
Dialect Similarity	-	0.173***	0.144***	0.135***	0.119***	0.094***	0.079***
		(0.017)	(0.017)	(0.017)	(0.016)	(0.016)	(0.016)
East/West Dummy	-	-	-	0.123***	-	0.133***	0.094***
				(0.017)		(0.015)	(0.015)
Different State Dummy	-	-	-	-	-0.788***	-0.792***	-0.777***
					(0.019)	(0.019)	(0.019)
Δ Reachability Next National Agglomeration	-	-	-	-	-	-	-0.001***
							(0.000)
Δ Reachability Next European Agglomeration	-	-	-	-	-	-	0.002***
							(0.000)
F-Statistic	***	***	***	***	***	***	***
R²	0.731	0.732	0.745	0.745	0.764	0.766	0.767
N	186,025	186,025	186,025	186,025	186,025	186,025	186,025

*Notes: Notes: Columns (1)–(3) of the table report the results from OLS regressions of travel distance in car minutes, language similarity, and geographic distance on the log of the number of German migrants from region i to j divided by the origin region's number of inhabitants. Columns (4)–(7) additionally control for distance-related reachability indicators that are calculated in minutes. All controls are calculated as absolute differences between region pair i and j. We use a fixed effects specification with fixed effects for both origin region and target region. Geographic Distance and Language Similarity are logs in all specifications. Robust standard errors are reported in parenthesis. *** statistically significant at the 1% level; ** statistically significant at the 5% level; * statistically significant at the 10% level.*

Table 4: FE-OLS Regressions with Controls for Historic Disparities (Entire Population)

	(1)	(2)	(3)	(4)	(5)	(6)
	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$
Geographic Distance	-1.251***	-1.251***	-1.251***	-1.263***	-1.263***	-1.580***
	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)	(0.016)
Dialect Similarity	0.094***	0.094***	0.094***	0.088***	0.087***	0.055**
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.022)
Δ Mineralogy of the Subsoil	0.094***	0.383***	-0.052***	-0.050***	-0.050***	-
	(0.016)	(0.010)	(0.010)	(0.010)	(0.010)	
Δ Parental Soil	0.017***	0.178***	-0.054***	-0.050***	-0.049***	-
	(0.006)	(0.008)	(0.007)	(0.007)	(0.007)	
Δ Slope	-	-0.201***	-0.019***	-0.020***	-0.020***	-
		(0.009)	(0.007)	(0.007)	(0.007)	
Δ Mining	-	-	0.007***	0.007***	0.007***	-
			(0.000)	(0.000)	(0.000)	
Δ Catholics (1890)	-	-	-	0.012***	0.012***	-
				(0.002)	(0.002)	
Crossing Historic Boarder Dummy	-	-	-	-	-0.010	-
					(0.013)	
Δ Expellees (1950)	-	-	-	-	-	0.102***
						(0.002)
Different State and East/West Control	YES	YES	YES	YES	YES	No
F Statistic	***	***	***	***	***	***
R^2	0.766	0.766	0.766	0.766	0.766	0.797
N	186,025	186,025	186,025	186,025	186,025	98,906

Notes: The table reports the results from OLS regressions of geographic distance and language similarity on the log of the number of German migrants from region i to j divided by the origin region's number of inhabitants conditional on controls for the type of soil, the log of the land slope measured as median maximum and minimum elevations in meters, the proximity to a mining school as an indicator for the exploitation of mineral resources, the share of Catholics, controls for historic and present borders, and the population share of expellees after WWII. All controls are calculated as differences between regions i and j. We use a fixed effects specification with fixed effects for both origin region and target region. Geographic Distance and Language Similarity are logs in all specifications. Robust standard errors are reported in parenthesis.

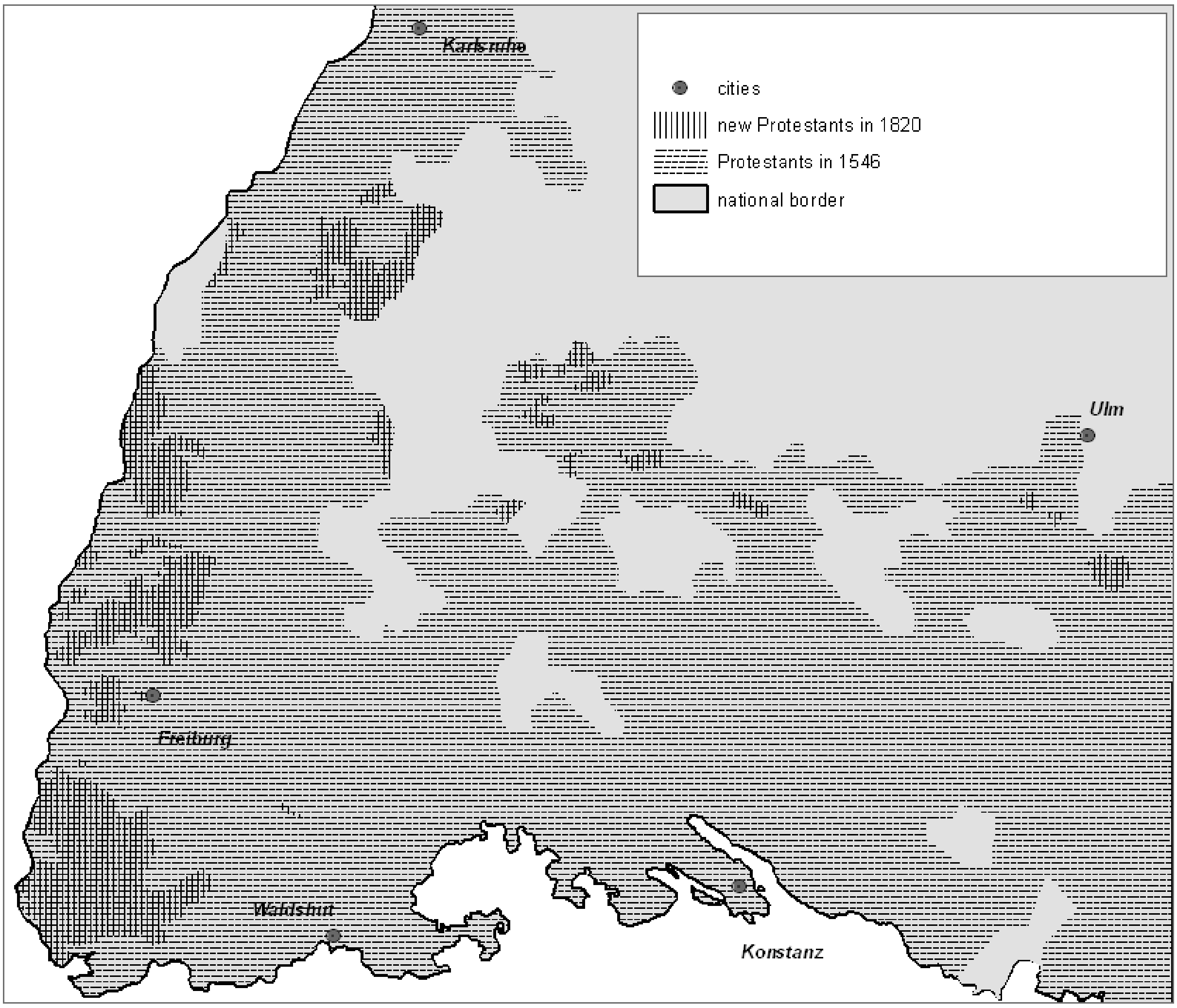
***** statistically significant at the 1% level; ** statistically significant at the 5% level; * statistically significant at the 10% level.**

Table 5: Regressions Coping with Zero Flows and Heteroskedasticity

	Entire Population			Working-Age Population (18–65)		
	(1) $\ln\left(\frac{1+M_{ij}}{L_i}\right)$	(2) $\ln\left(\frac{M_{ij}}{L_i}\right)$	(3) $\frac{M_{ij}}{L_i}$	(4) $\ln\left(\frac{1+M_{ij}}{L_i}\right)$	(5) $\ln\left(\frac{M_{ij}}{L_i}\right)$	(6) $\frac{M_{ij}}{L_i}$
Geographic Distance	-1.131***	-1.478***	-1.760***	-1.079***	-1.477***	-1.736***
	(0.009)	(0.004)	(0.012)	(0.009)	(0.003)	(0.011)
Dialect Similarity	0.072***	0.165***	0.340***	0.061***	0.167***	0.364***
	(0.012)	(0.008)	(0.044)	(0.012)	(0.007)	(0.038)
Zero Flow Dummy	-0.155***	-	-	-0.088****	-	-
	(0.013)			(0.012)		
First Stage						
Geographic Distance	-	-1.496***	-	-	-1.409***	-
		(0.035)			(0.030)	
Language Similarity	-	0.187***	-	-	0.214***	-
		(0.042)			(0.038)	
Mills Lambda	-	0.544***	-	-	0.669***	-
		(0.018)			(0.016)	
F-Statistic	***	-	-	***	-	-
R²	0.773	-	-	0.774	-	-
Chi²	-	***	***	-	***	***
Pseudo R²	-	-	0.195	-	-	0.199
Censored Observations	-	6,257	-	-	7615	-
N	192,282	192,282	192,282	192,282	192,282	192,282

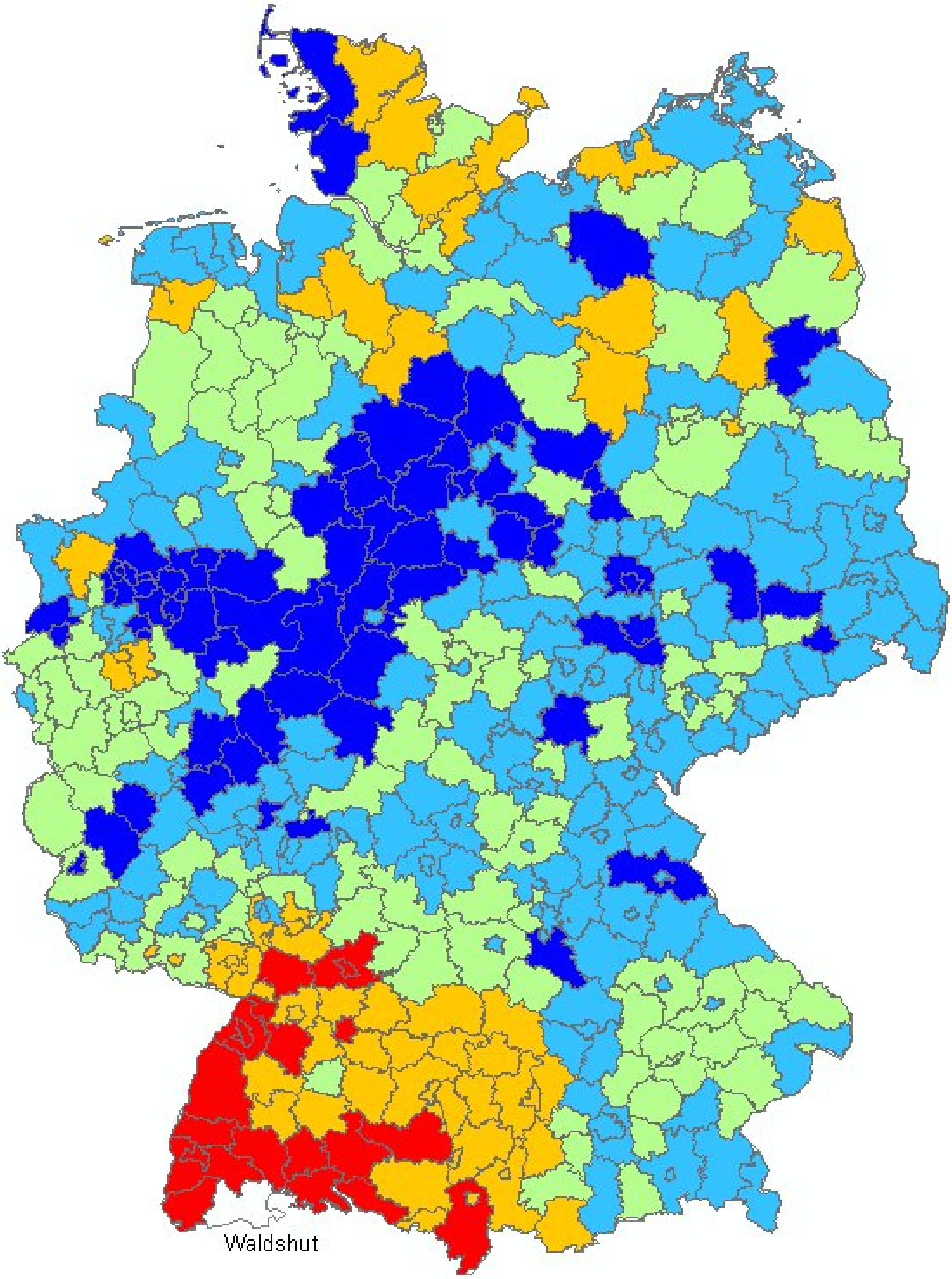
Notes: Columns (1) and (4) of the table report OLS regressions of geographic distance and language similarity on the log of the number of German migrants from region i to j divided by the origin region's number of inhabitants (working-age inhabitants). To keep zero flows in a log specification, zero flows are coded as 1 instead of while controlling for a zero flow dummy (Columns 1 and 4). Columns (2) and (5) of the table report the results from a Heckman selection model for estimations of geographic distance and language similarity on the log of the number of German migrants from region i to j divided by the origin region's number of inhabitants (working-age inhabitants) on geographic distance and language similarity. The first-stage selection considers the probability of a zero flow of migrants between region i and j. Columns (3) and (6) report Poisson regressions of geographic distance and language similarity on the number of German migrants from region i to j divided by the origin region's number of inhabitants (working-age inhabitants). We use a fixed effects specification with fixed effects for both origin region and target region. Geographic Distance and Language Similarity are logs in all specifications. Robust standard errors are reported in parenthesis

Figure 1: Distribution of religious faith in southern Germany



Sources: Steger, H., E. Gabriel, and V. Schupp (eds.) (1989 ff.): Südwestdeutscher Sprachatlas. Marburg: Elwert; Großer Historischer Weltatlas (1953 ff.). München: Bayerischer Schulbuch-Verlag.

Figure 2: Language Similarity to the Waldshut district



Notes: Similarity of all districts to the reference point Waldshut (white spot). Red indicates highest familiarity and yellow indicates higher familiarity while the green and blue indicate less familiarity.

Figure 3: Religious faith compared to language Similarity (Waldshut district)

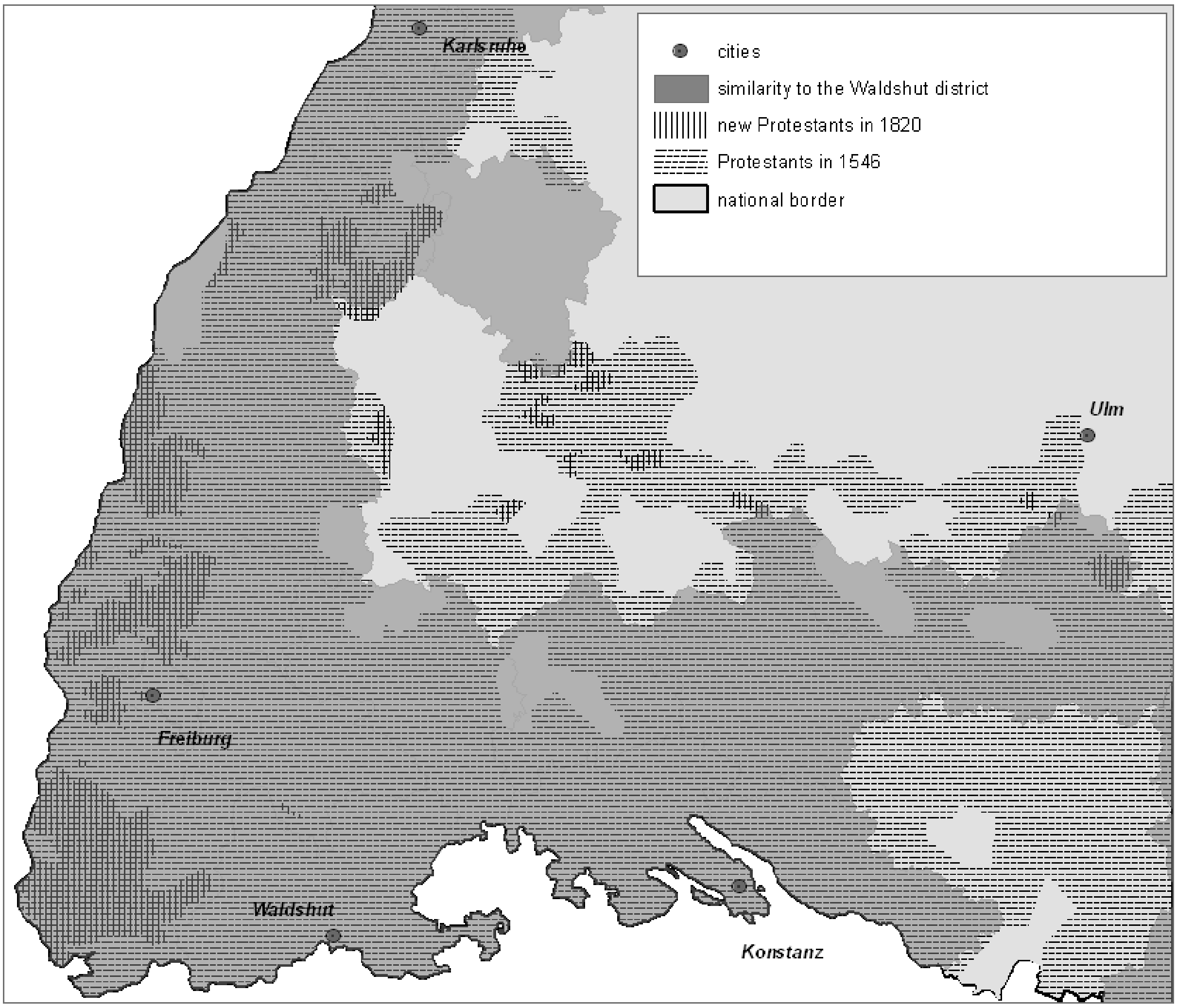
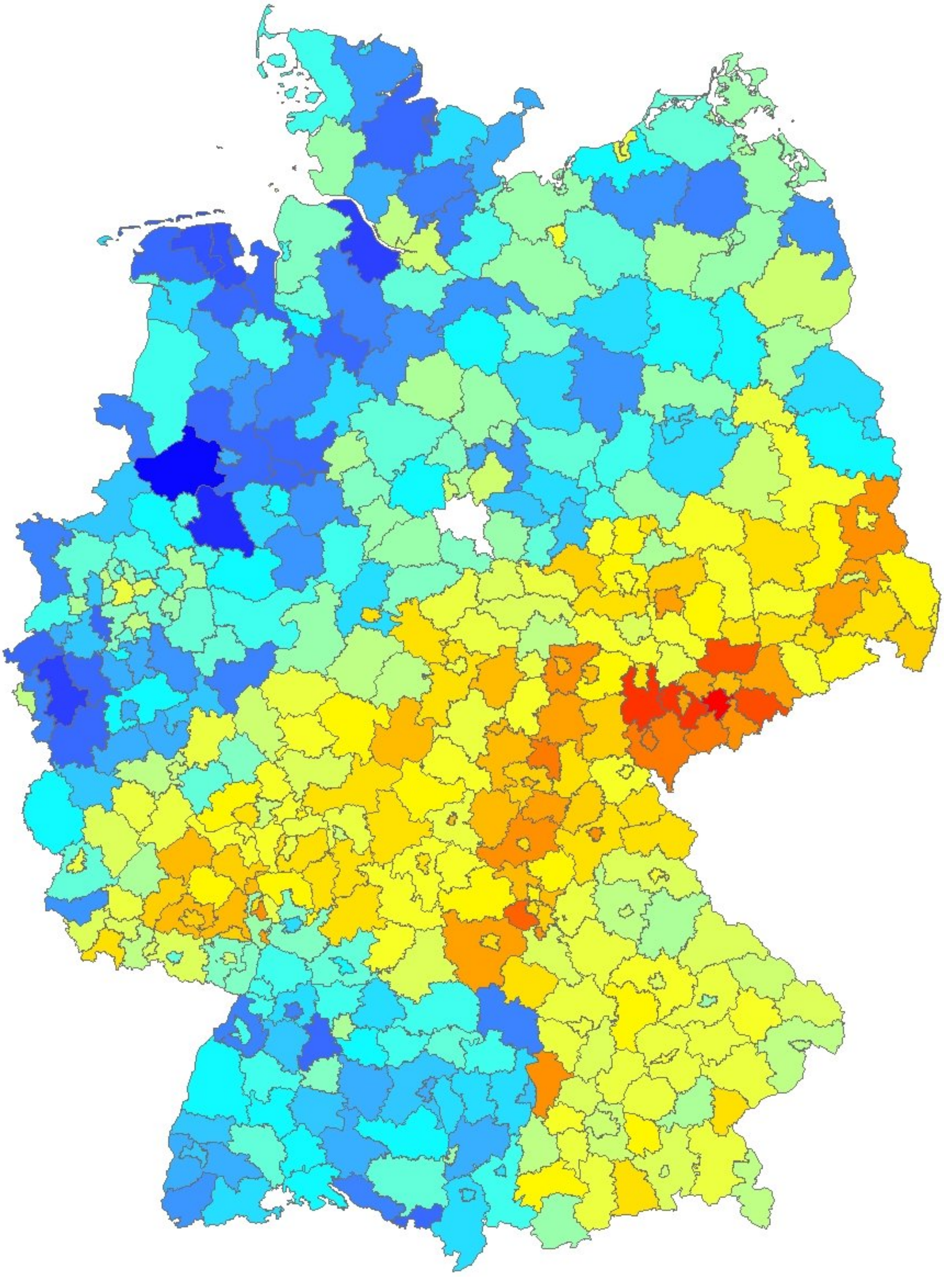
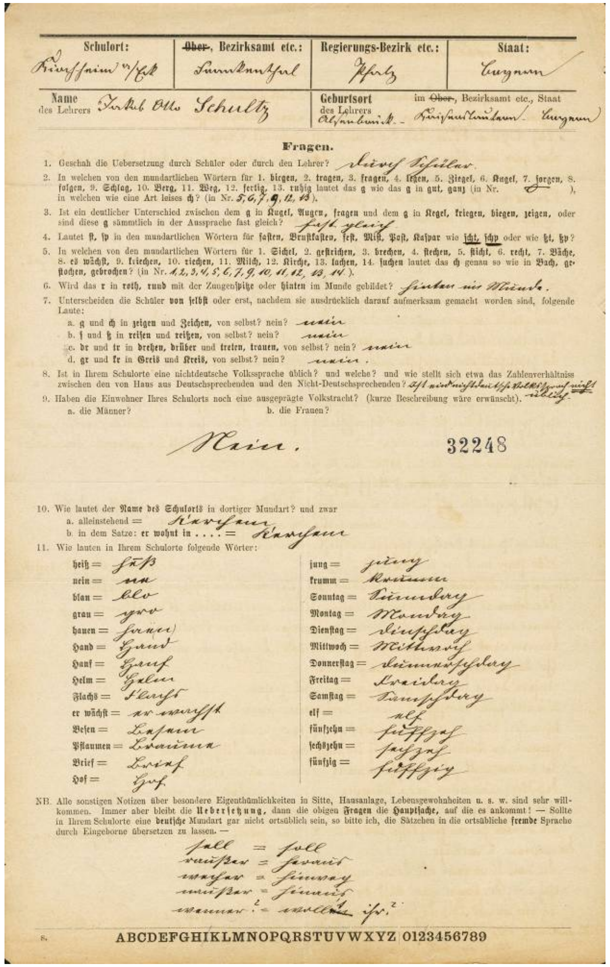


Figure 4: The language enclave Goslar



Notes: Similarity of all districts to the reference point Goslar (white spot). Red indicates highest familiarity and warmer tints (yellow and green) indicate higher familiarity while the bluish tints indicate less familiarity.

Figure A1: Exemplary Questionnaire of the Language Survey



Schulort:	~~Ober~~, Bezirksamt etc.:	Regierungs-Bezirk etc.:	Staat:
Kirchheim a/Eck	Frankenthal	Pfalz	Bayern

Name des Lehrers Jakob Otto Schultz

Geburtsort des Lehrers Alsenbrück – im ~~Ober~~, Bezirksamt etc., Staat Kirchheimbolanden – Bayern

Fragen.

1. Geschah die Uebersetzung durch Schüler oder durch den Lehrer? Durch Schüler.
2. In welchen von den mundartlichen Wörtern für 1. **biegen**, 2. **tragen**, 3. **fragen**, 4. **leſen**, 5. **Ziegel**, 6. **Kugel**, 7. **ſorgen**, 8. **folgen**, 9. **Schlag**, 10. **Berg**, 11. **Weg**, 12. **fertig**, 13. **ruhig** lautet das **g** wie das **g** in **gut**, **ganz** (in Nr. —), in welchen wie eine Art leiſes **ch**? (in Nr. 5, 6, 7, 9, 12, 13).
3. Ist ein deutlicher Unterschied zwischen dem **g** in **Kugel**, **Augen**, **fragen** und dem **g** in **Kegel**, **kriegen**, **biegen**, **zeigen**, oder sind diese **g** sämmtlich in der Aussprache fast gleich? fast gleich
4. Lautet **ſt**, **ſp** in den mundartlichen Wörtern für **faſten**, **Bruſtkaſten**, **feſt**, **Miſt**, **Poſt**, **Kaſpar** wie **ſcht**, **ſchp** oder wie **ßt**, **ßp**?
5. In welchen von den mundartlichen Wörtern für 1. **Sichel**, 2. **geſtrichen**, 3. **brechen**, 4. **ſtechen**, 5. **ſticht**, 6. **recht**, 7. **Bäche**, 8. **es wächſt**, 9. **kriechen**, 10. **riechen**, 11. **Milch**, 12. **Kirche**, 13. **lachen**, 14. **ſuchen** lautet das **ch** genau so wie in **Bach**, **geſtochen**, **gebrochen**? (in Nr. 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14).
6. Wird das **r** in **roth**, **rund** mit der Zungenſpitze oder **hinten** im Munde gebildet? hinten im Munde.
7. Unterscheiden die Schüler **von ſelbſt** oder erst, nachdem sie ausdrücklich darauf aufmerksam gemacht worden sind, folgende Laute:
   a. **g** und **ch** in **zeigen** und **Zeichen**, von selbst? nein? nein
   b. **ſ** und **ß** in **reiſen** und **reißen**, von selbst? nein? nein
   c. **dr** und **tr** in **drehen**, **drüber** und **treten**, **trauen**, von selbst? nein? nein
   d. **gr** und **kr** in **Greis** und **Kreis**, von selbst? nein? nein.
8. Ist in Ihrem Schulorte eine nichtdeutsche Volkssprache üblich? und welche? und wie stellt sich etwa das Zahlenverhältniss zwischen den von Haus aus Deutschsprechenden und den Nicht-Deutschsprechenden? Ist eine nichtdeutsche Volkssprache nicht üblich.
9. Haben die Einwohner Ihres Schulorts noch eine ausgeprägte Volkstracht? (kurze Beschreibung wäre erwünscht).
   a. die Männer? b. die Frauen?

Nein.

32248

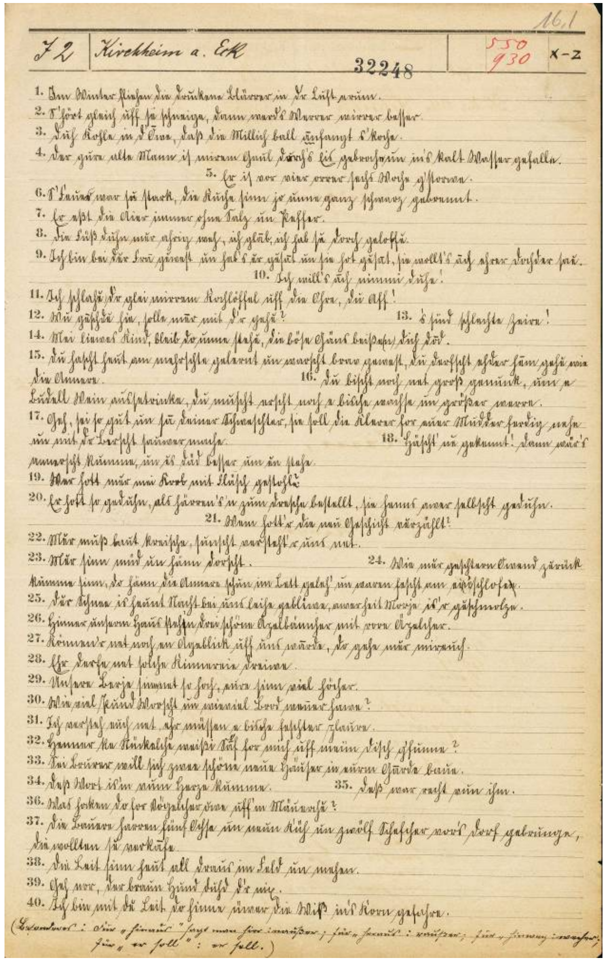
10. Wie lautet der **Name des Schulorts** in dortiger Mundart? und zwar
    a. alleinstehend = Karchem
    b. in dem Satze: **er wohnt in** . . . . = Karchem
11. Wie lauten in Ihrem Schulorte folgende Wörter:

heiß =	häß	**jung** =	jung
nein =	nä	**krumm** =	krumm
blau =	blo	**Sonntag** =	Sunndag
grau =	gro	**Montag** =	Mondag
hauen =	haue	**Dienſtag** =	Dinschdag
Hand =	Hand	**Mittwoch** =	Mittwoch
Hanf =	Hanf	**Donnerſtag** =	Dunnerschdag
Helm =	Helm	**Freitag** =	Freidag
Flachs =	Flachs	**Samſtag** =	Samschdag
er wächſt =	er wachst	**elf** =	elf
Beſen =	Besem	**fünfzehn** =	fuffzeh
Pflaumen =	Praume	**ſechszehn** =	sechzeh
Brief =	Brief	**fünfzig** =	fuffzig
Hof =	Hof		

NB. Alle sonstigen Notizen über besondere Eigenthümlichkeiten in Sitte, Hausanlage, Lebensgewohnheiten u. s. w. sind sehr willkommen. Immer aber bleibt die **Ueberſetzung**, dann die obigen **Fragen** die **Hauptſache**, auf die es ankommt! — Sollte in Ihrem Schulorte eine **deutſche** Mundart gar nicht ortsüblich sein, so bitte ich, die Sätzchen in die ortsübliche **fremde** Sprache durch Eingeborne übersetzen zu lassen. —

sall = soll
raußer = heraus
neiner = hinein
naußer = hinaus
wanner? = wollt ihr?

8. ABCDEFGHIKLMNOPQRSTUVWXYZ 0123456789



16,1

F 2	Kirchheim a. Eck 32248		550 930	X-Z

1. Im Winter fliehe die dürre Blärrer in dr Luft erum.
2. S'hört gleich uff ze schneie, dann werd's Werrer wirrer besser.
3. Duh Kohle in d'Ove, daß die Millich bald anfangt z'koche.
4. Der gute alte Mann is mirem Gaul dorch's Eis gebroche un in's kalt Wasser gefalle.
5. Er is vor vier orre sechs Woche g'storwe.
6. S'Feuer war ze stark, die Kuche sinn jo unne ganz schwarz gebrennt.
7. Er eßt die Eier immer ohne Salz un Peffer.
8. Die Füß duhn mer arig weh, ich glab, ich hab se dorch gelofe.
9. Ich bin bei der Fra geweßt un hab's er gesat un sie hot gesat, sie wollt's ach ihrer Dochter sa.
10. Ich will's ach nimmi duhe!
11. Ich schlage dr glei mirrem Kochlöffel uff die Ohre, du Aff!
12. Wu gehschde hin, solle mer mit dr gehe?
13. 's sind schlechte Zeite!
14. Mei liewes Kind, bleib do unne stehe, die böse Gäns beiße dich dod.
15. Du hascht heut am mehrschte gelernt un warscht brav geweßt, du derfscht ehder hem gehe wie die Annere.
16. Du bischt noch net groß genunk, um e Bottell Wein auszetrinke, du mußcht erscht noch e bische wachse un größer werre.
17. Geh, sei so gut un sa deiner Schweschter, sie soll die Klerer for eier Mudder ferdig mache un mit dr Berscht sauwer mache.
18. Hätscht'n gekannt! dann wär's annerscht kumme, un 's dät besser um'n stehe.
19. Wer hot mer mei Korb mit Fleisch gestohle?
20. Er hot so gedahn, als härren s'n zum Dresche bestellt, sie hann awer selbscht gedahn.
21. Wem hot'r die nei Geschicht verzählt?
22. Mer muß laut kreische, sunscht versteht'r uns net.
23. Mir sinn mied un hann Dorscht.
24. Wie mer geschtern Owend zerick kumme sinn, do hänn die Annere schun im Bett geleh' un waren fescht am eigeschlofe.
25. Der Schnee is heunt Nacht bei uns leije gebliwe, awer heit Morje is'r geschmolze.
26. Hinner unserm Haus stehn drei schöne Appelbäumcher mit rote Äppelcher.
27. Könnener net noch en Agablick uff uns warde, do gehe mer mit euch.
28. Ihr derfe net solche Kinnereie treiwe.
29. Unsere Berge sinn net so hoch, eire sinn viel höcher.
30. Wie viel Pfund Worscht un wieviel Brod wollener hawe?
31. Ich versteh euch net, ihr müsse e bische lauter schwätze.
32. Hanner ke Stückelche weißi Seef for mich uff mein Disch gfunne?
33. Sei Bruder will sich zwee schöne neue Häuser in eirem Garde baue.
34. Des Wort is'm vun Herze kumme.
35. Des war recht vun ihne.
36. Was hocke do for Vögelcher owe uff'm Mäuerche?
37. Die Bauere hawwe fünf Ochse un nein Küh un zwölf Schäfcher vor's Dorf gebrunge, die wollten se verkafe.
38. Die Leit sinn heut all drauß im Feld un mehe.
39. Geh nor, der braun Hund duht dr nix.
40. Ich bin mit de Leit do hinne üwer die Wiß ins Korn gefahre.

(Bemerkung: Für „hinaus" sagt man hier: naußer; für „heraus": raußer; für „hinein": neiner; für „er soll": er sall.)

Table A1: Historic Locations of Mining Academies

Location of Mining Academy	Year of Founding	District	Federal State
Aachen	1870	Aachen	North Rhine-Westphalia
Bochum	1816	Bochum	North Rhine-Westphalia
Clausthal	1775	Goslar	Lower Saxony
Eisleben	1798	Mansfeld-Südharz	Saxony-Anhalt
Essen	1808	Essen	North Rhine-Westphalia
Freiberg	1765	Mittelsachsen	Saxony
Königshütte	1803	Harz	Saxony-Anhalt
Saarbrücken	1816	Saarbrücken	Saarland
Siegen	1818	Siegen-Wittgenstein	North Rhine-Westphalia
Bad Steben	1793	Hof	Bavaria
Zwickau	1862	Zwickauer Land	Saxony
Tarnowitz (Upper Silesia)	1803	-	Poland
St. Joachimsthal (Bohemia)	1717	-	Czech Republic
Waldenburg, (Lower Silesia)	1838	-	Poland

Table A2: Regressions with Different Distance Measures (Working-Age Population)

	(1) $\ln\left(\frac{M_{ij}}{L_i}\right)$	(2) $\ln\left(\frac{M_{ij}}{L_i}\right)$	(3) $\ln\left(\frac{M_{ij}}{L_i}\right)$	(4) $\ln\left(\frac{M_{ij}}{L_i}\right)$	(5) $\ln\left(\frac{M_{ij}}{L_i}\right)$	(6) $\ln\left(\frac{M_{ij}}{L_i}\right)$	(7) $\ln\left(\frac{M_{ij}}{L_i}\right)$
Travel Distance	-1.759*** (0.014)	-1.705*** (0.015)	-0.250*** (0.045)	-	-	-	-
Geographic Distance	-	-	-1.241*** (0.036)	-1.494*** (0.011)	-1.197*** (0.013)	-1.252*** (0.011)	-1.252*** (0.011)
Dialect Similarity	-	0.172*** (0.017)	0.142*** (0.017)	0.121*** (0.017)	0.116*** (0.016)	0.080*** (0.015)	0.080*** (0.016)
East/West Dummy	-	-	-	0.184*** (0.017)	-	0.195*** (0.015)	0.195*** (0.015)
Different State Dummy	-	-	-	-	-0.805*** (0.019)	-0.810*** (0.019)	-0.810*** (0.019)
Δ Reachability Next National Agglomeration	-	-	-	-	-	-	0.388*** (0.009)
Δ Reachability Next European Agglomeration	-	-	-	-	-	-	0.021*** (0.006)
F-Statistic	***	***	***	***	***	***	***
R²	0.745	0.746	0.759	0.761	0.779	0.782	0.782
N	184,667	184,667	184,667	184,667	184,667	184,667	184,667

Notes: Notes: Columns (1)–(3) of the table report the results from OLS regressions of travel distance in car minutes, language similarity, and geographic distance on the log of the number of German migrants from region i to j divided by the origin region's number of inhabitants. Columns (4)–(7) additionally control for distance-related reachability indicators that are calculated in minutes. All controls are calculated as absolute differences between region pair i and j. We use a fixed effects specification with fixed effects for both origin region and target region. Geographic Distance and Language Similarity are logs in all specifications. Robust standard errors are reported in parenthesis.

**** statistically significant at the 1% level; ** statistically significant at the 5% level; * statistically significant at the 10% level.*

Table A3: FE-OLS Regressions with Alternative Cost Function Specifications

	Entire Population			Working-Age Population (18–65)		
	(1)	(2)	(3)	(4)	(5)	(6)
	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$
Geographic Distance	-1.452***	-	-	-1.440***	-	-
	(0.012)			(0.013)		
Dialect Similarity	0.157***	-	-	0.156***	-	-
	(0.017)			(0.017)		
Geographic Distance (no log)	-	-0.004***	-0.013***	-	-0.004***	-0.012***
		(0.000)	(0.000)		(0.000)	(0.000)
Language Similarity (no log)	-	0.034***	-0.072***	-	0.034***	-0.071***
		(0.001)	(0.003)		(0.001)	(0.003)
Geographic Distance² (no log)	-	-	0.000***	-	-	0.000***
			(0.000)			(0.000)
Language Similarity² (no log)	-	-	0.002***	-	-	0.002***
			(0.000)			(0.000)
F-Statistic	***	***	***	***	***	***
R²	0.744	0.639	0.706	0.758	0.652	0.720
N	186,025	186,025	186,025	184,667	184,667	184,667

Notes: The table reports the results from OLS regressions of geographic distance and language similarity on the log of the number of German migrants from region i to j divided by the origin region's number of inhabitants (working-age inhabitants). We use a fixed effects specification with fixed effects for both origin region and target region. Geographic Distance and Language Similarity are logs in the specifications reported in columns (1) and (4) and no logs in all other specifications. Robust standard errors are reported in parenthesis.

**** statistically significant at the 1% level; ** statistically significant at the 5% level; * statistically significant at the 10% level.*

Table A4: FE-OLS Regressions with Controls for Historic Disparities (Working-Age Population)

	(1)	(2)	(3)	(4)	(5)	(6)
	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$	$\ln\left(\frac{M_{ij}}{L_i}\right)$
Geographic Distance	-1.252*** (0.011)	-1.252*** (0.011)	-1.252*** (0.011)	-1.264*** (0.012)	-1.264*** (0.012)	-1.584*** (0.016)
Dialect Similarity	0.080*** (0.016)	0.080*** (0.016)	0.080*** (0.016)	0.074*** (0.015)	0.074*** (0.015)	0.045** (0.022)
Δ Mineralogy of the Subsoil	0.388*** (0.009)	0.414*** (0.009)	-0.031*** (0.009)	-0.029*** (0.009)	-0.029*** (0.009)	-
Δ Parental Soil	0.021*** (0.006)	0.184*** (0.008)	-0.047*** (0.007)	-0.043*** (0.007)	-0.043*** (0.007)	-
Δ Slope	-	-0.203*** (0.009)	-0.031*** (0.007)	-0.031*** (0.007)	-0.031*** (0.007)	-
Δ Mining	-	-	0.007*** (0.000)	0.007*** (0.000)	0.007*** (0.000)	-
Δ Catholics (1890)	-	-	-	0.012*** (0.002)	0.012*** (0.002)	-
Crossing Historic Border Dummy	-	-	-	-	0.002*** (0.013)	-
Δ Expellees (1950)	-	-	-	-	-	0.107*** (0.002)
Different State and East/West Dummy	YES	YES	YES	YES	YES	-
F Statistic	***	***	***	***	***	***
R²	0.782	0.782	0.782	0.782	0.782	0.812
N	184,667	184,667	184,667	184,667	184,667	98,313

Notes: The table reports the results from OLS regressions of geographic distance and language similarity on the log of the number of German migrants from region i to j divided by the origin region's number of inhabitants conditional on controls for the type of soil, the log of the land slope measured as median maximum and minimum elevations in meters, the proximity to a mining school as an indicator for the exploitation of mineral resources, the share of Catholics, controls for historic and present borders, and the population share of expellees after WWII. All controls are calculated as differences between regions i and j. We use a fixed effects specification with fixed effects for both origin region and target region. Geographic Distance and Language Similarity are logs in all specifications. Robust standard errors are reported in parenthesis.

**** statistically significant at the 1% level; ** statistically significant at the 5% level; * statistically significant at the 10% level.*

Table A5: Extended Data Description

Variable	**Description and Source**
Geographic Distance	The geographic distance between two districts is calculated as Eucledian distance between each pair of districts' centroids.
Travel Distance	The travel distance is calculated in car minutes from one district's capital to the other.
Historic Border Dummy	Historic borders refer to 38 member states and 4 independent cities that were part of the German Confederation at its foundation in 1815. Data are taken from a map in *Putzger – Historischer Weltatlas*, 89th edition, 1965.
Share of Expellees	The share of expellees is calculated as the number of expellees over the district's local population. Data stem from the population census in 1950. These data were published by the Minster for Expellees (*Bundesminister für Vertriebene*) in 1952.
Mineralogy of the Subsoil	This variable represents the minerals in the subsoil, i.e., the intermediate layer between the topsoil and the bedrock. This variable is a scale of the following eight characteristics (only five apply to Germany). 1. KQ = 1/1 Minerals + Quartz 2. KX = 1/1 Min. + Oxy. and Hydroxy. 3. MK = 2/1 and 1/1 Minerals 4. (M = 2/1 and 2/1/1 non swel. Minerals) 5. MS = Swel. and non swel. 2/1 Minerals 6. S = Swelling 2/1 Minerals 7. (TV = Vitric Minerals) 8. (TO = Andic Minerals) Data stem from the European Soil Database (esdb) and are compiled by the European Soil Data Centre. A more detailed description of the variable and its generation process is provided in Combes *et al.* (in press). We are deeply indebted to Gilles Duranton for providing these data.

Variable	Description and Source (continued)
Parental Soil	This variable represents the dominant parent material in the soil. This variable is a scale of the following eight characteristics: 1. consolidated-clastic-sedimentary rocks 2. sedimentary rocks (chemically precipitated, evaporated, or organogenic or biogenic in origin) 3. igneous rocks 4. metamorphic rocks 5. unconsolidated deposits (alluvium, weathering residuum, and slope deposits) 6. is unconsolidated glacial deposits/glacial drift 7. eolian deposits 8. organic materials Data from the European Soil Database (esdb) and are compiled by the European Soil Data Centre. A more detailed description of the variable and its generation process is provided in Combes *et al.* (in press). We are deeply indebted to Gilles Duranton for providing these data.
Slope	Slope is measured as the difference between the median maximum and minimum elevations in meters. We are deeply indebted to Gilles Duranton for providing these data.
Mining	Mining is calculated as the distance from the district's centoid to the closest mining academy that was founded before 1880. A list of mining academies is provided in Table A1.
Access to High-Speed Infrastructure	This indicator characterizes the availability of modern transportation systems measured as reachability in minutes by car. This indicator is published by the Federal Office for Building and Regional Planning (cf. Maretzke 2005).
Reachability Next European Agglomeration	This indicator measures the reachability of European metropolis in combined road and air traffic in minutes. This indicator is published by the Federal Office for Building and Regional Planning (cf. Maretzke 2005).
Reachability Next National Agglomeration	This indicator reports the reachability of the nearest three national or international agglomerations in combined road and rail traffic in minutes. This indicator is published by the Federal Office for Building and Regional Planning (cf. Maretzke 2005).
Share of Catholics (1890)	The districts' historic shares of Catholics in 1890 are calculated from a map in *Meyers Konversations Lexikon*, 4th edition, 1885–1892.
Share of Catholics (1987)	The districts' shares of Catholics are taken from the last population census in 1987.