Barcelona School of Economics

WORKING PAPERS

Borders within Europe

BSE Working Paper 1229

December 2020 (Revised: February 2024)

Marta Santamaría, Jaume Ventura, Ugur Yesilbayraktar

bse.eu/research

Borders within Europe *

Marta Santamaría, Jaume Ventura and Uğur Yeşilbayraktar

February 23, 2024

Abstract

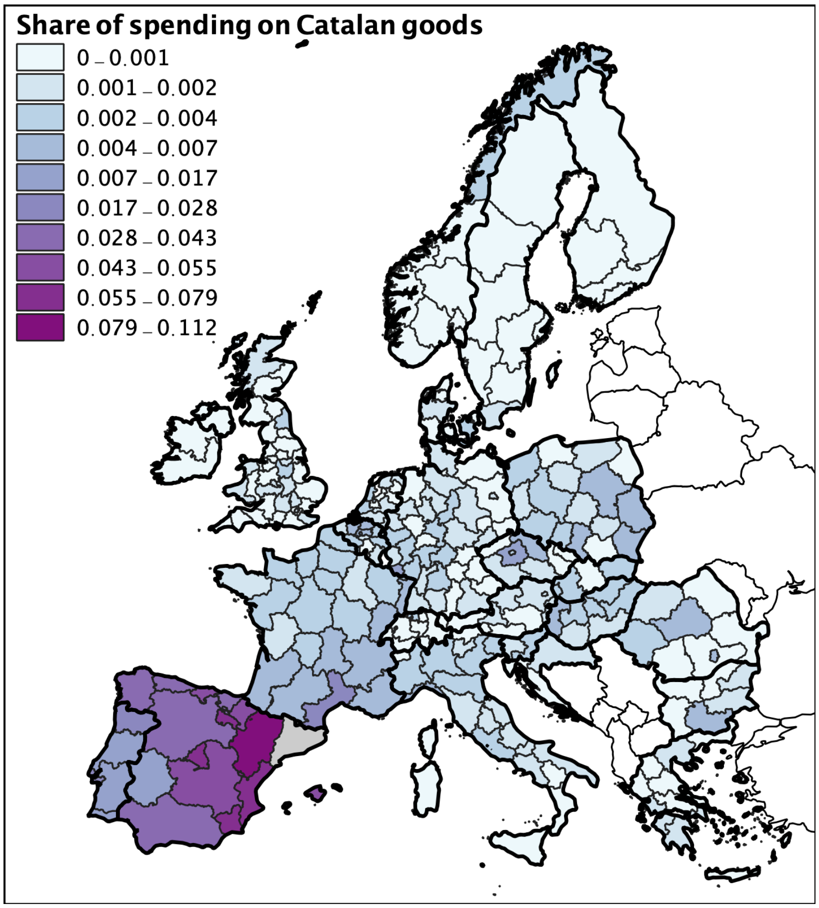
Are country borders still an impediment to trade flows within Europe? Using a microlevel survey with 3 million annual shipments of goods, we construct a matrix of bilateral trade for 269 European regions. Take two similar region pairs, one containing regions in different countries and the other containing regions in the same country. The market share of the origin region in the destination region for the international pair is 17.5 percent that of the intranational pair. Across industries, this estimate ranges from 12.3 to 38.9 percent. For post-1910 borders, this estimate is 28.8 percent. The implication is clear: Europe is far from having a single market.

JEL Classification: D71, F15, F55, H77, O57

Keywords: Border effect, European integration, regional trade.

*Marta Santamaría: University of Warwick. Jaume Ventura: CREI, Universitat Pompeu Fabra and BSE. Uğur Yeşilbayraktar: University of Duisburg-Essen. E-mail: marta.santamaria@warwick.ac.uk, jventura@crei.cat, ugur.yesilbayraktar@uni-due.de. We acknowledge financial support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program Advanced Grant (693512 - GEPPS), the Spanish Ministry of Economy and Competitiveness through the Severo Ochoa Programme for Centres of Excellence in RD (CEX2019-000915-S) and from the Generalitat de Catalunya through the CERCA and SGR Programme (2017-SGR-1393). We thank Alejandro Cuñat, Manuel García-Santana, Keith Head, Joan Monràs, David Nagy, Giacomo Ponzetto, Maria Ptashkina and seminar/conference participants at CREI, Universitat Pompeu Fabra, London School of Economics, University of Warwick, University of Michigan, University of Glasgow, CEMFI, Urban Economics Association (UEA), Workshop on International Economic Networks (WIEN), MIT, Yale University, Dartmouth College, Princeton University, Federal Reserve Bank of Philadelphia, NBER SI, Barcelona SE Summer Forum and European Research Workshop in International Trade (ERWIT), for their useful comments and suggestions. The results and conclusions of this study are ours and not those of Eurostat, the European Commission or any of the national statistical authorities whose data have been used.

Figure 1: Market shares of Catalonia in Europe



Notes: The figure shows the share of spending on Catalan goods in each European region. The shading represents the value of the market share, with darker shares representing larger market shares. The spending shares come from our newly built regional trade dataset (see Section 2).

1 Introduction

How do country borders affect trade flows within Europe? Figure 1 shows sales from Catalonia (shown in grey) to 268 European regions as a share of total spending in each destination region. A striking aspect of these market shares is their national bias. Catalonia's total share of Spanish markets, excluding Catalonia, is 5.8 percent; while its total share of non-Spanish markets is only 0.26 percent. Catalonia is not special in this regard, though. A similar national bias emerges when we examine market shares for other European regions. For the average region (whose size is about 25 percent that of Catalonia) the intranational and international market shares are 2.2 and 0.08 percent respectively.

To what extent is this bias caused by country borders?[1] Comparing intranational and

[1]We say that there is a border between two regions if they belong to different countries. Thus, we adopt

international trade could be misleading. As Figure 1 shows, Spanish regions are on average closer to Catalonia than non-Spanish regions. Since geographical distance raises transport costs and reduces trade, this creates an identification problem. A cleaner strategy would be to compare neighbouring regions. For instance, the market share of Catalonia in Languedoc-Rousillon (in France just north of Catalonia) is almost three times smaller than the market share of Catalonia in Valencia (in Spain just south of Catalonia). Is this difference caused by the French-Spanish border or the Pyrenees mountain range that coincides with it? We need to make comparisons that control for factors, such as distance and mountain ranges, that influenced the placement of borders in the past and may influence trade outcomes today.

To search for these confounding factors, normalize market shares by their average and think about them as deviations from the predictions of a naïve gravity model:[2]

$$\ln\left(\frac{n\text{'s share of market } m}{n\text{'s share of all markets}}\right) = \ln\left(n\text{'s sales to } m\right) - \ln\left(\frac{n\text{'s total sales} \times m\text{'s total spending}}{\text{spending in all markets}}\right)$$

where n and m are the origin and destination regions, respectively. The LHS is the (log) normalized market share, while the RHS is the difference between the actual (log) sales and the predicted (log) sales using a naïve gravity model. Naïve gravity applies if (i) regions produce differentiated products; (ii) regions have common homothetic preferences, and (iii) trade costs are negligible. Under these assumptions, all regions purchase the same proportions of all goods and, as a result, these proportions must be the average ones:

$$\frac{n\text{'s sales to } m}{m\text{'s total spending}} = \frac{n\text{'s total sales}}{\text{spending in all markets}}$$

Since assuming that regions produce differentiated products is uncontroversial, our search for confounding factors must focus on differences in preferences and trade costs.

There is a national bias in preferences if, for a common set of prices across regions, spending falls disproportionally on national goods, i.e. a violation of assumption (ii). One reason for such a bias is the behavior of governments. Eager for political support, governments prefer to award procurement contracts to expensive domestic suppliers instead of cheap foreign ones.[3] Another reason for a national bias in preferences is the behavior of individuals,

a purely political view of borders, i.e. having a border means not sharing a country government.

[2]To see this relationship, simply note that (i) n's share of market m equals n's sales to m divided by m's spending; and (ii) n's share of all markets equals n's total sales divided by spending in all markets.

[3]Herz and Varela-Irimia (2020) examine European public procurement contracts published in the EU's Tenders Electronic Daily database and find that local firms (located in the same region of the contracting agency) are 900 times more likely to win contracts than foreign firms. García-Santana and Santamaría (2023) use a similar dataset to explore the role of governments and find that this home bias in favor of local firms is stronger for subnational governments.

who often prefer expensive domestic goods than cheap foreign ones. Over the last couple of centuries, national governments have made massive efforts aimed at creating a common national identity. Policies such as the adoption of a single official language, the advancement of shared interpretations of history and traditions, the homogenization of educational systems and the promotion of internal migration, have all contributed to the creation of a national culture and, together with it, a preference for national goods. We treat this behavior of governments and individuals as endogenous to the border, as channels through which country borders affect trade.

There is a national cost advantage if trade costs are lower for intranational than for international trade, i.e. a violation of assumption (iii). Although tariffs have been eliminated and technical regulations have been de jure harmonized within Europe, many de facto trade barriers remain. National courts ruling on contract disputes tend to favor national firms, raising the costs of foreign firms to operate in the domestic market. National regulators tend to impede conformity assessments of foreign products to favor domestic firms. National agencies create infrastructure systems that favor intranational mobility, often at the expense of international mobility. These factors are endogenous to the border, additional channels through which country borders affect trade.

There is an important part of the national cost advantage, however, that is due to geography and cannot be attributed to country borders. The cost of transporting goods grows with distance and the presence of geographical obstacles, such as mountain ranges or seas; and it shrinks with the presence of geographical advantages, such as navigable rivers or plains. Individual spending falls disproportionally on goods with low transport costs, and these tend to be lower for intranational trade than for international trade. Interestingly, geography might also contribute to the national bias in preferences. Even if technological improvements were to eliminate transport costs, the effects of geography would still be felt as past transport costs interact with habit formation to shape present individual preferences. Since geography precedes borders and causes them (as we shall show formally later), we need an empirical strategy that effectively controls for geographical factors and produces an unbiased estimator of the causal effect of country borders on trade.

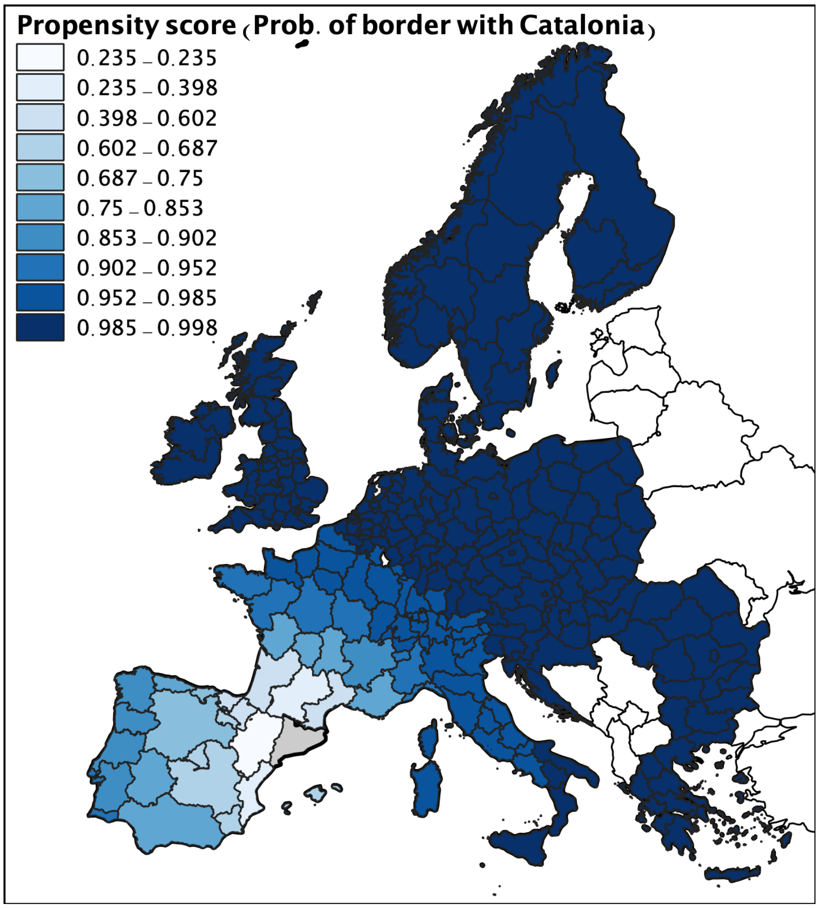
The first step in our empirical strategy is to find the appropriate dataset to work with. Measuring the border effect essentially amounts to comparing trade within and across national borders. Although there is plenty of data on trade across national borders, there is a surprising scarcity of reliable data on trade within national borders. A first contribution of this paper is to build a dataset of trade in goods for 269 regions from 24 European countries, using the European Road Freight Transport survey collected by Eurostat. This survey annually records around 3 million shipments of goods by road across Europe. For each shipment, we observe its origin and destination regions, the industry of the goods shipped, the weight of the shipment and the distance covered. We aggregate these shipments and impute export prices to build matrices of bilateral trade flows for 12 industries covering the period 2011 to 2017. This dataset provides the first integrated view of regional trade within Europe. Figure 1, for instance, was simply not known or available before.

The second step in our empirical strategy is to use the causal inference framework (see Imbens and Rubin (2015)) to design a credible identification strategy. We first estimate the probability of having a border (or propensity score) as a function of distance, insularity, remoteness and the presence of mountain ranges and river basins. These covariates explain almost half of the border assignment. Figure 2 shows the distribution of propensity scores for Catalonia (again shown in grey). Interestingly, we find regions in Spain, Portugal and France that have similar propensity scores, i.e. for which the border assignment was equally likely ex-ante even though ex-post some have a border with Catalonia and some do not.

We want an estimator that is not only unbiased, but also has a small sampling variance. Imbens and Rubin (2015) argue that there are two factors that reduce the sampling variance: (i) the number of observations (region pairs); and (ii) the balance or overlap of propensity scores between treated (region pairs separated by a border) and control (region pairs not separated by a border) groups. We first examine the entire sample and find that it is too unbalanced to produce reliable estimates. This should be apparent by looking at Figure 2. For almost all non-Spanish regions the probability of a border with Catalonia is higher than 90 percent. Thus, we trim the sample, eliminating extreme observations with propensity scores close to zero or one, to achieve a much better overlap of propensity score distributions between treated and control pairs. We then use the trimmed sample to construct a blocking estimator. That is, we build subsamples or blocks of region pairs with similar propensity scores, we estimate the border effect within these blocks and we weigh the block estimates to produce an average border effect. Since the probability of having a border is similar between treated and control pairs within each block, the difference in trade between them can be interpreted as the causal effect of the border.

Take two similar region pairs, the first one containing regions in different countries and the second one containing regions in the same country. The main result of this paper is that the market share of the origin region in the destination region for the international pair is only 17.5 percent that of the intranational pair. We refer to this estimate as the average border effect, and we say that country borders cause reductions in market shares of 0.175. This estimate is quite precise and remarkably similar across blocks, i.e. at different levels of

Figure 2: Probability of having a border with Catalonia



Notes: The figure shows the probability of finding a border between Catalonia and each European region based on a set of geographical covariates (propensity score). The shading represents the value of this probability, with darker shares representing probabilities closer to one.

the propensity score. Thus, the specific weighting scheme chosen for the blocking estimator has little effect on the final estimate. We do find some variation, though, when we estimate the border effect for each industry separately. In particular, we find that borders cause reductions in market shares that range from 0.123 to 0.389.

Is our estimate of the border effect large? The answer to this question naturally depends on one's own priors. But we can gain some intuition by being more specific about the counterfactual. After the War of Spanish Succession (1701-1714), the first Bourbon king of Spain Philip V incorporated Catalonia as a province of the kingdom of Spain. What would have happened if, instead, it would have been the French Bourbon king Louis XIV who incorporated Catalonia as a province of the kingdom of France? It is not too far-fetched to think that this would have made Catalonia quite different from what it is today. French would co-exist with Catalan and Spanish would be considered a foreign language, Catalans would exhibit a taste for French goods and traditions rather than Spanish ones, transport systems would foster mobility north rather than south, many Catalans would have their origins and family ties in other French regions rather than in Spanish ones, and so on. Is it surprising to find that, in this scenario, Catalonia would be trading 5.714 times more with other French regions and 0.175 times less with Spanish regions today?

An important observation is that our estimate should be treated as an "average" border effect. One potential source of heterogeneity is the age of the border. It takes a long time to build a common national identity, or an infrastructure system aimed at promoting internal interactions. It takes less time to implement a procurement system that favors domestic firms or to enact laws and regulations that protect them from foreign competition. Thus, borders with different ages might have different effects. To determine this, we exploit the process of political fragmentation that Europe has experienced since 1910. In our sample, about one third of the region pairs that shared a government in 1910 no longer share a government in 2010. Using the methodology explained above, we find that post-1910 borders reduce market shares to 28.3 percent of their potential. This estimate is still large, but substantially smaller than our estimate of 17.5 percent obtained by pooling pre- and post-1910 borders.

So what is behind this large border effect? A detailed analysis of the mechanisms through which borders affect trade is outside the scope of this paper. But we can use our methodology to estimate the causal effect of borders on bilateral variables that we know that are correlated with bilateral trade and we think that they might cause trade. For instance, we estimate that a border raises by 71.4 percent the probability of having different languages. We also estimate that a border causes a 6.2 percent increase in distance traveled, presumably through the choice of infrastructure. We also find that borders increase by 30 percent the disagreement in values between regions (relative to the average *Disagreement Index*), presumably through the efforts of national governments to promote internal migration and create a common culture. The new evidence reported here supports the view that these factors might be behind the large border effect.

The paper is organized as follows. Section 2 describes how we construct the dataset. Section 3 explains our identification strategy. Section 4 presents our results. Section 5 examines the effects of borders on sharing a language, distance traveled and values disagreement. Section 6 concludes. Before all of this, we review previous efforts to estimate the border effect.

Literature review: In his pioneering study, McCallum (1995) estimated a gravity equation (that is, a linear regression of bilateral trade on economic size and distance) extended to include a border dummy. The estimated coefficient indicated that, after controlling for economic size and distance, trade between Canadian provinces was on average 22 times larger than trade between Canadian provinces and US states. Although the notion that borders hinder trade was not surprising, the magnitude of the effect came as a shock, as model-based explanations using conventional trade barriers seemed unable to account for the size of the border coefficient.

A first reaction to McCallum's result was mostly methodological, and it centered on how to estimate gravity equations that are consistent with the theory. In an influential paper, Anderson and Van Wincoop (2003) showed that controlling for differences in price levels, something that McCallum (1995) had not done, reduced McCallum's estimate from 22 to 5. This estimate implies that the US-Canada border reduces market shares to 20 percent of their potential. Feenstra (2002) proposed a much simpler fixed-effects strategy that soon became the standard to estimate gravity equations. This did not affect, though, the finding that controlling for price levels reduces McCallum's estimate from 22 to 5. The methodology to estimate gravity equations evolved rapidly over the next few years.[4] But this has not led to a revision of the effect of the US-Canada border.

The first contribution of our paper is to use the causal inference framework to estimate the border effect. The gravity equation has proven to be a very useful theoretical and empirical tool in many settings. But the coefficient of a border dummy in a gravity equation cannot be interpreted as causal. Borders reduce the spending on goods produced by a region, lowering its income. And yet gravity equations include incomes as independent variables (either directly or, more often, through the use of fixed effects) alongside the border dummy. This creates a classic "bad-control" problem when we try to interpret the coefficient of the border dummy as causal. A similar problem applies to bilateral variables that are typically included in gravity equations, such as dummies indicating a common language or a common currency. Dealing with this problem turns out to be important quantitatively. We show that using a fixed-effects gravity regression overestimates the border effect. In particular, this procedure suggests that borders reduce market shares to 10.7 percent of their potential, instead of the

[4] The use of log-linear OLS came under scrutiny due to concerns regarding its performance in the presence of heteroskedasticity (Silva and Tenreyro, 2006) and its inability to incorporate zero trade flows (Helpman et al., 2008). As a consequence, more flexible estimation methods such as Poisson-Pseudo Maximum Likelihood and Gamma-Pseudo Maximum Likelihood became customary. Baier and Bergstrand (2009) proposed an approximation to control for multilateral resistance terms that allowed the researcher to perform comparative statics while still using OLS for estimation. Head and Mayer (2014) provide a review of these developments.

17.5 percent that we find.

A second reaction to McCallum's result was to go beyond the US-Canada border and look at the effects of other borders. A major obstacle, though, was the absence of readily available datasets on regional trade for other country pairs. Wei (1996) and Nitsch (2000) computed intranational trade as national production minus exports and compared it to international trade for OECD and European countries, respectively. Later studies measured intranational trade using data at the region-region level and international trade using data at the region-country level (See, for instance, Gil-Pareja et al. (2005) and Coughlin and Novy (2021)). This was indeed an improvement, although comparisons between different units are still far from ideal.[5] Gallego and Llano (2015) use region-region level data, but they focus exclusively on Spanish regions. Not surprisingly, given the large variation in methodologies and samples used, these papers have estimated a border effect that varies between reducing market shares to 30 percent of their potential (Coughlin and Novy (2021)) up to reducing market shares to 7 percent of their potential (Gallego and Llano (2015)).

The second contribution of our paper is to estimate the border effect using the most comprehensive region-to-region dataset to date, which includes 269 regions and 24 countries. For instance, studies that focus on North America typically have 60 regions and 2 countries. The dataset that we construct here constitutes a major leap forward in the analysis of the border effect.

There are some studies that have explored the specific channels through which borders affect trade. For instance, Turrini and van Ypersele (2010) explore the effects of judicial systems, Bailey et al. (2020), Combes et al. (2005), Fukao and Okubo (2004) and Rauch (2001) explore the role of social and business networks, Schulze and Wolf (2009) focus on ethno-linguistic factors, Evans (2003) studies the relative importance of substitutability between goods vis-a-vis price wedges and Chen (2004) analyzes technical barriers to trade and product-specific information costs that increase the effect of borders on trade.

For these variables to be relevant, one must show the causal effect of borders on them. The third contribution of this paper is to show that borders do indeed cause linguistic differences, increase distance traveled and create values disagreement.

[5] The problem is aggravated because working with the wrong units also makes it difficult to measure distance. Head and Mayer (2009) showed that accurate measurement of distance is critical to having a precise estimate of the border coefficient. Moreover, Hillberry and Hummels (2008) and Coughlin and Novy (2021) have shown that using large geographical units overlooks the non-linear effect of distance on trade, generating an upward bias on the border coefficient.

2 European regional trade: a new dataset

The European Road Freight Transport survey (ERFT) is a micro-level survey of freight road shipments collected by the statistical office of the European Union, Eurostat. The ERFT data is collected from a survey of shippers in the industry, and is therefore similar in nature to the Community Flow Survey data available for the United States that has been used in a number of empirical studies. This section describes the main features of the ERFT survey and shows how we use it to build our dataset.

A natural question is whether freight road shipments are representative of all trade flows. According to Eurostat's own statistics, between 2011 and 2017 road freight accounted for about 49 percent of all intra-EU trade in tonne-km terms, while the share of maritime short-sea shipping and rail transport were 32 percent and 11 percent respectively (the other modes of transportation reported are inland waterways 4, pipelines 3 and air 0.1). Thus, we think that our dataset measures a sizeable fraction of intra-European trade.

2.1 From road shipments to regional trade weights

The ERFT survey covers shipments by road aggregated every year from micro-data collected by a total of 29 European countries, all European Union members except for Malta plus Norway and Switzerland.[6] Each participating country chooses a stratified sample of vehicles from the national register of road freight vehicles, following Eurostat guidelines.[7] The operators of the sampled vehicle are required to report, for a limited number of days in a month, the characteristics of all the shipments completed.

The survey requests information at the level of the vehicle, the journey and the specific goods shipped. At the level of the vehicle, the survey records vehicle characteristics such as age, type of vehicle and ownership. At the journey level, the questionnaire records whether the journey is loaded or unloaded, the type of transport (hired or own account) and the type

[6] The European Union adopted in 1998 regulation to provide a legal base for the collection of a wide range of data on road freight transport ((EC) 1172/98), laying the emphasis on quality and comparability of statistical information. This regulation has introduced major changes in the data collected in order to describe the regional origin and destination of intra-European Union transport on the same basis as national transportation (Road Freight Transport methodology, 2016 edition).

[7] The selection of the sample is made to ensure that the raw survey results are representative of the total numbers recorded on the vehicle register. In countries where such a registry is not available or sufficiently reliable, a register of persons licensed to operate as road hauliers (company/registered owner for private hauliers) or a business register of companies could be considered. In this case, the sampling unit could be the vehicle operators or transport companies. (Road Freight Transport methodology, 2016 edition) Further details are provided in the ERFT survey documentation.

of journey.[8] At the goods level, the record includes the shipment's weight (kg), the type of goods carried according to the 2 digit NST 2007 classification, the region of origin and destination (at NUTS3 level), the actual shipping distance covered and a sampling weight for each shipment.[9] Eurostat aggregates the origin and destination of each shipment into larger regions (at NUTS2 level) for anonymity reasons. We use the ERFT survey for the period 2011 to 2017. Using this micro-dataset has several advantages relative to using aggregate trade data. It also requires us to make some adjustments.

A first advantage of the survey is that it allows us to overcome one of the main challenges to estimate the border effect: the lack of subnational trade data. The ERFT survey allows us to distinguish between flows within a region and flows between regions in the same country for all countries surveyed except for five one-region countries: Cyprus, Estonia, Latvia, Lithuania and Luxembourg. For this reason, we drop these countries from the dataset. This leaves us with 24 countries in our sample: the remaining 22 European Union countries plus Norway and Switzerland.

A second advantage of the survey is that it is collected from a stratified sample of actual shippers rather than imputed from different aggregated data sources. This means that our data captures, with higher accuracy, the movement of goods within countries. The survey includes two types of flows: shipments that move goods between producers and consumers and shipments that move goods from a producer to an intermediary or from intermediary to intermediary. What the survey actually captures is the region to region distribution of goods. In most cases, these shipments will take goods from the origin to the destination region. Yet, in other cases, these shipments will be a middle step in a longer distribution chain across European regions, not coinciding with the observed origin and destination of the trade flow.

To address this limitation, we restrict our sample in three ways. First, we use the detailed information in the survey to drop journeys that are classified as distribution journeys. These journeys are characterised by the existence of several stops between the origin and the destination to load and/or unload goods. Dropping these journeys seeks to bring our shipment data closer to trade data.

Second, we restrict the number of industries in the analysis. The shipments are classified

[8] The type of journey records whether the journey involved one single transport operation, several transport operations or a collection/distribution of goods, with many stopping points for loading and/or unloading in the course of a single journey.

[9] The weight of shipments is calculated by multiplying reported estimates by the inverse of the sampling weight. The industry classification followed in the survey is the NST 2007 classification, the "statistical classification of economic activities in the European Community".

into 20 industries enumerated in Table B.1 in the Appendix. We adopt two criteria for industry coverage: (i) the industry must be unambiguously associated with trade; and (ii) transport by road must be an important mode of transport for the industry. The first criterion leads us to discard seven industries.[10] The second criterion leads us to discard one additional industry.[11] Thus, we are left with twelve industries.

Finally, we want to make sure that the survey on road shipments is representative of aggregate trade. This would not be the case for regions with a very small share of shipments traveling by road. To ensure this, we restrict the number of regions by dropping insular regions very far from continental Europe. For these small and far away regions, shipments by road are not likely to be representative.[12]

After all these adjustments, our dataset contains 269 regions (in 24 countries) and 12 industries. We use the dataset to construct a set of industry-year matrices:

$$W^{it} = \left[W^{it}_{nm}\right]_{269\times 269}$$

where W^{it}_{nm} is the weight (kg) from industry i shipped from region n to region m in year t. Since our dataset contains 12 industries and 7 years, we have 84 such matrices.

Figure 3 plots exports (kg) across the countries in our sample in the Y-axis against bilateral shipments (kg) obtained by aggregating the survey data at the country level on the X-axis. As we can see, most observations concentrate along the 45 degree line (Rsq=0.55), showing that our data is very correlated with aggregate exports data from Eurostat. Figures A.1, A.2 and A.3 in the appendix plot the same relationship, year-by-year and industry-by-industry. These figures show that this correlation is also strong when we use data disaggregated by industry and/or year.[13]

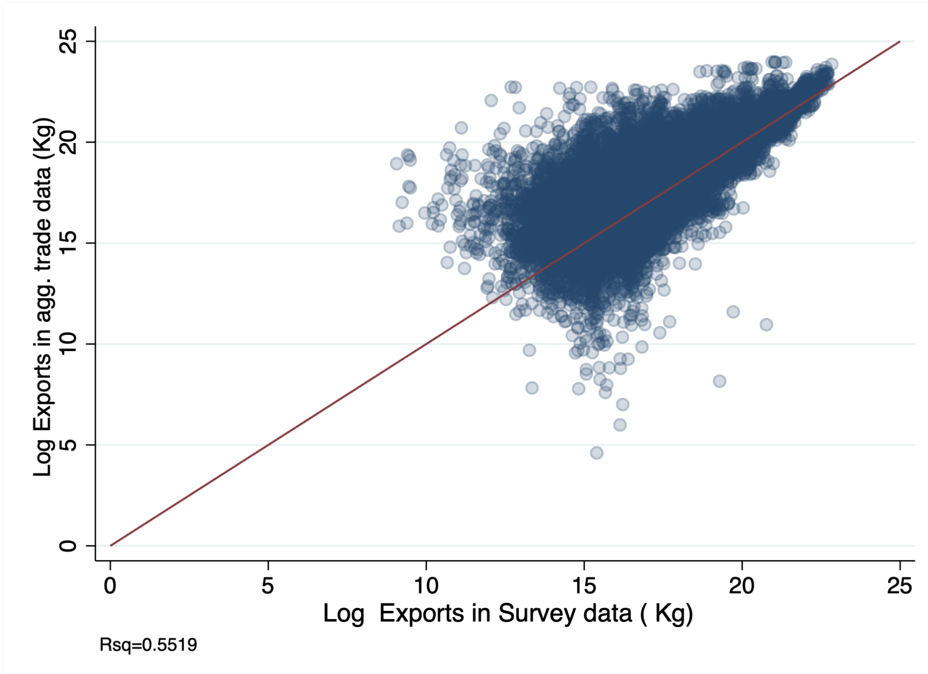
[10] These industries are: 14 Secondary materials, municipal wastes and other wastes; 15 Mail, parcels; 16 Equipment and materials utilized in the transport of goods; 17 Goods moved in the course of household and office removals, 18 Grouped goods; 19 Unidentifiable goods; and 20 Other goods n.e.c. It is unclear to us what fraction of the shipments included in these categories can be safely classified as trade in goods. For instance, disposing of waste, distributing mail or moving furniture is clearly not associated with trade.

[11] This industry is: 2 Coal and lignite, crude petroleum and natural gas. A large fraction of trade in this industry is transported by railways or through pipelines.

[12] We keep large, close-by islands like Sardinia or Sicily. The survey includes shipments taken by truck when the truck is loaded on a ship and unloaded after crossing to an island. Therefore, we can include these larger islands since their trade is well represented in the survey. A table with the list of all regions can be provided upon request.

[13] The R-square for these correlations stays above 50% for all years and all industries except for industry 3:"Metal ores, mining and other quarrying products, Uranium" (Rsq=0.42); industry 7: "Coke and refined petroleum products" (Rsq=0.24) and industry 13: "Miscellaneous: Furniture and other manufacturing products" (Rsq=0.47), probably due to the particular nature of these industries.

Figure 3: Correlation with aggregate international trade data



Notes: The figure shows the correlation between exports and shipments in the ERFT survey in kilograms. The Y-axis represents (log) bilateral trade (kg) by country-pair-industry-year using international trade data from Eurostat. The X-axis shows bilateral shipments by road (kg) aggregated by country-pair-industry-year obtained from the ERFT survey.

2.2 From trade weights to trade values

The survey provides trade weights, and we would like to convert weights into values. Thus, we look for other data sources. The statistical agencies of France, Germany, Spain and United Kingdom release data of exports from individual regions to foreign countries in value and volume. These data allows us to observe export flows from 66 regions in our sample (belonging to the four countries mentioned above) to all the remaining countries in our sample. For these export flows, we observe the value in euros and the quantity in kilograms of export flows, allowing us to compute the price per kilo of exports. Unfortunately, similar data could not be collected for the remaining countries in our sample. The reason why such regional level data on exports is not available for other countries is unknown to us and, hopefully, not systematically related to the price of exports in those regions. Therefore, we think of our data as incomplete data in which the price of exports is missing for part of the sample.

Imputation methods replace missing values by suitable estimates and then apply standard methods to the filled-in data. Imputations are means or draws from a predictive distribution of the missing values, and require a method for creating a predictive distribution for the imputation that is based on the observed data. We choose an explicit modelling approach, where the distribution is based on a formal statistical model. In particular, we use regression imputation, a standard choice of conditional mean imputation. First, the regression of the variable with missing values on other covariates is estimated from the complete cases, and then, the resulting prediction equation is used to impute the conditional mean of the missing values. Regression imputation is a plausible method, particularly when the chosen covariates explain most of the variation of the variable with missing values.

Our preferred specification is to pool all time periods and industries to estimate a linear regression for the (log) of the price of exports, calculated as the ratio between the value of exports and the weight of exports for each industry, origin, destination and year. As explanatory variables, we use a vector of origin and destination characteristics. The only bilateral variable that we use is distance.[14] We also include industry-time dummies to allow for different time trends in prices across industries. Table C.2 in the appendix contains the full list of variables included in the price regressions.

Our regression model seems to perform well, as shown in Table B.2 in the Appendix. The R-squared in the above specifications is higher than 50 percent. Since the collected variables explain a large share of the variation in export prices in the subsample with no missing values, we can use the estimated coefficients from the linear regression to impute the values that are missing.[15]

With our estimated prices per unit, we can finally construct the trade value data for each industry i and year t as follows:

$$V^{it} = \left[V^{it}_{nm}\right]_{269\times 269} \qquad \text{where } V^{it}_{nm} = P^{it}_{nm} \cdot W^{it}_{nm}$$

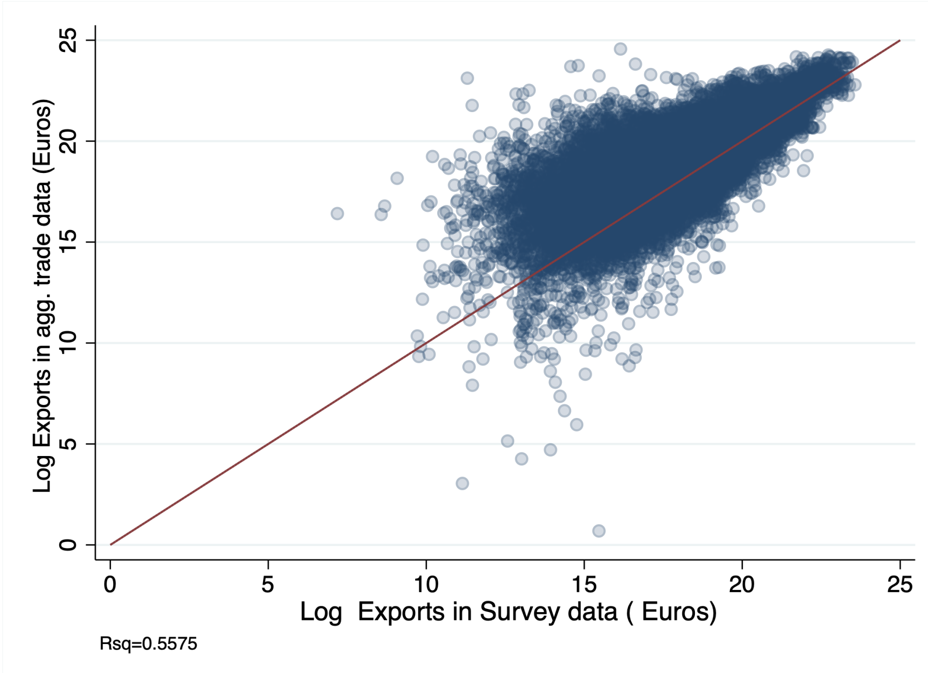
where V^{it}_{nm} is the value (euros) from industry i shipped from region n to region m in year t.

Figure 4 plots exports (euros) across the countries in our sample in the Y-axis against

[14] As shown in Hummels and Skiba (2004), the presence of transport costs leads firms to ship high-quality goods abroad while keeping low-quality goods for the domestic market. This is known as the "Alchian and Allen conjecture" (see Alchian and Allen (1964)). Another reason why export prices per kilogram could increase with distance is transport costs. However, our export prices are Free On Board (F.O.B), meaning that they are net from transport and insurance costs.

[15] In order to further assess the accuracy of our imputed prices we perform two sets of checks. First, we perform a series of out-of-sample estimations where we drop one of the four countries for which we observe regional export prices and we predict export prices for this dropped country. We then compare our out-of-sample estimates with the actual regional prices (See Figure A.4 in the Appendix). Second, we collect export value and weights from Eurostat for all European countries and compute unit export prices for every country-pair at the industry and year level. We aggregate our region-pair estimated prices to a country-pair level and compare them to the country-pair price of exports from international trade data (See Figure A.5 in the Appendix). Both tests suggest that our imputed prices are reasonable.

Figure 4: Correlation with aggregate international trade data



Notes: The figure shows the correlation between exports and shipments in the ERFT survey in euros. The Y-axis represents (log) bilateral trade (euros) by country-pair-industry-year using international trade data from Eurostat. The X-axis shows bilateral shipments by road (euros) aggregated by country-pair-industry-year obtained from the ERFT survey after imputing missing prices.

bilateral shipments (euros) obtained by aggregating the survey data at the country level on the X-axis. As we can see, most observations concentrate along the 45 degree line (R-squared = 0.55), showing that our data is very correlated with aggregate exports data that come from Eurostat when we use values. Figures A.6, A.7 and A.8 plot the same relationship, industry-by-industry and year-by-year. These figures show that this correlation is also strong when we use data disaggregated by industry and/or year.

2.3 European Regional trade: A first look at the data

Our dataset contains region pairs such that: (i) origin and destination regions belong to the same country; and (ii) origin and destination regions belong to different countries. We refer to these two types of trade as intranational and international, respectively.[16] Out of a total of 72,092 region pairs in our sample, 4,958 are intranational, and 67,134 are international.[17]

[16]We exclude from our sample pairs for which the origin region is the same as the destination region. Therefore, intranational trade does not include trade within a region.

[17]These numbers take into account origin and destination. Thus, we count region pair (n, m) as different than (m, n).

Table 1: Summary statistics

Trade type	Intranational trade Mean	International trade Mean
Panel A: Unconditional		
Value (Mill. euros)	553.52	18.61
Weight (Mill. Kg)	601.49	9.98
Normalized Market share	10.87	0.27
Panel B: Zero trade observations		
Region pairs	4958	67134
Region pairs with no trade	157	25699
Regions pairs with positive trade	4801	41435
Panel C: Conditional on positive trade		
Value (Mill. euros)	571.62	30.15
Weight (Mill. Kg)	621.15	16.17
Normalized Market share	11.22	0.44

Notes: This table reports the (unweighted) average bilateral trade flow (euros and kilos) and the (unweighted) average normalised market share in our new European regional dataset. Column 1 reports the average flow between intranational region pairs (origin and destination in regions in the same country) and column 2 reports the average flow between international region pairs (origin and destination regions in different countries). Panel A reports unconditional statistics. Panel B repors the number of region pairs that display positive trade and zero trade. Panel C reports statistics conditional on trading.

Panel A of Table 1 shows the average values of the two types of trade at the region-pair and annual level. We see that the average value of trade among intranational pairs is almost 30 times larger than among international pairs. This average is unweighted, and one might think that it could be affected by differences in economic size between groups. We obtain a similar picture, however, when we look at normalized market shares.

Panel B of Table 1 shows another important feature of our data, the prevalence of region pairs that do not trade. Among intranational pairs, 96.8 percent exhibit positive trade. The picture is quite different when we look at international pairs. Among them, only 61.7 percent of pairs trade with each other. Taking this into account, Panel C of Table 1 shows the same statistics as in Panel A but now conditional on observing a positive flow of goods. Not surprisingly, this increases the average trade values among international pairs, without affecting much the average trade values of the other group. The main takeaway is that the national bias manifests itself both on the intensive and the extensive margins.

3 Identifying the border effect

The causal relationship of interest is the effect of country borders on trade. In this section, we describe our empirical strategy to identify this effect which draws heavily from the causal inference framework (see Imbens and Rubin (2015)). We use as an outcome variable, the normalized market share:

$$S_{nm} \equiv \frac{V_{nm}/E_m}{Y_n/E} \tag{1}$$

where $Y_n = \sum_m V_{nm}$ are the total sales or income of region n; $E_m = \sum_n V_{nm}$ are the total purchases or spending of region m, and $E = \sum_m E_m$ is total spending by all regions. The variable S_{nm} measures region n's share of region m's market normalized by region n's share of all markets, including its own. If market m has an average importance to producers of region n, i.e. $V_{nm}/E_m \approx Y_n/E$; the market share is one. If instead market m has a larger (smaller) than average importance, the market share is above (below) one. Unlike trade values, normalized market shares are not affected mechanically by the economic size of origin and destination regions.[18] This makes them more helpful than trade values to infer preference biases and trade costs.

3.1 The border effect

The French-Spanish border runs across Catalonia and Languedoc-Roussillon, and not across Catalonia and Valencia. Catalonia's average market share in all the 269 regions in our sample is 1.5 percent. Given how close Catalonia is geographically and culturally to Languedoc-Roussillon and Valencia, it is not surprising that these two markets be specially important for Catalan exporters. Indeed, the normalized share of Catalonia in the Languedoc-Roussillon market is well above one, 1.79, implying that $1.79 \times 1.5 = 2.7$ percent of all the spending of Languedoc-Roussillon is on products that come from Catalonia. Yet Catalonia's normalized share of the Valencia market is almost three times larger than this, 5.21, implying that $5.21 \times 1.5 = 7.9$ percent of all the spending of Valencia is on products that come from Catalonia. To what extent is this difference caused by the French-Spanish border? What would have happened if this border were southwest of Catalonia instead of north? How much would Catalonia's share of the Languedoc-Roussillon market grow? How much would Catalonia's share of the Valencia market shrink?

Answering these questions involves comparing observed market shares with the counter-

[18] To see this, assume trade is balanced, i.e. $E_m = Y_m$ and E = Y. Then, we have that:

$$lnS_{nm} = lnV_{nm} - lnY_n - lnY_m + lnY$$

Since $Y_n = \sum_m V_{nm}$, one might think that $\ln S_{nm}$ is obtained by taking out fixed-effects from $\ln V_{nm}$. This is close, but not quite right. To construct $\ln S_{nm}$, we subtract and add the logs of the means to $\ln V_{nm}$, and not the means of the logs.

factual market shares that would have occurred if the French-Spanish border were southwest of Catalonia. More formally, let (n, m) be a region pair, and let $B_{nm} \in \{0, 1\}$ be a dummy variable that takes value one if the regions in the pair belong to different countries, and zero otherwise. Let S_{nm} be the observed market share for region pair (n, m) in our sample. We define two potential market shares as follows:

$$S_{nm} = \begin{cases} S_{nm}(1) & \text{if } B_{nm} = 1 \\ S_{nm}(0) & \text{if } B_{nm} = 0 \end{cases} \tag{2}$$

where $S_{nm}(1)$ and $S_{nm}(0)$ are region n's share of market m with a border (active treatment) and without a border (control treatment), respectively. For each region pair, we observe only one potential outcome. For instance, we observe $S_{CAT,L-R}(1) = 1.79$ for the pair (Catalonia, Languedoc-Roussillon) and $S_{CAT,VAL}(0) = 5.21$ for the pair (Catalonia, Valencia). Unfortunately, we do not observe $S_{CAT,L-R}(0)$ or $S_{CAT,VAL}(1)$.

We define the border effect β_{nm} as the log change in market shares caused by the border:

$$\beta_{nm} = \ln \frac{S_{nm}(1)}{S_{nm}(0)} \tag{3}$$

Since one potential outcome is unobserved, we cannot observe border effects. It is tempting however to assume that, if the French-Spanish border were southwest of Catalonia, the roles of these two markets for Catalan exporters would reverse, that is, $S_{CAT,L-R}(0) = S_{CAT,VAL}(0)$ and $S_{CAT,VAL}(1) = S_{CAT,L-R}(1)$. This identification assumption allows us to estimate a common border effect for the two region pairs as follows:

$$\beta = \ln \frac{S_{CAT,L-R}(1)}{S_{CAT,VAL}(0)} = -1.07 \tag{4}$$

That is, the French-Spanish border reduces Catalonia's share of the Languedoc-Roussillon market to a third of its potential: $100 \times e^{-1.07} = 34.3$ percent. Should we take this estimate very seriously? How good is the identification assumption that underlies it? The main challenge we face in this paper is to construct samples for which this type of comparisons can be interpreted as causal.

There are a couple of assumptions embedded in our notation worth mentioning explicitly. The first one is that the unobserved potential outcome is unique. As mentioned, moving Catalonia to France would remove the border between Catalonia and Languedoc-Roussillon. But so would moving Languedoc-Roussillon to Spain, or creating a new country containing both regions. Our framework implies that $S_{CAT,L-R}(0)$ is the same in all these cases and, indeed, in any other possible case. This assumption captures the view that, to a first-order approximation, what matters is whether there is a border or not. The specific type of border only matters to a second or third-order approximation. We think this is quite a reasonable view.

Our notation also embeds the notion that the difference in potential outcomes measures the effect of changing the border for one region pair, *keeping all other borders constant*. This partial-equilibrium clause, which is standard in micro studies that use the causal framework, has an added force in this context. It still contains the standard requirement that region pairs be small so that "treating" one of them does not have general equilibrium effects on European trade. But this is not enough in this context. The units of observation are region pairs, but borders are not bilateral variables. It is not possible in general to "treat" one region pair only, leaving all other pairs "untreated". Consider again moving the French-Spanish border southwest of Catalonia. This experiment would remove the border between Catalonia and 22 French regions and create a border between Catalonia and 15 Spanish regions. Thus, it would produce 37 border changes. Since these border changes affect only 0.001% of all region pairs, it seems safe to assume they would have a minor impact on European trade and the partial-equilibrium assumption holds.[19]

Since we cannot experiment with borders, we must rely on observational data to estimate an average border effect. In particular, we define the average border effect β as the average log change in market shares caused by the border as:

$$\beta = E\left(\ln S_{nm}(1) - \ln S_{nm}(0) \middle| S_{nm}(1) > 0,\ B_{nm} = 1\right) \tag{5}$$

The value of β is expected to be negative since the border is expected to reduce trade. The larger is $|\beta|$, the larger is the average reduction in market shares caused by the border. Throughout, we assume that there are no region pairs such that $S_{nm}(1) > 0$ and $S_{nm}(0) = 0$. Obviously, this cannot be verified.

The causal inference framework shows that we can use observational data as if it came from an experiment if the assignment of treatment is (i) probabilistic, (ii) individualistic and (iii) uncounfounded. If the assignment mechanism satisfies these conditions, the comparison of units with different treatments but identical pre-treatment covariates can be given a causal

[19] This partial equilibrium clause also applies to the standard fixed-effects gravity regressions. The coefficient of a border dummy in these regressions is often interpreted as the effect of removing a border on the trade of a 'small' region pair, *keeping all other borders constant*. As our estimate, the gravity-based estimate is of a partial-equilibrium nature since removing many borders at once would also change incomes and price levels around the entire system, and this would also affect trade beyond the effect measured by the coefficient.

interpretation.

We believe that the first two conditions hold in our setting. Probabilistic assignment requires a nonzero probability for each treatment value, for every unit. The probability that two far-away regions belong to the same country might be very small, but it is not zero. Individualistic assignment requires limited dependence of a particular unit's assignment probability on the values of covariates and potential outcomes for other units. This is the partial-equilibrium clause mentioned above, which we argued is a reasonable one.

The last condition, unconfounded assignment, deserves much more attention. Under unconfoundedness, all the assignment probabilities are free from dependence on potential outcomes, after conditioning on a vector of pre-treatment covariates. This assumption is often referred to as the Conditional Independence Assumption (see Dawid (1979)) and written as $B_{nm} \perp S_{nm}(0), S_{nm}(1) | X_{nm}$. In our setting, unconfoundedness means that the assignment of borders must be independent of potential trade outcomes across regions, after conditioning on a vector of pre-treatment geographical covariates X_{nm}. We describe this vector and explain our control strategy in the next couple of sections.

Let us assume for now that we have a vector of pre-treatement geographical covariates X_{nm} such that, after conditioning for them, the border assignment is unconfounded. This allows us to interpret comparisons between units with different treatments as causal. Does this mean that we can estimate the border effect by simply comparing the average market shares of international and intranational pairs with the same covariate values $X_{nm} = x$? The answer, unfortunately, is negative. The following estimator makes exactly this comparison:

$$\begin{aligned} \hat{\beta} \;=\; & E\left(\ln S_{nm}(1) \middle| \mathcal{S}_{nm}(1) > 0, B_{nm} = 1, X_{nm} = x\right) \\ & - E\left(\ln S_{nm}(0) \middle| \mathcal{S}_{nm}(0) > 0, B_{nm} = 0, X_{nm} = x\right) \end{aligned} \tag{6}$$

It is straightforward to see that $\hat{\beta}$ suffers from two potential sources of selection bias:

$$\begin{aligned} \hat{\beta} - \beta = & \underbrace{\begin{array}{c} E\left(\ln S_{nm}(0) \middle| \mathcal{S}_{nm}(1) > 0, B_{nm} = 1, X_{nm} = x\right) \\ - E\left(\ln S_{nm}(0) \middle| \mathcal{S}_{nm}(1) > 0, B_{nm} = 0, X_{nm} = x\right) \end{array}}_{\text{Selection bias due to the number of borders}} \\ + & \underbrace{\begin{array}{c} E\left(\ln S_{nm}(0) \middle| \mathcal{S}_{nm}(1) > 0, B_{nm} = 0, X_{nm} = x\right) \\ - E\left(\ln S_{nm}(0) \middle| \mathcal{S}_{nm}(0) > 0, B_{nm} = 0, X_{nm} = x\right) \end{array}}_{\text{Selection bias due to changes in participation}} \end{aligned} \tag{7}$$

Consider first the selection bias due to the number of borders, which is the first term of Equation (7). It might seem surprising that we condition on the border after assuming that the border assignment is unconfounded. But there is a subtle source of selection bias that arises from any random border assignment, including those that are unconfounded. To understand its nature, consider a world with 6 regions and 2 countries. The six regions are identical in any possible way, except for the border assignment. The latter is random, with all regions being equally likely to belong to any country. Let us assume that the realization of the border assignment is such that regions 1 and 2 belong to country A, while regions 3, 4, 5 and 6 belong to country B. This introduces the only source of asymmetry in this world: regions in A have four borders, while regions in B have only two borders. Assume there are no trade costs other than those caused by the border, which result in the same percentage reduction in market shares for all pairs:

$$\beta = \ln \frac{S_{nm}(1)}{S_{nm}(0)} \quad \text{for all } n, m \tag{8}$$

Let S_A^D and S_B^D be the market share of any region in A and B in a domestic market (including itself), respectively. Symmetry and the absence of non-border related trade costs ensure that, within each country, these shares are identical for all relevant pairs. Let S_A^F and S_B^F to be the market share of any region in A and B in a foreign market, respectively. Symmetry and the absence of non-border related trade costs also ensure that, within each country, these shares are identical for all relevant pairs. By construction, normalized market shares must add to one. Thus, we have that

$$2S_A^D(0) + 4S_A^F(1) = 4S_B^D(0) + 2S_B^F(1) = 1 \tag{9}$$

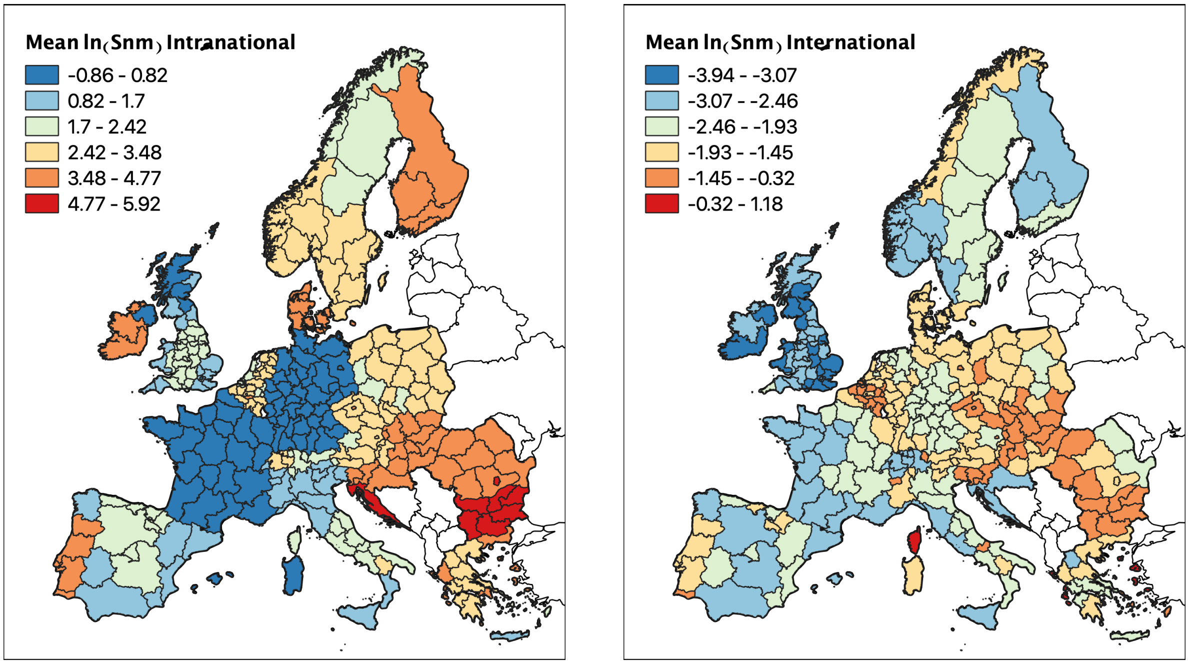
It is straightforward to show that Equations (8) and (9) imply that:

$$\frac{S_A^D(0)}{S_B^D(0)} = \frac{S_A^F(1)}{S_B^F(1)} = \frac{2+e^{\beta}}{2e^{\beta}+1} > 1 \tag{10}$$

for any value of $\beta < 1$. That is, regions with many borders have larger market shares. The key observation is that region pairs with many borders tend to be over-represented among international pairs and under-represented among intranational pairs. This creates a positive selection bias that makes the observed difference in average market shares smaller (in absolute value) than the true average border effect.[20]

[20]The existence of this type of selection bias was noted first by Anderson and Van Wincoop (2003). In their sample, however, the group of intranational pairs contained only Canadian provinces, i.e. regions with many borders; while the group of international region pairs contained mostly US states, i.e. regions with

Figure 5: Average market share and number of borders



A) Intranational market share **B) International market share**

Notes: The figure shows the average market share of each region with its intranational partners (panel A) and with its international partners (panel B). The color shading represents the value of this average, with cooler colours representing lower market shares and warmer colors representing higher market shares

Fortunately, there is a simple solution to this problem, namely, to estimate border effects conditioning on the number of borders. We shall show later that this type of selection bias is important empirically. But one can already suspect this by looking at Figure 5, which shows average intranational and international market shares in panels A and B, respectively. The color of a region represents the value of the average normalized share, with dark blue shades representing the smallest values and dark red shades representing the highest values. In countries with many regions, such as United Kingdom or Germany, regions have smaller than average intranational and international market shares (predominantly blue shades). In countries with few regions, such as Belgium, Slovenia, or Portugal, regions have larger than average intranational and international market shares (predominantly red shades).

Consider next the selection bias due to changes in participation, which is the second term of Equation (7). This type of selection bias arises because some region pairs trade without a border, $S_{nm}(0) > 0$, but would not trade with a border, $S_{nm}(1) = 0$. Let us refer to these pairs as switchers. Average market shares for intranational pairs include switchers, while

few borders. Thus, they found that this type of selection bias leads to overstating the average border effect. Here, with a balanced sample, this selection bias leads to understating the border effect.

average market shares for international pairs do not. If average market shares for switchers and non-switchers were the same, there would be no selection bias and the second term in Equation (7) would be zero. But it is reasonable to expect average market shares for switchers to be lower than those of pairs that always trade. This creates a positive selection bias that makes the observed difference in average market shares smaller (in absolute value) than the true average border effect.

The importance of this bias depends on the fraction of switchers in the sample. Without this information, we must treat $\hat{\beta}$ as a lower bound for the border effect. We show later, however, that the fraction of switchers must be quite small in the samples we work with. This means that the bias due to changes in participation cannot be important quantitatively and, as a result, $\hat{\beta}$ provides a good estimate for the border effect.

To sum up, if the border assignment is probabilistic, individualistic and unconfounded, we can compare intranational and international pairs and be confident to obtain a good estimate of the border effect if (i) we condition on the number of borders; and (ii) we check that the fraction of switchers is small.

3.2 Understanding the border assignment

Geography affects trade costs and market shares. Since geography precedes borders, this poses an identification problem if the border assignment is also affected by geography. But it is easy to see that this is indeed the case. Our comparison of the (Catalonia, Languedoc-Rousillon) and (Catalonia, Valencia) region pairs shows how difficult it is to escape from this conclusion. Both pairs are contiguous, continental and located on the Mediterranean coast. Thus, comparing their market shares already 'controls' for some of the most relevant geographical factors. But even then, we cannot conclude that the location of the French-Spanish border is unrelated to geographical factors that also affect trade. On its north, Catalonia is separated from Languedoc-Roussillon by the Pyrenees mountain range. On its south, Catalonia shares the Ebro river basin with Valencia. This geographical difference, which affects trade costs, might have also contributed to the French-Spanish border being north of Catalonia rather than southwest.

To satisfy the unconfoundedness condition, causal inference must be conditional on those factors that precede and influence both the treatment assignment and the outcome variable. In our framework, these include geographical covariates that affect the border assignment and trade outcomes simultaneously. Ideally, we would also include additional covariates such as language, religion, and cultural values. Unfortunately, we do not have enough information on these variables *before* the current borders were assigned. And we know, as explained in the introduction, that borders have a large influence on language use, the choice of religion or the adoption of cultural values. With this restriction in mind, we collect the following set of covariates for each region pair:

1. *Distance*. Length of the curve linking the central point of the origin region (centroid) and the central point of the destination region, in kilometers. We use a curve since we take into account the curvature of earth's surface.

2. *Insularity*. Dummy variable taking value one if there is the need to cross a sea to reach from one region to the other, and zero otherwise.

3. *Mountain ranges*. Largest altitude difference between two regions, computed as the difference between the highest altitude point and the lowest altitude point along the straight line that joins the centre of the origin region (centroid) and the centre of the destination region.

4. *River basin*. Dummy variable taking value 1 if both regions belong to the same river basin. We consider the largest rivers in Europe. A map of the areas covered by each river basin can be found in figure D.1 in the Appendix.

5. *Remoteness*. We calculate the remoteness of a region as the sum of the bilateral distance from that region to every other region in the sample. Then, we calculate the remoteness of a pair as the average remoteness of both regions.

All these covariates are known to affect bilateral trade, and they can be treated as pre-treatment covariates when considering the border assignment. The next question is whether these covariates also affect the border assignment. Unlike the theory of bilateral trade, which is quite sophisticated and developed at this time, the theory of borders is rough and underdeveloped. Thus, we are forced to rely on some basic conjectures about how these geographical factors affect the costs and benefits of sharing a government.[21]

It seems reasonable to think that distance, insularity and the presence of mountain ranges all raise the costs and lower the benefits of sharing a government. Thus, we would expect these variables to raise the probability of a border assignment. It is less clear however to predict the effects of sharing a river basin. Rivers could be a geographical obstacle such as

[21] The relevant costs and benefits are those borne by whomever decides borders. The decision-maker(s) might be regions in the pair, or other regions elsewhere. Admittedly, the discussion here is quite superficial. See Kitamura and Lagerlöf (2020) for a recent empirical study that examines the effects of geography on European borders.

Table 2: Covariate distributions across treatment groups

	Treatment group mean	Control group mean	Difference (t-stat)
Distance	1213.62	315.64	-898.0
			(-71.79)
Insularity	0.32	0.06	-0.258
			(-27.23)
Mountain Ranges	1473.66	496.08	-977.6
			(-37.95)
River Basin	0.04	0.19	0.153
			(35.81)
Remoteness	1157.47	1075.85	-81.62
			(-17.19)
N	33567	2479	36046

Notes: This table reports the average value of each geographical covariate in the treatment group (column 1) and in the control group (column 2). The last column reports the difference in means (defined as control minus treated). The t-statistics in parentheses.*Distance* is bilateral distance between origin and destination in kilometers, *Insularity* takes value 1 if one of the regions is an island. *Mountain Ranges* is the highest difference in elevation between two regions in metres (difference between highest point and lowest point), *River Basin* takes value 1 if the region pair shares a river basin and *Remoteness* is the average remoteness of the origin and the destination regions.

mountain ranges, but they could also provide a geographical mobility advantage or create externalities that raise the benefits of a shared government. Thus, we do not know a priori whether being in the same river basin raises or lowers the probability of a border assignment. Unconditionally, we would expect remote region pairs to have more borders because they are farther away from each other. Conditioning on other geographical variables, however, we would expect the probability of a border assignment for a region pair to increase with their remoteness. For instance, the probability that a region pair 1000 kilometers apart in the center of Europe has a border is larger than for a region pair 1000 kilometers apart in Scandinavia.

Table 2 provides summary statistics of these geographical covariates in the treatment and control groups. Intranational pairs are closer to each other, less likely to be insular or separated by a mountain range, more likely to share a river basin, and on average less remote. These differences are significant, and have the expected sign.[22]

To obtain a more convincing assessment of the role of geographical covariates on the border assignment, we estimate the propensity score.[23] In particular, we estimate a logistic

[22]The positive sign on the river basin variable is not informative. International pairs are more distant than intranational ones, making it unlikely that the former be located in the same river basin. One needs to control for distance to determine how sharing a river basin affects the border assignment.

[23]The propensity score at covariate values x is the average probability of border assignment for region

regression model, where the log odds ratio of receiving the treatment is modeled as linear in a number of the geographical covariates, with unknown coefficients. We estimate the coefficients by maximum likelihood. To choose how many of our geographical covariates to include in the logistic regression, we follow the recursive procedure recommended in Imbens and Rubin (2015). We find that all the covariates described above should be included.

Table 3, column (1) presents the estimation results from the logit model. The coefficients of the covariates are all significative at the 1 percent level and the model has an R-squared of 0.476. As expected, distance, insularity and mountain ranges raise the probability of a border assignment, while remoteness lowers it. Interestingly, we find that being in the same river basin raises the probability of having a border. It seems thus that rivers promote borders rather than the opposite.

Table 3: Propensity Models

Dependent Variable: Border	Full sample (1)	Trimmed sample (2)
Distance	2.998***	1.893***
	(0.056)	(0.078)
Insularity	1.096***	1.059***
	(0.096)	(0.128)
Mountain Ranges	0.179***	0.283***
	(0.030)	(0.031)
River Basin	0.767***	0.420***
	(0.089)	(0.089)
Remoteness	-3.857***	-3.341***
	(0.155)	(0.168)
Constant	9.129***	11.180***
	(0.992)	(1.029)
N	36046	6110
Pseudo R^2	0.476	0.143

Notes: This table reports the estimation of the logistic regression model, where the log odds ratio of receiving the treatment (having a border) is modeled as linear in a number of the geographical covariates. *Distance* is (log) bilateral distance between origin and destination in kilometers, *Insularity* takes value 1 if one of the regions is an island. *Mountain Ranges* is the highest difference in elevation between two regions in metres (difference between highest point and lowest point, in logs), *River Basin* takes value 1 if the region pair shares a river basin and *Remoteness* is the log of the average remoteness of the origin and the destination regions.

pairs (n, m) with covariates $X_{nm} = x$.

By its own nature, the unconfoundedness assumption cannot be proved formally. But economic theory identifies as potential confounding factors a set of geographical covariates that precede the border assignment and affect trade costs. We have shown that, indeed, these covariates affect the border assignment. Thus, comparisons of units with different treatments can be given a causal interpretation only if we condition for these pre-treatment covariates. The next step is to find the right way to do this necessary conditioning.

3.3 Constructing the 'right' samples

To measure the border effect we estimate a linear regression model of normalized market shares on the border dummy, controlling for the number of borders and the set of geographical covariates:

$$\ln S_{nm} = \alpha + \beta \cdot B_{nm} + \gamma \cdot N_{nm} + \lambda' \cdot X_{nm} + u_{nm} \tag{11}$$

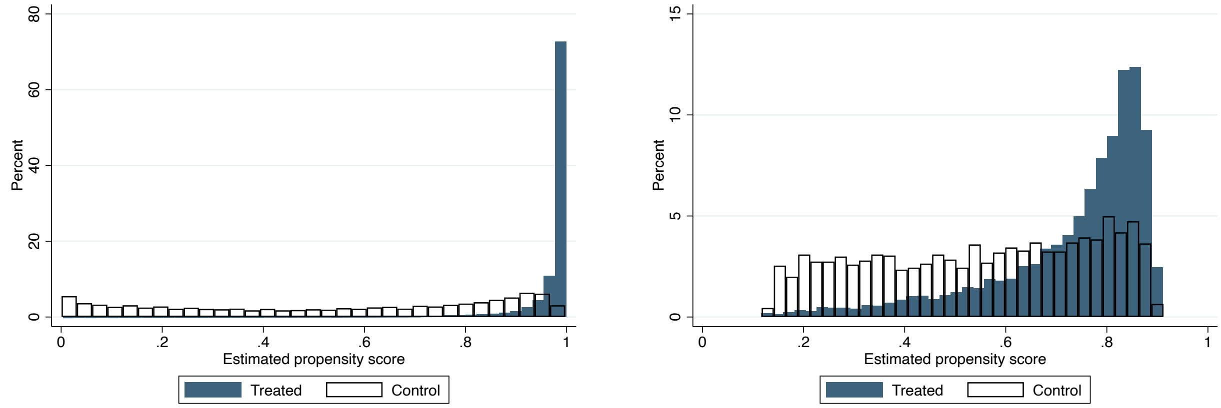
where N_{nm} is the log of the number of borders faced by the region pair, and u_{nm} is a zero-mean error term uncorrelated with the regressors.[24] Since this regression controls for both the number of borders and the pre-treatment covariates, we can use the estimated value $\hat{\beta}$ as a lower bound for the border effect. If we are also able to show that the fraction of switchers is small, then $\hat{\beta}$ is an unbiased estimate of the border effect.

The question we address now is that of choosing the right sample to estimate the regression model in Equation (11). One might initially think that we should use the entire sample. After all, using all the information available is a principled way to proceed. However, Imbens and Rubin (2015) show that the sampling variance of the estimator $\hat{\beta}$ will be large if the population distribution of covariates is unbalanced between treated and control units. Before using regression methods on the entire sample, one needs to ensure that there is enough balance or overlap in the two covariate distributions.

To determine whether there is sufficient overlap in our entire dataset, the left panel in Figure 6 plots the distribution of the estimated propensity score for control units (empty bars) and for treated units (blue shaded bars). The overlap of the propensity score distribution for treated and control units is small. Thus, we trim the data to drop units with extreme values for the estimated propensity score, following the procedure recommended by Crump et al. (2009). This trimming procedure amounts to dropping all observations for which the

[24]The number of borders of a given region equals to 268 minus the number of regions within its country plus 1. The smallest number of borders corresponds to the 38 regions of Germany, with 231 borders. The largest number of borders corresponds to the 2 regions of Slovenia and Croatia, with 267 borders. The variable N_{nm} is the (log) sum of the borders of the region pair. Thus, the values of N_{nm} lie between $\ln(231 \times 2) = 6.1355$ and $\ln(267 \times 2) = 6.2804$.

Figure 6: Histogram of propensity score



A) All region pairs **B) Trimmed sample**

Notes: This figure shows the distribution of the estimated propensity score, probability of having a border, for control units (empty bars) and for treated units (blue shaded bars). Panel A reports the results using the full sample while panel B reports the results using the trimmed sample (dropping region pairs with extreme estimated probability of having border).

propensity score is above or below a threshold determined following a variance criterion.[25] We apply this methodology to our sample and obtain a value of the threshold equal to 6.5 percent. We trim the sample accordingly and re-estimate the propensity score. Column (2) in Table 3 presents the results. The R-squared is now smaller, showing that our covariates explain now a smaller fraction of the variation in the border assignment, as expected after dropping observations in the extremes of the propensity score distribution. The right panel of Figure 6 shows that the distribution of the propensity score across control and treated pairs has a much higher overlap after trimming the initial sample.

To further improve the samples used for inference, we compute a blocking estimator, based on grouping region pairs with similar propensity score values. We call these subsamples blocks. To create them, we follow the procedure recommended by Imbens and Rubin (2015), using the algorithm in Becker and Ichino (2002). This algorithm starts by splitting the sample into 5 equally spaced intervals of the propensity score and then testing whether the average propensity score of treated and control units does not differ much within blocks. If it does, the algorithm splits the interval in half and tests again, until the average propensity

[25] The idea in Crump et al. (2009) is to choose a subset A of the covariate space X so that there is substantial overlap between the covariate distribution for the treated and control units. Crump et al. (2009) use the asymptotic efficiency bound for the efficient estimator for the treatment effect in subset A to choose the trimming threshold. The intuition is that if there is a value of the covariate space such that there are few treated units relative to the number of controls, for this value the variance for an estimator for the average treatment effect will be large. Therefore, excluding units with such covariate values should improve the asymptotic variance of the efficient estimator.

Table 4: Summary statistics of covariates by block

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd
Distance	154.36	186.07	240.35	298.82	349.83	383.02	440.94	480.01	446.70
	61.03	74.23	93.43	121.79	143.55	143.03	161.45	136.84	61.64
Insularity	0.01	0.01	0.01	0.02	0.04	0.07	0.08	0.12	0.22
	0.08	0.12	0.12	0.15	0.20	0.25	0.28	0.33	0.42
Mountain Ranges	208.38	291.05	351.19	466.84	533.75	549.99	596.98	735.32	1244.59
	232.38	320.38	376.25	457.99	528.13	545.14	561.71	681.78	888.16
River Basin	0.29	0.28	0.21	0.19	0.17	0.14	0.12	0.10	0.06
	0.45	0.45	0.41	0.39	0.37	0.35	0.32	0.31	0.24
Remoteness	1169.05	1097.32	1092.09	1087.40	1081.35	1051.59	1038.82	1002.73	938.72
	307.02	268.01	273.50	276.93	275.84	249.16	229.19	187.51	140.79
Propensity score	**0.20**	**0.31**	**0.44**	**0.57**	**0.66**	**0.72**	**0.78**	**0.84**	**0.89**
	0.04	0.04	0.04	0.04	0.02	0.02	0.02	0.02	0.01
N	323	408	515	698	507	660	1062	1582	354

Notes: This table reports the mean and standard deviation of each geographical covariate and the propensity score in each block. *Distance* is bilateral distance between origin and destination in kilometers, *Insularity* takes value 1 if one of the regions is an island. *Mountain Ranges* is the highest difference in elevation between two regions in metres (difference between highest point and lowest point), *River Basin* takes value 1 if the region pair shares a river basin and *Remoteness* is the average remoteness of the origin and the destination regions.

score of treated and control units no longer differs within blocks. Starting from the trimmed sample, this procedure delivers nine blocks. We have ordered these blocks such that the propensity score is increasing.

Table 4 reports the summary statistics of the covariates and the propensity score by block. Recall that there are two factors that reduce the sampling variance of the estimates: (i) the number of observations; and (ii) the balance between treated and control groups. The number of observations varies substantially across blocks, ranging from 323 in Block 1 to 1582 in Block 8. Blocks also vary substantially in terms of their propensity score, ranging from 20 percent in the first block to 89 percent in the ninth one. Blocks 3, 4 and 5 are the most balanced ones with a propensity score of 44, 57 percent and 66 percent, respectively.

Table 5 reports the t-statistic from a difference in means test between treated and controls (test is defined as control mean minus treatment mean). Covariates are well balanced within blocks, with only small differences in means that do not seem to follow a systematic pattern. If the covariates were perfectly balanced within blocks, we could estimate causal effects as if assignment was random within each block. That is, we could compare the means of the international and intranational pairs controlling only for the number of borders. Since three out of five covariates are continuous, however, it is unavoidable to have some small variation in covariates within blocks. In this case, Imbens and Rubin (2015) recommend that these comparisons also control for covariates. Thus, we shall estimate the regression model in Equation (11) for each of the blocks.

To give a sense of the composition of the blocks in terms of regions, Figure 7 shows

Table 5: Balancing test of covariates by block

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Distance	-22.24***	8.207	5.049	4.693	17.13	-11.79	-24.16**	-33.09***	28.87***
	(8.077)	(8.126)	(8.290)	(9.269)	(13.42)	(12.40)	(12.07)	(9.763)	(9.636)
Insularity	-0.00990	0.0206	0.0166	0.0187*	0.0302	0.0125	-0.00573	-0.0613***	-0.00663
	(0.0105)	(0.0132)	(0.0103)	(0.0110)	(0.0190)	(0.0216)	(0.0208)	(0.0234)	(0.0660)
Mountain Ranges	-31.46	25.23	-16.62	-120.6***	-148.1***	-114.3**	-96.62**	-139.1***	45.43
	(31.06)	(35.09)	(33.39)	(34.56)	(49.01)	(47.07)	(41.96)	(48.70)	(140.6)
River Basin	0.0528	-0.0328	-0.00768	0.0366	0.0247	-0.0101	-0.00522	-0.0304	0.0285
	(0.0608)	(0.0495)	(0.0362)	(0.0296)	(0.0350)	(0.0303)	(0.0242)	(0.0219)	(0.0382)
Remoteness	-109.7***	51.41*	30.83	20.83	44.75*	-5.539	-21.53	-54.87***	59.66***
	(40.65)	(29.26)	(24.24)	(21.06)	(25.75)	(21.61)	(17.15)	(13.36)	(22.06)
N	323	408	515	698	507	660	1062	1582	354

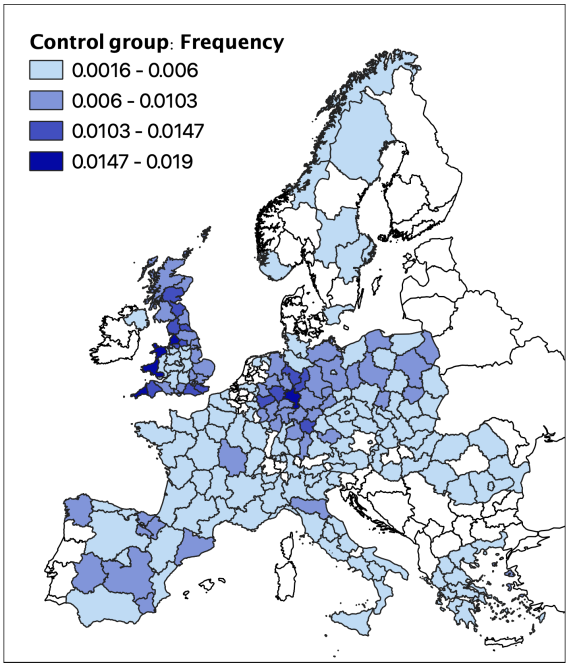
Notes: This table reports the difference in means between treated and control region pairs for each geographical covariate by block (defined as control minus treated). Standard errors in parenthesis, significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.. *Distance* is bilateral distance between origin and destination in kilometers, *Insularity* takes value 1 if one of the regions is an island. *Mountain Ranges* is the highest difference in elevation between two regions in metres (difference between highest point and lowest point), *River Basin* takes value 1 if the region pair shares a river basin and *Remoteness* is the average remoteness of the origin and the destination regions.

the frequency with which each region appears (as a part of a pair) within the control and treated groups in block 4. This block has an average propensity score of 57 percent. That is, region pairs within this block had roughly an equal chance of having a border than not having one. In this block we find regions from all around Europe both in the treated and in the control units. The composition of regions changes across blocks. The figures for all the blocks can be found in the Appendix. As we would expect, blocks 1 and 2 source mostly from region-pairs that are at short distances while blocks 7, 8 and 9 contain regions located in the largest countries, since region-pairs are, on average, further away.[26]

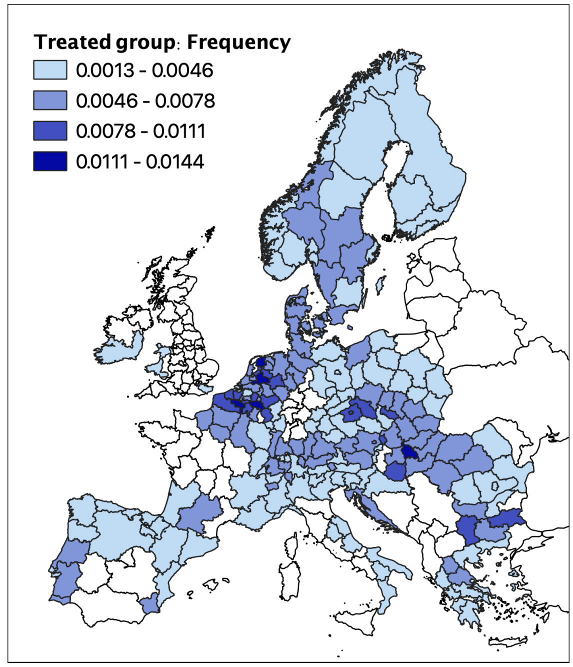
Let us go back to our example of Catalonia, Languedoc-Roussillon and Valencia. Figure 8 shows all the pairs that contain Catalonia (shown in grey) in our sample. The color of each region represents the block in which the corresponding pair is located. White-colored regions are pairs that have been dropped after trimming, for which the probability of a border was close to 1. There is no pair that includes Catalonia in block 1, indicating that the probability of Catalonia having a border with any of its neighbours was always 20 percent or larger. Languedoc-Roussillon is in block 5, where the average probability of a border is about 66 percent; and Valencia is in block 3, where the average probability of a border is about 44 percent.

[26]We say that a region pair has a border if they do not share the same government, independently of whether the regions are in contiguous or non-contiguous countries. It turns out, however, that many of the treated region pairs in our blocks are in contiguous countries. In particular, the percentage of treated pairs that are in contiguous countries are more than 90 percent in blocks 1-4, 85 percent in block 5, 78 percent in block 6, 70 percent in block 7, 64 percent in block 8 and 62 percent in block 9.

Figure 7: Composition of regions in block 4



A) Control group



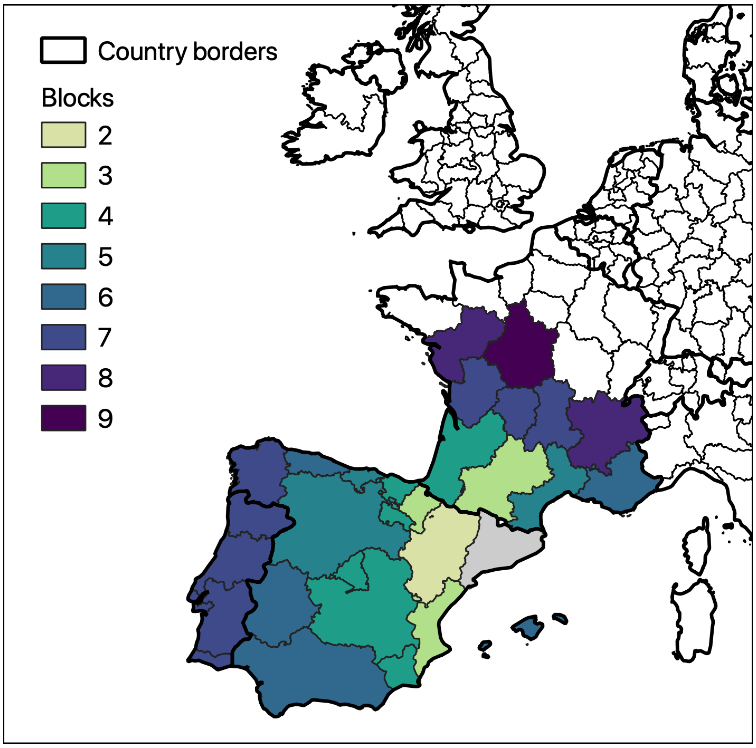
B) Treated group

Notes: This figure shows the regions that are part of the block. The shading represents the frequency with which each region appears (as a part of a pair) in the control group (first panel) and treated group (second panel) in the block.

Figure 8 allows us to illustrate our identification strategy, and the motivation behind our approach. Notice that block 7 contains intranational pairs, in Spain, as well as international pairs, in France and Portugal. The former will be used as control units, while the latter will be used as treated units. Region pairs in block 7 have a probability close to 78 percent of being separated by a border. Given that this probability is very similar across treated and control units, the difference in trade between them can be interpreted as the causal effect of the border.

We have now constructed the samples we needed to estimate the border effect. Before using them, though, we need to assess how important is the participation bias in these samples (recall Equation (7) and the discussion after it). Table 6 shows how participation rates differ between treated and control groups in the entire sample, the trimmed sample and in each of the blocks. Participation rates among control units are high in all the samples. In the entire sample, however, the participation rate among treated units is only 61.7 percent. This must be due to the fact that many international pairs are far away and likely to have a border. Indeed, participation rates in the trimmed sample increase dramatically among the treated, becoming quite close to those in the control group. The participation rates within blocks are even more balanced. Thus, we conclude that the participation bias cannot be

Figure 8: Distribution of Blocks for region-pairs with Catalonia



Notes: This figure shows the regions that are part of a pair that includes Catalonia in the trimmed sample. The colors represent the block in which each region pair is included. The blocks are ordered as increasing in the propensity score. Darker shading represents higher probability of having a border.

large within these blocks. Remarkably, our construction of blocks has achieved an almost perfect balance in participation rates without using any outcome variables in the procedure. This provides additional support for our chosen empirical strategy.

4 Causal effect of borders on trade

Finally we are ready to present our results. We show first our estimation of the average border effect and we continue with the estimation of the border effect across industries. Finally, we present our estimation of the effect of recent borders.

4.1 Average Border effect

Table 7 shows the results of estimating Equation (11) for each of the blocks.[27] Recall that the estimated coefficient on the border dummy is the log reduction in the normalized market

[27]Table B.3 in the appendix reports all the coefficients, including the geographical covariates.

Table 6: Participation rate: Control vs. Treated

	All	Trimmed	Blocks								
			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Part. rate control	0.968	0.976	1	.997	.993	.987	.968	.968	.936	.952	.915
Part. rate treated	0.617	0.946	.993	.996	.996	.969	.947	.957	.95	.928	.894
N	72092	12220	646	816	1030	1396	1014	1320	2124	3164	710

Notes: This table reports the share of region pairs that engage in positive trade in our regional trade dataset (participation rate) for the region pairs in the treated and control groups.

Table 7: Average border effect

Dep. Var: $ln(S_{n,m})$	Block 1 (1)	Block 2 (2)	Block 3 (3)	Block 4 (4)	Block 5 (5)	Block 6 (6)	Block 7 (7)	Block 8 (8)	Block 9 (9)
Border	-1.786***	-1.721***	-1.699***	-1.768***	-1.686***	-1.796***	-1.687***	-1.754***	-1.858***
	(0.182)	(0.178)	(0.175)	(0.175)	(0.238)	(0.289)	(0.268)	(0.290)	(0.201)
Number of Borders	7.058***	6.695***	7.041***	10.779***	11.294***	11.833***	9.234***	8.091***	0.420
	(1.756)	(1.970)	(2.034)	(1.730)	(2.064)	(2.783)	(2.792)	(3.063)	(2.944)
Geographic covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	645	813	1024	1364	968	1267	2011	2948	637
R^2	.572	.533	.501	.47	.375	.388	.31	.285	.299

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable is the (log) normalized market share of n in m. *Border* is a dummy for international border. *Number of borders* is the (log) sum of the number of borders that are faced by n and m.

share caused by the border, that is, the average border effect within the block. This effect is large, statistically significant at the one percent level, and it varies little across blocks. The border effect ranges from a minimum of -1.686 in block 5 to a maximum of -1.858 in block 9, which indicate that borders reduce trade to somewhere between $18.5\,(= \exp\{-1.686\})$ and $15.6\,(= \exp\{-1.858\})$ of their potential.

Table 7 also shows the effect on normalized market shares caused by the number of borders. Recall that the coefficient on this variable measures the elasticity of the normalized market share with respect to the number of borders. This elasticity varies across blocks, ranging from 6.695 in block 2 to 11.833 in block 6. Since $N_{nm} \in [6.1355, 6.2804]$ in our sample, we have that the difference in market shares caused by differences in the number of borders might be substantial. To put an upper bound to this difference, compare the region pair containing the two Slovenian regions, which is in block 1, with a region pair containing two German regions in the same block. According to our estimates, the normalized market share for the Slovenian pair is about $2.78\,(= \exp\{7.058 \times 0.1449\})$ larger than that of the German pair. Thus, our estimates reveal an additional important channel through which the border assignment affects trade. It is not only whether a border is assigned to a specific region pair that matters, but also how many borders are assigned to each region in the pair.

Let us now use these results to be a bit more precise about the counterfactual scenario discussed in the introduction, in which the French-Spanish border is southwest rather than north of Catalonia. Recall that the region pair (Catalonia, Languedoc-Roussillon) is in block 5, and that the change in the French-Spanish border reduces the number of borders of Catalonia by 7 and for Languedoc-Roussillon by 1. Then, we can compute the effect of this change in the border as the product of two separate effects: (i) the average border effect which increases the market share by a factor $5.398\,(=\exp\{1.686\})$; and (ii) the number-of-borders effect which lowers the market share by a factor $0.839\,(=\exp\{11.294(-0.0155)\})$. Thus, our estimates indicate that Catalonia's market share of the Languedoc-Roussillon market would be $4.530(=5.398\times 0.838)$ larger than it is today. Since the region pair (Catalonia, Valencia) is in block 3 and the change in the French-Spanish border increases the number of borders of Valencia by 1, Catalonia's share of the Valencia market would be $0.165\,(=\exp\{-1.699+7.041(-0.0119)\})$ smaller than it is today. These numbers are a bit different from those we showed in the introduction because the latter did not take into account the number-of-borders effect.

Our results refer to the effect of a border dividing two contiguous countries. Our methodology does not allow us to speak about how the market share would be affected in the case of regions that are divided by more than one border (non-contiguous countries). The reason is that, due to our propensity score estimation and blocking procedure, most of the region-pairs with a border that remain in our sample are regions in contiguous countries (by definition control pairs are in the same country and always contiguous). Only in the case of blocks 7, 8 and 9, for which the bilateral distance between regions is higher than 400 kilometers, we include a non-negligible fraction of region-pairs that belong to different and non-contiguous countries. Reassuringly, keeping only region-pairs that belong to contiguous countries to estimate the border effect does not change our results, as can be seen in table B.4 in the Appendix.

Table 8 reports the average border effect, after aggregating our regression results by block. We present two possible average treatment effects, weighting the coefficients by the

Table 8: Average Border Effect (Average treatment effect)

	Estimated β^{ATE}	
	All controls	Without number of borders
Weights: Size of blocks	-1.744	-1.299
Weights: Treated pairs	-1.747	-1.303

Notes: Average treatment effect calculated by computing the weighted average of the estimated coefficient of the *Border* dummy. The first row uses the number of observations in each blocks as weights, while the second row uses the number of treated units in each block.

size of the block (row 1) and weighting by the number of treated units in each block (row 2) (see Imbens and Rubin (2015)). The average effect of the border is negative and large in magnitude, and the weighting method does not make much of a difference. Our findings suggest that the border reduces trade between two regions to 17.5 percent of what they would trade without the border ($\exp\{-1.744\} = 0.175$).

A key step in our identification strategy is to control for the number of borders. This matters not only in itself as argued already, but also to avoid a selection bias problem when estimating the average border effect. As discussed in section 3.1, region pairs with many borders tend to have larger market shares and tend to be over-represented among international pairs and under-represented among intranational pairs. This creates a positive selection bias that makes the observed difference in average market shares smaller (in absolute value) than the true average border effect. To show that this source of selection bias is relevant, the second column of Table 8 reports the estimated average border effect that we would obtain if we failed to control for the number of borders.[28] This biased estimate of -1.299, would lead us to believe that the border reduces normalized market shares to 27.3 percent of their potential instead of the true estimate of 17.5 percent.

Table 9 provides some insights as to the importance of our identification strategy. The first two columns show the results that are obtained if, instead of using our blocking estimator, we use the trimmed and full samples, respectively. Most of the benefit of our procedure comes from trimming the data, rather than using the blocking estimator. For the full sample we obtain an estimate of -1.968, which would lead us to believe that the border reduces the normalized market share to 14 percent of its potential. For the trimmed sample we obtain an estimate of -1.716 which is essentially the same as the one provided by the blocking estimator. This is not surprising ex-post since we have seen that the border estimates do not vary much across blocks. The last two columns use our data to estimate standard structural gravity models with fixed effects. In column 3, we show the estimates including all geographical covariates. In column 4, we show the estimates using only distance, which is the only geographical variable that is included in this type of regression. The gravity estimates are larger than those obtained with the blocking estimator. This suggests that taking into account border endogeneity and constructing balanced samples is important to obtain a causal estimate of the border effect.[29]

[28] Table B.5 reports the regression coefficients for each block.

[29] Since this study uses previously unavailable regional-level trade data, it is important to compare our results with the results of alternative methodologies applied to the same data. Not doing this would make it challenging to disentangle the importance of our methodological contribution from the use of the newer, more granular dataset we use in this study. We do not include time-varying country fixed-effects because the data is aggregated across years as explained in previous sections.

Table 9: Average Border effect using alternative samples and fixed-effects gravity

	Trimmed sample	Full sample	Fixed-effects gravity	
Dependent Var.	$\log(S_{nm})$ (1)	$\log(S_{nm})$ (2)	log(Value) (3)	log(Value) (4)
Border	-1.716***	-1.968***	-1.907***	-2.229***
	(0.184)	(0.211)	(0.159)	(0.250)
Distance	-1.343***	-1.064***	-1.059***	-1.268***
	(0.124)	(0.109)	(0.084)	(0.067)
Number of Borders	8.346***	7.944***		
	(1.647)	(1.807)		
Geographical Covariates	Yes	Yes	Yes	No
Origin FE	No	No	Yes	Yes
Dest. FE	No	No	Yes	Yes
N	11677	46236	46236	46236
R^2	.642	.482	.687	.677

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable is the (log) normalized market share of n in m in columns 1 and 2, and (log) of trade value in euros in columns 3 and 4. *Border* is a dummy for international border. *Distance* is the log of bilateral distance, in kilometers. *Number of borders* is the log of the total number of borders that are faced by n and m. Geographical covariates include (log) bilateral distance, elevation difference, average remoteness, insularity dummy and same river dummy.

4.2 Border effect across industries

The average border effect may hide some cross-industry heterogeneity.[30] We report now the results of estimating Equation (11) industry by industry. Importantly, we can use the estimated propensity score and the same blocks, since both are constructed from region-pair covariates that are constant across industries.

Table 10 presents the results for all industries. The border effect is negative and statistically significant in all blocks in all industries (coefficients represented with confidence intervals in figure A.17 in the Appendix). As we could anticipate, the average border effect masks some heterogeneity. The industry "Food, Beverage and Tobacco", in column (10) of row (3), has a weighted coefficient of -2.095, meaning that the border effect is 0.123. The industry "Textiles", in column (10) of row (4), has a weighted coefficient of -.945, implying that the border effect is 0.389.

Our industries are very aggregated and it is difficult to say much about these differences in

[30]Using total trade flows misses the fact that industries have varying trade cost elasticities (Chen and Novy, 2012) and select into geographies taking into account border related costs. Therefore estimates that employ aggregated data at the industry level risk suffering from compositional bias (Hillberry, 1999).

Table 10: Border effect across industries and blocks

INDUSTRY	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8	Block 9	ATE: W	ATE: T
1. AGRI	-1.851***	-1.813***	-1.659***	-1.384***	-1.241***	-1.611***	-1.413***	-1.620***	-1.995***	-1.578	-1.559
2. MINE	-1.714***	-2.017***	-1.607***	-1.592***	-1.413***	-1.374***	-1.160***	-1.019***	-2.054***	-1.471	-1.395
3. FBT	-2.488***	-2.464***	-2.163***	-2.084***	-2.034***	-2.024***	-1.977***	-1.954***	-2.196***	-2.095	-2.047
4. TEX	-1.333***	-1.195***	-0.714***	-1.053***	-0.830***	-0.915***	-0.714***	-0.839***	-1.307***	-0.945	-0.904
5. WOOD	-1.532***	-1.641***	-1.366***	-1.429***	-1.369***	-1.360***	-1.488***	-1.588***	-1.828***	-1.499	-1.505
6. COKE/PET	-2.025***	-1.314***	-1.221***	-0.787***	-0.702***	-0.776***	-0.507***	-0.601***	-1.592***	-0.995	-0.866
7. CHEM	-1.373***	-1.278***	-1.206***	-1.388***	-1.080***	-1.267***	-1.298***	-1.308***	-1.249***	-1.282	-1.280
8. NON-MET	-1.936***	-1.975***	-1.850***	-2.030***	-1.767***	-1.951***	-1.739***	-1.834***	-2.122***	-1.886	-1.874
9. MET	-1.239***	-1.254***	-1.372***	-1.514***	-1.400***	-1.363***	-1.218***	-1.459***	-1.719***	-1.384	-1.400
10. MACH	-2.260***	-1.841***	-1.834***	-1.698***	-1.286***	-1.511***	-1.364***	-1.619***	-1.430***	-1.627	-1.565
11. VEH	-1.545***	-1.303***	-1.366***	-1.406***	-1.091***	-1.210***	-1.233***	-1.338***	-1.762***	-1.330	-1.321
12. OTHER	-2.029***	-1.589***	-1.361***	-1.494***	-1.372***	-1.283***	-1.272***	-1.165***	-1.716***	-1.406	-1.348
Aggregate BE	-1.786	-1.721	-1.699	-1.768	-1.686	-1.796	-1.687	-1.754	-1.858	-1.744	-1.747

Notes: This table reports the estimated border effect (coefficient on dummy Border, in regression equation (11)) by industry (rows) and block (column).The last two columns report the average border effect computed using as weights the size of the block (ATE: W) and the number of treated region pairs (ATE: T). The last row (Aggregate BE) reports the average border effect across industries, as reported in table 7.

the border effect. But we do notice that less negative coefficients, of around -1.4 are estimated in Chemicals, Metals and Vehicles. While more negative coefficients, of around -1.6, are found in Wood and Cork Products and Paper, Non Metals, Machinery and Agriculture. This is suggestive of an increasing border effect for more differentiated or more transformed goods.

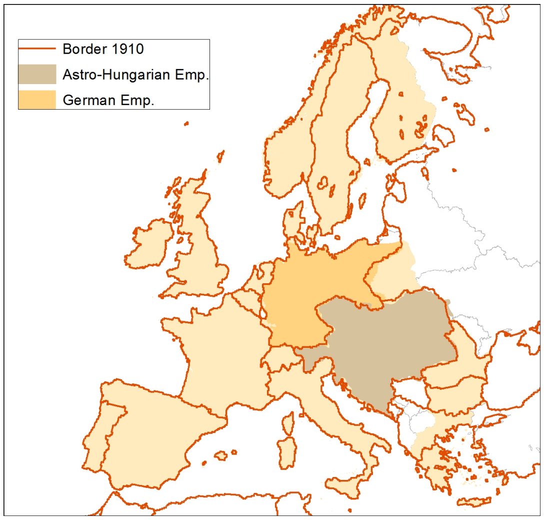
The last row of Table 10 reports the average border effect estimated in the previous subsection. In all industries but two this average effect is larger than the industry border effect. In the first blocks, columns 1 to 4, the estimates of the border effect for some industries are below the average while for some others they are above the average. However, in blocks 5 to 8 we see that the estimates of the border effect for almost all industries are below the average border effect estimated with the aggregated data. At first sight, this seems puzzling, since the average border effect is estimated by aggregating the industry-level data. The explanation for this observation is the imbalance in participation rates between treated and controls in this second set of blocks. As explained in the previous section, differences in participation between treated and control region pairs generate a participation bias that leads to an underestimation of the border effect.[31]

4.3 Effects of post-1910 borders

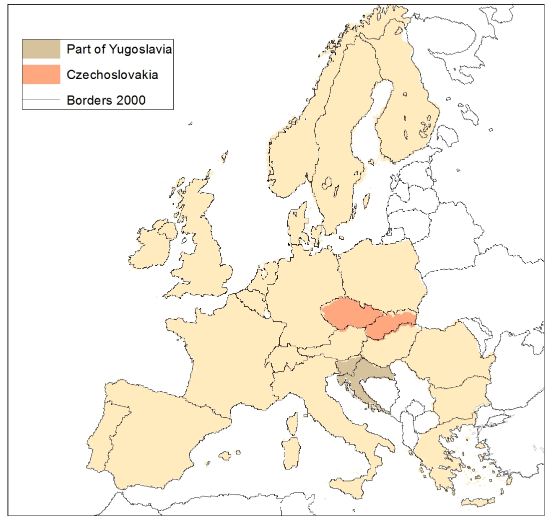
We next examine whether the border effect varies with the age of the border. Our sample contains borders that were created several centuries ago, such as the French-Spanish border, together with borders that were put in place only some decades ago, like the border between the Czech Republic and Slovakia that was established in 1993. It is plausible to think that

[31]Figure A.18 in the Appendix plots the differences in participation (share of trading pairs) between treated and control units in each industry and block. As expected, participation rates are very similar in all industries in blocks 1 to 3, but much larger for control pairs in other blocks.

Figure 9: Recent and old borders



A) Borders in 1910



B) Borders in 2010

Notes: This figure shows European borders in 1910 (panel A) and in 2010 (panel B).

the effects of these borders might be quite different.

Figure 9 shows borders in Europe in 1910 and 2010. The 1910 set of borders is the culmination of a process of political integration that included, for instance, the unification of Italy and Germany. After 1910, this trend reversed. The 2010 set of borders shows the effects of a process of political disintegration which included, for instance, the collapse of the Austro-Hungarian empire and of the former Yugoslavia and Czechoslovakia. Indeed, about one third of the region pairs that shared a government in 1910 no longer share a government in 2010.

We take 1910 as our reference year and split our sample of region pairs into four groups, according to their border history. The largest group consists of regions that are in different countries both in 1910 and in 2010, and contains 90 percent of our observations. The second largest group consists of regions that have always been in the same country, and contains 6.3 percent of our observations. The third largest group consists of regions that were in the same country in 1910, but are no longer in the same country in 2010. This group contains about 3.1 percent of our observations. The final and smallest group consists of regions that were in different countries in 1910 and now are in the same country. This group contains only 0.5 percent of our observations.

To measure the effects of adding a new border, we compare outcomes between the groups that were in the same country in 1910. As mentioned, about a third of the regions who shared a country in 1910, no longer do so in 2010. Thus, we have a good balance between

Table 11: Propensity Models for region pair with border 1910=0

Dependent Variable: Border	Full sample (1)	Trimmed sample (2)
Distance	2.254	2.414
	(0.108)	(0.134)
Insularity	0.270	0.257
	(0.186)	(0.192)
Mountain Ranges	0.071	-0.007
	(0.052)	(0.055)
Same River Basin	1.835	1.914
	(0.120)	(0.136)
Remoteness	-2.293	-2.215
	(0.272)	(0.299)
Constant	1.127	0.065
	(1.844)	(1.965)
N	3422	2630
Pseudo R^2	0.222	0.139

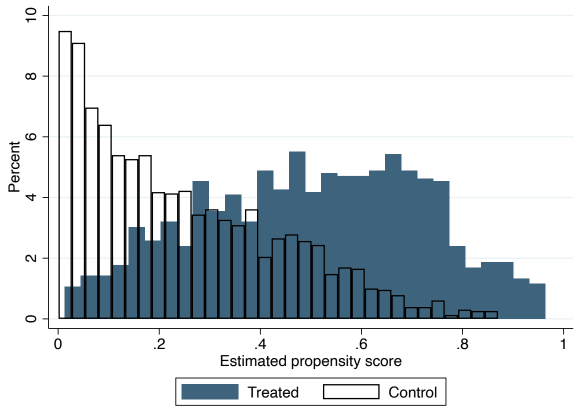
Notes: This table reports the estimation of the logistic regression model, where the log odds ratio of receiving the treatment (having a border) is linear in the geographical covariates. *Distance* is (log) bilateral distance between origin and destination in km, *Insularity* takes value 1 if one of the regions is an island. *Mountain Ranges* is the highest difference in elevation between two regions in metres (difference between highest point and lowest point, in logs), *River Basin* takes value 1 if the region pair shares a river basin and *Remoteness* is the log of the average remoteness of the origin and the destination regions.

treated and controls to perform inference. It would be interesting also to measure the effects of removing an old border by comparing outcomes between the groups that were in a different country in 1910. Unfortunately for our purposes, almost none of the regions in these two groups share a country today. There is simply too much imbalance between treated and controls to perform inference.[32]

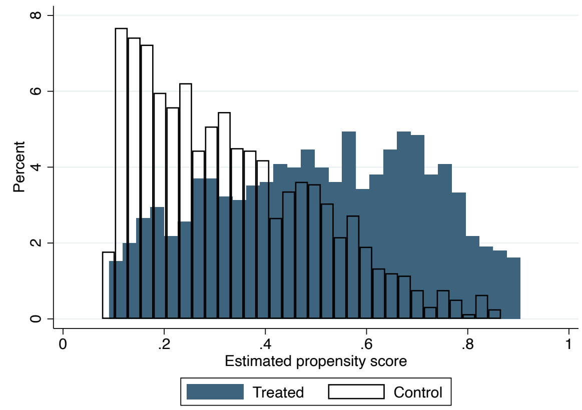
We start with a sample containing the two groups that were in the same country in 1910. Starting from this sample, we repeat the steps explained in section 3. We re-estimate the propensity score and we trim the sample to achieve a good overlap between treated and control units. Table 11 reports the estimation of the propensity score model for the full sample and the trimmed sample, whereas Figure 10 shows the distribution of the propensity score among treated and control units. We then create blocks and report the summary statistics of the covariates and the balancing test in Tables B.6 and B.7 in the Appendix.

[32]Previous studies in the literature have found persistent effects of bygone borders on trade. Nitsch and Wolf (2013) find persistence of the former inner German border on current intra-German trade by road, although the estimated border effect has been declining over time. Beestermöller and Rauch (2018) explore how the trading capital accumulated between members of the Astro-Hungarian empire still drives preferential trade between European countries even after the Fall of the Iron Curtain.

Figure 10: Histogram of propensity score



A) All region pairs



B) Trimmed sample

Notes: This figure shows the distribution of the estimated propensity score, probability of having a border, for control units (empty bars) and for treated units (blue shaded bars). Panel A reports the results using the full sample while panel B reports the results using the trimmed sample (dropping region pairs with extreme estimated probability of having border).

This procedure now generates 6 blocks.

Table 12 reports the results of estimating Equation (11) with this subsample. We find a negative and significant border effect for post-1910 borders, albeit smaller than the average border effect without conditioning on historical borders. The average border effect is -1.261 (-1.221) weighting by the number of region pairs (number of treated). This means that the border reduces the market share to 28.3 percent (29.5 percent) of its potential. These findings show that borders that have been in place for less than a century have large trade reducing effects, although smaller than those of older borders.

5 Language, distance traveled and values disagreement

So the border effect is large. But how does it work? The border effect is the national bias in trade that results from country governments adopting policies that either foster intranational trade or hamper international trade. What are the specific policies that create the border effect? What are the mediating factors through which these policies work? Recall the discussion in the introduction on how governments create a national bias in preferences and a national cost advantage.

We consider three potential mediating factors: language, distance traveled and values disagreement. As we shall discuss shortly, there are ample reasons to believe that (i) borders affect these variables; and (ii) these variables cause trade. If we could establish and measure

Table 12: Average border effect when Border in 1910=1

Dep. Var: $ln(S_{n,m})$	Block 1 (1)	Block 2 (2)	Block 3 (3)	Block 4 (4)	Block 5 (5)	Block 6 (6)
Border	-1.439***	-1.165***	-1.129***	-1.290***	-1.169***	-1.189***
	(0.259)	(0.305)	(0.301)	(0.415)	(0.322)	(0.405)
Number of borders	7.503***	7.325**	7.714**	7.239	8.364*	14.124***
	(2.120)	(2.765)	(3.502)	(4.762)	(4.696)	(4.240)
Geographical covariates	Yes	Yes	Yes	Yes	Yes	Yes
N	1530	1082	894	703	554	298
R^2	.612	.505	.432	.443	.353	.418

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable is the (log) normalized market share of n in m. *Border* is a dummy for international border. *Number of borders* is the (log) sum of the number of borders that are faced by n and m.

the strength of these two links, we would be able to determine the extent to which the border effect is mediated through language, distance traveled and values disagreement. Both of these links are needed to establish that these variables are a mediating factor through which borders cause trade. Each of these links is a necessary condition so that, if one link fails, we can discard the corresponding variable.

Here we provide novel evidence that borders cause differences in language, distance traveled and values. That is, we prove that link (i) is operative for these variables. The very same strategy that we have designed to determine the causal effects of borders on trade can be used to determine the causal effects of borders on other outcome variables. The trimming and blocking procedures developed in section 3 allowed us to build groups of region-pairs that are comparable in terms of geography and among which the assignment of borders is "as-good-as-random". Within these blocks, we can measure the causal effect of borders on any outcome variable by using it as the dependent variable (instead of trade) in Equation (11). This is what we do in this section.

Unfortunately, we cannot prove here that language, distance traveled and values disagreement cause trade (no matter how sensible and probable this might be!). That is, we cannot prove link (ii). This requires us to isolate exogenous variation in the mediating factor of interest to identify its causal effect on trade. This is a major effort that lies outside the scope of this paper. There are a couple of simple and apparently intuitive shortcuts that are often suggested as an alternative. But these shortcuts do not work. The first one is to add the mediating factor alongside the border dummy in Equation (11). This strategy fails because the mediating factor is caused by the border and constitutes a textbook case of bad control. The second shortcut is to instrument the border dummy with that part of the mediating factor that is predicted by the border when estimating Equation (11). This strategy fails because the mediating factor also causes trade directly and, consequently, this instrument does not satisfy the exclusion restriction.

5.1 Language

Several studies have highlighted how governments use policies to homogeneize language in order to create or strengthen a national identity (see, for instance Alesina et al. (2020) and Alesina et al. (2021)). Thus, region-pairs that share a country government may converge in their spoken languages. These studies provide anecdotal evidence, historical narratives and some data correlations that strongly suggest that this is indeed the case in Europe. But they fall short of formally showing that language convergence is caused by a specific government policy. One notable exception is the study by Blanc and Kubo (2023), which uses a regression discontinuity design to show that uniform schooling policies implemented by the French government caused the adoption of the French language.

The notion that language sharing fosters trade is almost taken for granted by much of the literature. Indeed, it has become standard to use a common in standard gravity equations. But we have found only one study that estimates the causal effect of language on trade: Egger and Lassmann (2015). This study focuses on Switzerland and employs a spatial regression discontinuity design to show that having a common native language causes regional trade.[33]

Sharing a language can affect bilateral trade flows in several ways. First, sharing a language reduces communication costs, which are an important part of trade costs. If two regions do not share a language, we expect trade costs to be higher between them. Second, sharing a language may also create familiarity and facilitate trust. People speaking the same language may feel culturally closer to one another in addition to facing lower communication costs. Therefore, if two regions share a language, they may also be culturally closer or share some familiarity that can increase trade through closeness in preferences. If sharing a language also affects trade through culture and trust, sharing 100% of their spoken languages will have stronger effects on trade than sharing 50% of their spoken languages.

To estimate the causal effect of borders on language, we manually collect the language or languages spoken currently in each of the 269 European regions in our sample using language maps from the Encyclopedia Britannica. We measure the degree of language sharing in two ways. First, we use the standard Language Dummy (LD) variable that measures whether a region pair shares any language. This is the traditional measure included in gravity equations.

[33] A related contribution is Melitz and Toubal (2014). This paper does not aim at establishing causality. But it extends the standard analysis by considering different forms of common language, i.e, official, native, spoken.

Table 13: Language Measures by block

Language measure	(1) mean/sd	(2) mean/sd	(3) mean/sd	(4) mean/sd	(5) mean/sd	(6) mean/sd	(7) mean/sd	(8) mean/sd	(9) mean/sd
Language dummy	0.87	0.82	0.70	0.62	0.48	0.43	0.34	0.31	0.36
	0.34	0.39	0.46	0.49	0.50	0.49	0.47	0.46	0.48
Language share	0.79	0.72	0.59	0.51	0.38	0.34	0.27	0.24	0.28
	0.36	0.40	0.44	0.45	0.44	0.43	0.41	0.39	0.40
N	645	813	1024	1364	968	1267	2011	2948	637

Table 14: Effect of borders on common language

DV: $Language_{nm}$	Block1 (1)	Block2 (2)	Block3 (3)	Block4 (4)	Block5 (5)	Block6 (6)	Block7 (7)	Block8 (8)	Block9 (9)
Border Effect	-0.618***	-0.573***	-0.671***	-0.709***	-0.806***	-0.747***	-0.744***	-0.720***	-0.756***
	(0.0933)	(0.0997)	(0.0763)	(0.0568)	(0.0706)	(0.0625)	(0.0719)	(0.0841)	(0.148)
Distance (Geo)	0.0920	-0.184	-0.213	-0.550***	-0.0665	-0.205	-0.325**	-0.313**	-0.0226
	(0.103)	(0.166)	(0.174)	(0.188)	(0.395)	(0.300)	(0.153)	(0.128)	(0.665)
Geo Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	645	813	1024	1364	968	1267	2011	2948	637
R^2	.624	.469	.513	.542	.611	.572	.556	.45	.36

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable, $Language_{n,m}$ is a language dummy, that takes value 1 if regions n and m share a language. *Border* is a dummy for international border. All regressions include as control the *Number of borders*, the (log) sum of the number of borders that are faced by n and m.

Second, we compute a Language Share (LS) variable that measures the number of languages a region pair have in common, as a share of the total number of languages spoken within the region pair. Table 13 reports the mean and standard deviation of these two variables for each block. As expected, blocks of region-pairs closer to each other (blocks 1-5) have higher values of both variables, while blocks of region-pairs further away from each other (blocks 6-9) have lower values. Almost 90% of the region pairs share a language in block 1 while only 36% do so in block 9. The average share of languages in common is almost 80% in block 1 and only 28% in block 9.

We estimate Equation (11) using our two measures of language sharing as the dependent variable. Table 14 reports the border effect on the language dummy, while Table B.9 in the appendix shows the results for the language share. The main finding is that the border has a strong causal effect on language sharing, regardless of the measure used. The average treatment effect indicates that the border reduces the probability of sharing a language by 71.4 percent, and reduces the share of common languages by 70.3 percent. This effect of the border on language sharing is present in all blocks, but it seems to be larger in those blocks with more distant region-pairs. Reassuringly, our geographical covariates are not significant in most blocks (see Tables B.8 and B.9 the Appendix), showing that the balancing of control and treated region-pairs achieved by our procedure is effective.

Thus, we conclude that borders reduce language sharing. Moreover, the effects are quantitatively large. To the extent that language sharing facilitates trade, this evidence suggests that language is likely to be a mediating factor through which borders cause trade. To provide a quantitative assessment, however, we would need also an estimate of the causal effect of languages on trade, which we do not have for our sample.

5.2 Distance traveled

The notion that distance causes trade is not controversial in the literature of international trade. However, distance is often assumed to be exogenous. And yet the distance traveled between two regions also depends, for instance, on the quality and the design of the transport infrastructure available. As discussed in the introduction, a national bias in infrastructure investments creates a national cost advantage that contributes to the border effect. For instance, governments might underinvest near the country borders and over-invest in domestic connections (see Felbermayr and Tarasov (2022)) for a quantitative assessment of this bias). A few studies have explored how borders shape infrastructure investments. For instance, Santamaria (2020) documents that the division of Germany led to the reshaping of highway investments by the West German government in response to the appearance of the Inner German border; and Loumeau (2023) has shown that regional borders shape the transport network between French departments.

Our dataset provides a unique opportunity to test whether borders affect distance. The reason is that the ERFT survey measures the actual distance traveled by individual shipments, reported by each truck. We leverage this granular information in the survey to compute, for each region pair, the distance traveled as the average distance covered by all the trucks moving goods between them, weighted by the size of the shipments. Distance traveled depends on geographical factors (including, of course, geodesic distance) but also on the quality, design and congestion along the transport network. This measure is highly endogenous. This is why we do not use it in our estimation of the effects of borders on trade. However, this variable is ideal for our current purposes. Controlling for geodesic distance and other geographical variables, a larger distance traveled by the truck shipments indicates less direct transport connections between regions.[34]

Table 15 reports, for each block, mean and standard deviation of our measure of distance traveled. As expected, blocks of region-pairs closer to each other (blocks 1-5) have low values

[34]There is one exception, though. For this estimation we drop region-pairs composed by at least one island. This is because the distance traveled reported in the survey is the distance covered by the truck and does not include the distance covered by sea, when the truck is loaded onto a ship. This is not a problem in the rest of the paper because we use the geodesic distance that does not have this problem.

Table 15: Distance Traveled and Geodesic Distance by block

Distance measure	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd
Traveled dist.	209.07	245.68	321.42	403.43	465.36	511.92	591.21	651.39	644.81
	100.07	113.03	142.53	178.68	196.27	200.88	221.12	205.56	117.08
Geodesic dist.	154.51	185.57	239.50	296.90	345.43	380.30	438.53	477.36	448.54
	61.20	74.44	93.62	120.78	141.90	139.58	159.26	126.16	55.42
N	641	803	1010	1340	932	1190	1858	2606	505

Table 16: Border effect on distance traveled

DV: Dist Traveled$_{nm}$	Block1	Block2	Block3	Block4	Block5	Block6	Block7	Block8	Block9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Border	0.158**	0.0988***	0.101***	0.0717***	0.0612*	0.0168	0.0478	0.0398	0.0545**
	(0.0604)	(0.0292)	(0.0329)	(0.0206)	(0.0347)	(0.0429)	(0.0296)	(0.0344)	(0.0265)
Distance (Geo)	0.710***	0.952***	1.015***	1.039***	0.720***	0.987***	0.890***	0.865***	0.521**
	(0.122)	(0.104)	(0.110)	(0.0686)	(0.176)	(0.131)	(0.0917)	(0.0657)	(0.244)
Geo Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	641	803	1010	1340	932	1190	1858	2606	505
R^2	.733	.822	.803	.838	.76	.711	.724	.584	.372

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable, Dist traveled$_{n,m}$ is the average distance traveled by a truck taking a shipment from n to m. *Border* is a dummy for international border. All regressions include as control the *Number of borders*, the (log) sum of the number of borders that are faced by n and m.

for distance traveled, while blocks of region-pairs further away from each other (blocks 6-9) have a higher values. To facilitate comparisons, Table 15 also provides the mean and standard deviation of the geodesic distance by block. As expected, the survey distance is larger than the geodesic distance. And these differences seem to be almost constant across blocks.

We estimate Equation (11) using distance traveled as the dependent variable. Table 16 reports the results. Borders significantly increase the distance traveled by trucks when delivering goods for region-pairs located within 400 kilometers from each other (in geographical distance). The effect goes down in blocks of region-pairs further from each other, and it is not significant in blocks 6, 7 and 8. The average treatment effect tells us that this effect is large: for the average region pair, a truck that crosses an international border would to travel a 6.2 percent longer distance. As expected, geodesic distance is a very important determinant of the distance traveled, with regression coefficients of around 1 and highly significant. Other covariates seem much less important (as shown in Table B.10 in the Appendix).

Thus, we conclude that borders cause an increase in distance traveled. While the European Union has been working towards an integration of the transport networks, border effects are still visible in the distance traveled by truck shipments across Europe.

5.3 Values disagreement

The literature on nation-building has highlighted how governments have historically affected values and beliefs through institutions and policies. In a historical setting, Cinnirella and Schueler (2018) show that state spending on primary education in Imperial Germany affected participation in general elections and votes for pro-nationalist parties. Using a survey of European citizens, Becker et al. (2016) use a regression discontinuity to show that citizens in regions that were part of the Astro-Hungarian empire have a higher trust in the government and other national institutions, than neighbouring regions that remained outside of the Astro-Hungarian empire's border. This common belief on national institutions is a legacy of the high-skilled, efficiently-operating bureaucracy established during the Habsburg Empire.

Culture shapes economic interactions (Guiso et al., 2006). Guiso et al. (2009), for instance, show that cultural relationships affect trust and are an important omitted factor that determines international trade, foreign direct investments and portfolio investments between European countries.[35] Thus, differences in values, attitudes and beliefs across countries may affect trade participation and trade flows. Citizens may not like brands that they identify with foreign values. Entrepreneurs may have difficulties approaching or trusting other entrepreneurs with different attitudes or beliefs.

We use the European Values Survey to measure attitudes towards economic and political questions across European countries, which is available at the NUTS2 level (level of disaggregation of our regional dataset). We use waves 2017-2022 to compute a disagreement index for 32 relevant questions that cover attitudes towards different democratic elements, attitudes towards foreigners and immigrants and views about work, environmental protection and the justifiability of different anti-social behaviours.[36] Using all these variables, we compute a region-pair disagreement index as follows:

$$\text{Disagreement}_{nm} = \sum_{i=1}^{32} |score_{i,n} - score_{i,m}| \tag{12}$$

This disagreement measure goes from 0 (total agreement) to 32 (total disagreement). Table 17 reports the mean and standard deviation of this variable for each block. As expected,

[35]A set of core common values and beliefs are at the heart of many projects of international cooperation and integration such as the European Union. A clear example is provided by Article 2 of the Maastricht Treaty: "The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail."

[36]Table B.12 in the Appendix lists the variables used in the Disagreement Index

Table 17: Disagreement in Values by block

	(1) mean/sd	(2) mean/sd	(3) mean/sd	(4) mean/sd	(5) mean/sd	(6) mean/sd	(7) mean/sd	(8) mean/sd	(9) mean/sd
Disagreement Index	1.83	1.91	2.11	2.33	2.47	2.47	2.33	2.42	2.37
	1.02	1.03	1.06	1.05	1.08	.862	.758	.726	.808
N	442	476	546	646	452	522	750	1010	216

Table 18: Border effect on disagreement in values

DV: Disagreement$_{nm}$	Block1 (1)	Block2 (2)	Block3 (3)	Block4 (4)	Block5 (5)	Block6 (6)	Block7 (7)	Block8 (8)	Block9 (9)
Border	0.489	0.671**	0.637**	0.469	1.102***	0.870***	0.796***	0.348**	0.188
	(0.364)	(0.283)	(0.292)	(0.429)	(0.303)	(0.279)	(0.251)	(0.172)	(0.342)
Distance (Geo)	-0.191	-0.821**	-0.187	0.049	-1.902	0.478	0.177	-0.359	2.159
	(0.341)	(0.346)	(0.764)	(0.568)	(1.467)	(0.677)	(0.570)	(0.613)	(2.321)
Geo Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	442	476	546	646	452	522	750	1010	216
R^2	.333	.315	.344	.192	.286	.215	.143	.0913	.135

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable is the value disagreement between n and m. *Border* is a dummy for international border. All regressions include as control the *Number of borders*, the (log) sum of the number of borders that are faced by n and m.

blocks of region-pairs closer to each other (blocks 1, 2 and 3) have lower disagreement values, while blocks of region-pairs further away from each other (blocks 4-9) have higher disagreement values. Notice that the mean disagreement in values is 2.2, and the disagreement measure ranges from 0.42 to 7.13. These numbers tell us that disagreement in values seems to be small across European regions today.

We estimate equation (11) in each of our 9 blocks using as dependent variable our measure of disagreement in values. The results are reported in Table 18. The first row shows the effect of borders on value disagreement. The main finding is that, in most blocks, having a border increases the disagreement measure between a region-pair. The effect of the border is not constant. The border effect goes from increasing disagreement by 0.348 points, an increase in disagreement of 15 percent of the disagreement mean (in block 8), to increasing disagreement by 1.1, an increase of 50 percent of the disagreement mean (in block 5). In blocks 1, 4 and 9, by contrast, the border coefficient is positive but of a smaller magnitude and not statistically significant. The average treatment effect of borders on the value disagreement index is of 0.618 points. This means that, for the average region-pair, having a border increases disagreement by 0.618 points, which represents an increase of 30 percent of the average disagreement (average disagreement = 2.2). As we observed before, we again see that the geographical covariates do not seem to be significant in any of the blocks (table B.11 in the Appendix).

Thus, we conclude that borders cause disagreement in values. Moreover, the effects are substantially large. To the extent that sharing values and beliefs about political and economic issues may facilitate trade interactions, our findings suggest that value disagreement is likely to be a mediating factor through which borders shape trade patterns.

6 Concluding remarks

In this paper we have built a European regional trade dataset and we have estimated the average border effect on trade flows using a new identification framework. Our results show that the effects of country borders on trade flows within Europe are large. Take two similar region pairs, the first one containing regions in different countries and the second one containing regions in the same country. The market share of the origin region in the destination region for the international pair is only 17.5 percent that of the intranational pair. We refer to this estimate as the average border effect. It seems, then, that we are still far from having a single market in Europe. Country borders have created a national bias in preferences and a national cost advantage that penalize international trade and foster intranational trade. How do country borders affect trade flows? What are the welfare implications? Providing satisfactory answers to these questions is a major research goal on its own, one which is likely to deliver important policy implications for Europe.

We view our contribution as part of a broader research program on the effects of country borders within Europe. To start with, we are currently using our new dataset and the empirical framework developed here to measure the effect of regional governments. These also make decisions about procurement, infrastructure, laws and regulations and so on. What is the effect of regional borders on trade? This project will allow us to obtain a more detailed and precise picture of the effects of different types of political borders.[37]

The broader research program we envision should go beyond estimating the size of border effects, and also try to disentangle the relative importance of the different channels through which country borders affect trade. Here we have provided some evidence suggesting that borders cause language differences, raise distance traveled and foster value disagreements. Some further insight can be obtained by looking at differences in the estimates across industries and between new and old borders provided here. But this only scratches the surface. One would like to have precise answers to questions such as: How much would the border effect be reduced if the European Union were able to eliminate the large observed national bias in government procurement? How much would the border effect be reduced if the European

[37]There are a few papers that have looked at the effects of regional borders using the gravity framework. For instance Wolf (2000), Coughlin and Novy (2012) and Garmendia et al. (2012).

Union were able to build a truly European transportation network?

The research program we have in mind should also go beyond trade flows and examine the effects of country borders on other economic and social interactions. Country borders have implications that go far beyond trade flows. The approach developed here could also be used to measure the effect of borders on migration and investment flows, cultural values, travel and tourism, cooperation in research projects, joint sports activities, and so on. It would be useful to have a broader picture of how country borders within Europe affect economic and social interactions among its regions.

Carrying out this project also made it clear to us that we need a richer theory. Our results suggest that modeling borders is crucial to understand the patterns of intranational and international trade. We have wonderful quantitative theories of trade that realistically model the incentives and constraints faced by consumers and firms. But these quantitative theories rarely include a realistic description of the incentives and constraints faced by governments. If modeled at all, governments either act mechanically or solve some unrealistic social planner problem. How are procurement decisions made? How are infrastructures chosen? How are laws and regulations decided and enforced? Only a realistic and detailed modeling of the behavior of governments can shed light on the channels through which political borders affect trade and welfare. Fortunately, there is a lot of excellent work on the political economy of trade policy to draw upon for this purpose (See, for instance, Grossman and Helpman (2001)).

Much less developed is the theory of country borders. It is here where we have felt more at sea when working on this project. Understanding the border assignment is key to develop a sound identification strategy. And yet there does not exists a theory of borders that is developed at the same level of sophistication, say, than the theory of international trade. There exist some classic approaches to modeling and understanding country formation (see Spolaore and Alesina (2003)); and some recent ones too (see Cervellati et al. (2019) and Gancia et al. (2020)). But these theoretical frameworks can only be seen as promising prototypes, much work is needed to develop them into a fully fledged theory capable of guiding quantitative and empirical research.

References

Alchian, A. A. and W. R. Allen (1964). *Exchange and Production; Theory in Use*. Wadsworth Publishing Company.

Alesina, A., P. Giuliano, and B. Reich (2021). Nation-Building and Education. *The Economic Journal 131*(638), 2273–2303.

Alesina, A., B. Reich, and A. Riboni (2020, December). Nation-building, nationalism, and wars. *Journal of Economic Growth 25*(4), 381–430.

Anderson, J. E. and E. Van Wincoop (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review 93*(1), 170–192.

Baier, S. L. and J. H. Bergstrand (2009, February). Bonus vetus OLS: A simple method for approximating international trade-cost effects using the gravity equation. *Journal of International Economics 77*(1), 77–85.

Bailey, M., A. Gupta, S. Hillenbrand, T. Kuchler, R. Richmond, and J. Stroebel (2020). International trade and social connectedness. *Journal of International Economics*, 103418.

Becker, S. O., K. Boeckh, C. Hainz, and L. Woessmann (2016). The empire is dead, long live the empire! long-run persistence of trust and corruption in the bureaucracy. *The Economic Journal 126*(590), 40–74.

Becker, S. O. and A. Ichino (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal 2*(4), 358–377.

Beestermöller, M. and F. Rauch (2018). A dissection of trading capital: Trade in the aftermath of the Fall of the Iron Curtain. *The Journal of Economic History 78*(2), 358–393.

Blanc, G. and M. Kubo (2023, November). French. (2308).

Cervellati, M., S. Lazzaroni, G. Prarolo, and P. Vanin (2019). Political geography and pre-industrial development: A theory and evidence for Europe 1000-1850. London, CEPR.

Chen, N. (2004). Intra-national versus international trade in the European Union: Why do national borders matter? *Journal of International Economics 63*(1), 93–118.

Chen, N. and D. Novy (2012). On the measurement of trade costs: Direct vs. indirect approaches to quantifying standards and technical regulations. *World Trade Review 11*, 401.

Cinnirella, F. and R. Schueler (2018). Nation building: The role of central spending in education. *Explorations in Economic History 67*, 18–39.

Combes, P.-P., M. Lafourcade, and T. Mayer (2005). The trade-creating effects of business and social networks: Evidence from France. *Journal of International Economics 66*(1), 1–29.

Coughlin, C. C. and D. Novy (2012). Is the International Border Effect Larger than the Domestic Border Effect? Evidence from U.S. Trade. CEP discussion papers, Centre for Economic Performance, LSE.

Coughlin, C. C. and D. Novy (2021). Estimating border effects: The impact of spatial aggregation. *International Economic Review 62*(4), 1453–1487.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with Limited Overlap in Estimation of Average Treatment Effects. *Biometrika 96*(1), 187–199.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological) 41*(1), 1–15.

Egger, P. H. and A. Lassmann (2015, 05). The Causal Impact of Common Native Language on International Trade: Evidence from a Spatial Regression Discontinuity Design. *The Economic Journal 125*(584), 699–745.

Evans, C. (2003). The economic significance of national border effects. *American Economic Review 93*(4), 1291–1312.

Feenstra, R. C. (2002). Border effects and the gravity equation: Consistent methods for estimation. *Scottish Journal of Political Economy 49*(5), 491–506.

Felbermayr, G. and A. Tarasov (2022). Trade and the spatial distribution of transport infrastructure. *Journal of Urban Economics 130*(C), S009411902200050X.

Fukao, K. and T. Okubo (2004). Why has the border effect in the japanese market declined? the role of business networks in East Asia. Technical report, Graduate Institute of International Studies.

Gallego, N. and C. Llano (2015). Thick and thin borders in the European Union: How deep internal integration is within countries, and how shallow between them. *The World Economy 38*(12), 1850–1879.

Gancia, G., G. A. Ponzetto, and J. Ventura (2020). Globalization and political structure. *CEPR Discussion Paper No. DP11159*.

García-Santana, M. and M. Santamaría (2023). Understanding home bias in procurement.

Garmendia, A., C. Llano, A. Minondo, and F. Requena (2012). Networks and the disappearance of the intranational home bias. *Economics Letters 116*(2), 178 – 182.

Gil-Pareja, S., R. Llorca-Vivero, J. A. Martínez-Serrano, and J. Oliver-Alonso (2005). The border effect in Spain. *World Economy 28*(11), 1617–1631.

Grossman, G. M. and E. Helpman (2001). *Special Interest Politics*. MIT press.

Guiso, L., P. Sapienza, and L. Zingales (2006, May). 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? *The Quarterly Journal of Economics 124*(3), 1095–1131.

Head, K. and T. Mayer (2009). Illusory border effects: distance mismeasurement inflates estimates of home bias in trade. In *In: The Gravity Model in International Trade: Advances and Applications. Editors: Bergeijk and Brakman*. Citeseer.

Head, K. and T. Mayer (2014). Gravity equations: Workhorse, toolkit, and cookbook. In *Handbook of International Economics*, Volume 4, pp. 131–195. Elsevier.

Helpman, E., M. Melitz, and Y. Rubinstein (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics 123*(2), 441–487.

Herz, B. and X.-L. Varela-Irimia (2020). Border Effects in European Public Procurement. *Journal of Economic Geography*.

Hillberry, R. (1999). Explaining the 'border effect': what can we learn from disaggregated commodity flow data. *Indiana University Graduate Student Economics Working Paper Series*.

Hillberry, R. and D. Hummels (2008). Trade responses to geographic frictions: A decomposition using micro-data. *European Economic Review 52*(3), 527–550.

Hummels, D. and A. Skiba (2004). Shipping the good apples out? an empirical confirmation of the alchian-allen conjecture. *Journal of political Economy 112*(6), 1384–1402.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kitamura, S. and N.-P. Lagerlöf (2020, 07). Geography and State Fragmentation. *Journal of the European Economic Association 18*(4), 1726–1769.

Loumeau, G. (2023, 07). Regional Borders, Commuting and Transport Network Integration. *The Review of Economics and Statistics*, 1–45.

McCallum, J. (1995). National borders matter: Canada-US regional trade patterns. *The American Economic Review 85*(3), 615–623.

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

Nitsch, V. (2000). National borders and international trade: evidence from the European Union. *Canadian Journal of Economics/Revue canadienne d'économique 33*(4), 1091–1105.

Nitsch, V. and N. Wolf (2013). Tear down this wall: on the persistence of borders in trade. *Canadian Journal of Economics/Revue canadienne d'économique 46*(1), 154–179.

Rauch, J. E. (2001). Business and social networks in international trade. *Journal of Economic Literature 39*(4), 1177–1203.

Santamaria, M. (2020). Reshaping Infrastructure: Evidence from the division of Germany. The Warwick Economics Research Paper Series (TWERPS) 1244, University of Warwick, Department of Economics.

Schulze, M.-S. and N. Wolf (2009). On the origins of border effects: insights from the Habsburg Empire. *Journal of Economic Geography 9*(1), 117–136.

Silva, J. S. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and Statistics 88*(4), 641–658.

Spolaore, E. and A. Alesina (2003). *The Size of Nations.* Mit Press Cambridge, MA.

Turrini, A. and T. van Ypersele (2010). Traders, courts, and the border effect puzzle. *Regional Science and Urban Economics 40*(2-3), 81–91.

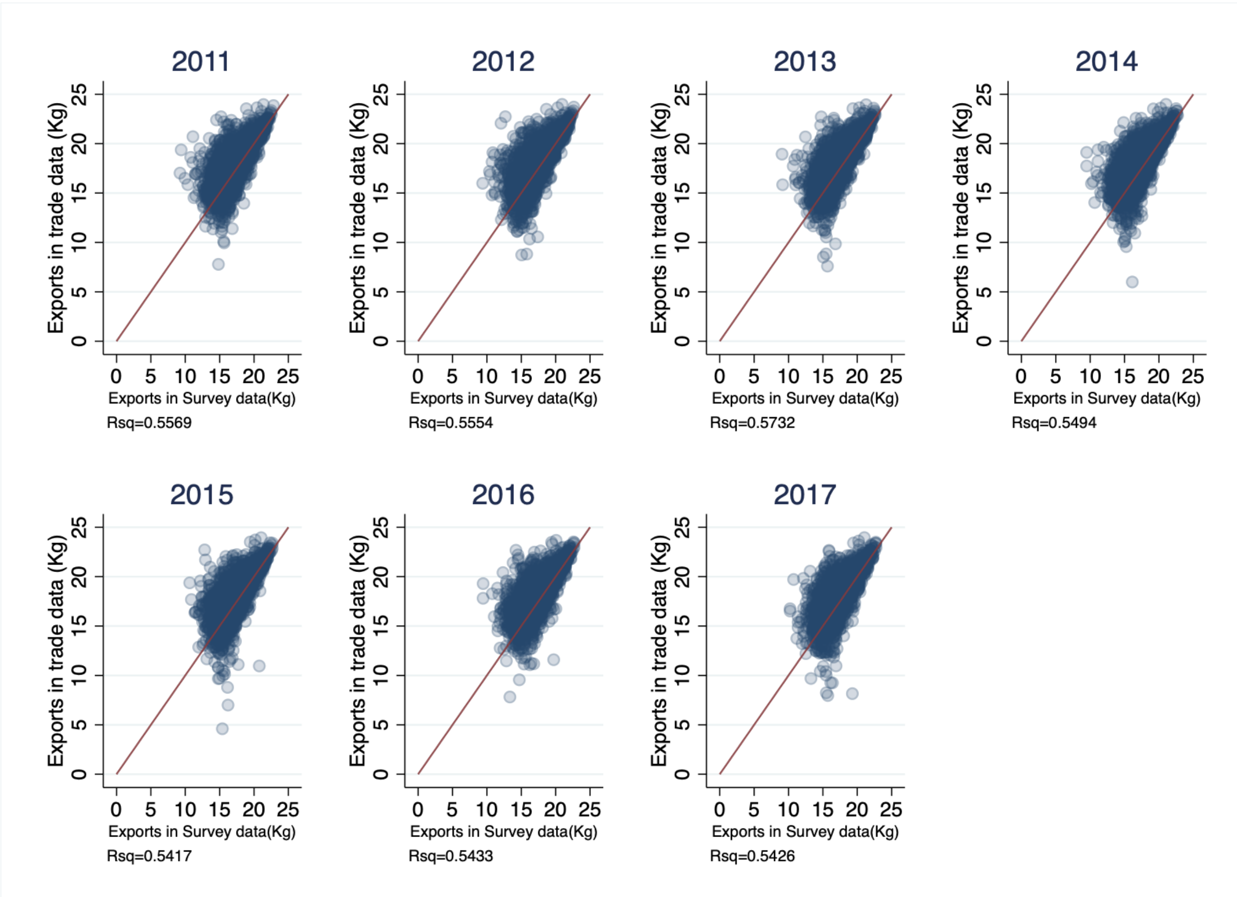
Wei, S.-J. (1996). Intra-national versus international trade: how stubborn are nations in global integration? Technical report, National Bureau of Economic Research.

Wolf, H. C. (2000). Intranational home bias in trade. *Review of Economics and Statistics 82*(4), 555–563.

SUPPLEMENTARY APPENDIX (For online publication only)

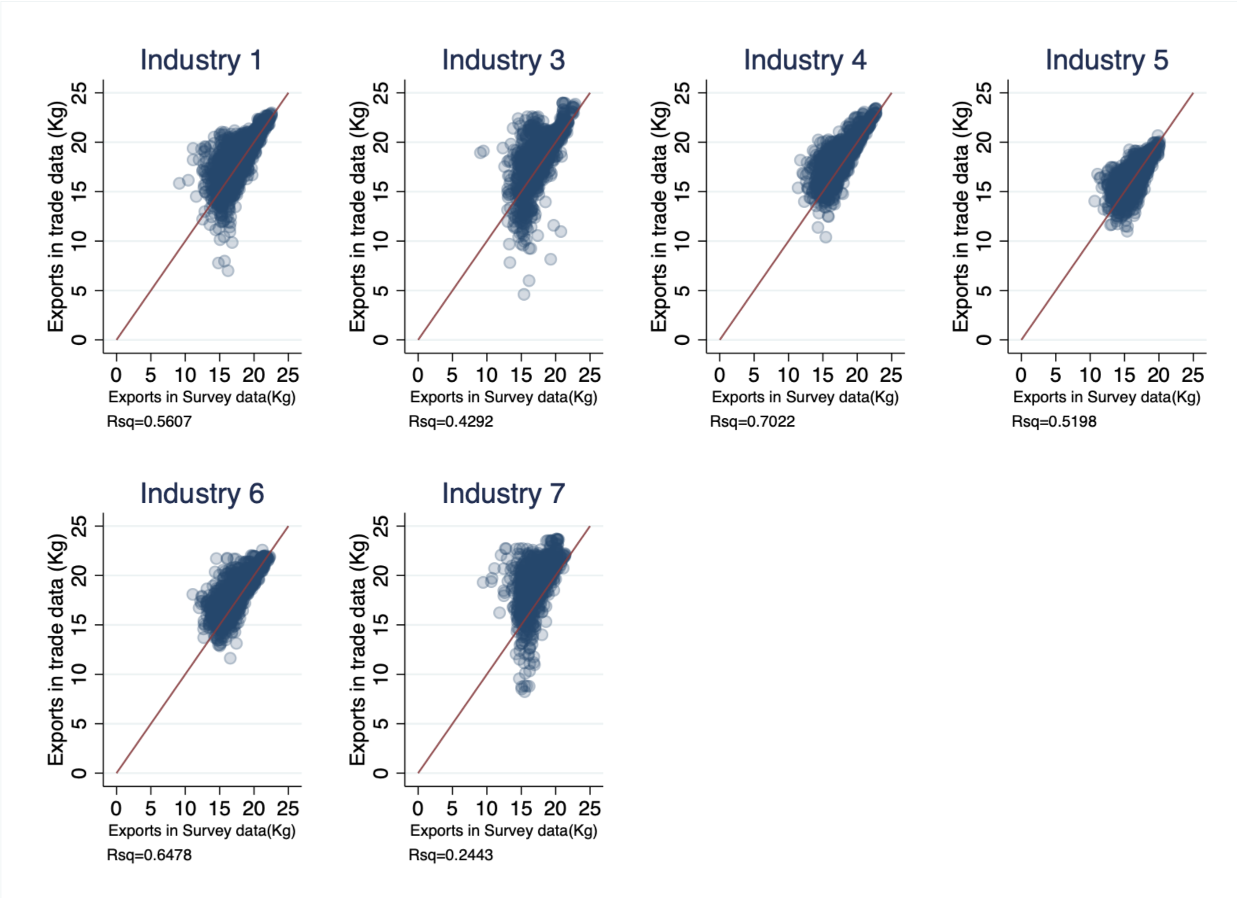
A Additional Figures

Figure A.1: Correlation with aggregate international trade data



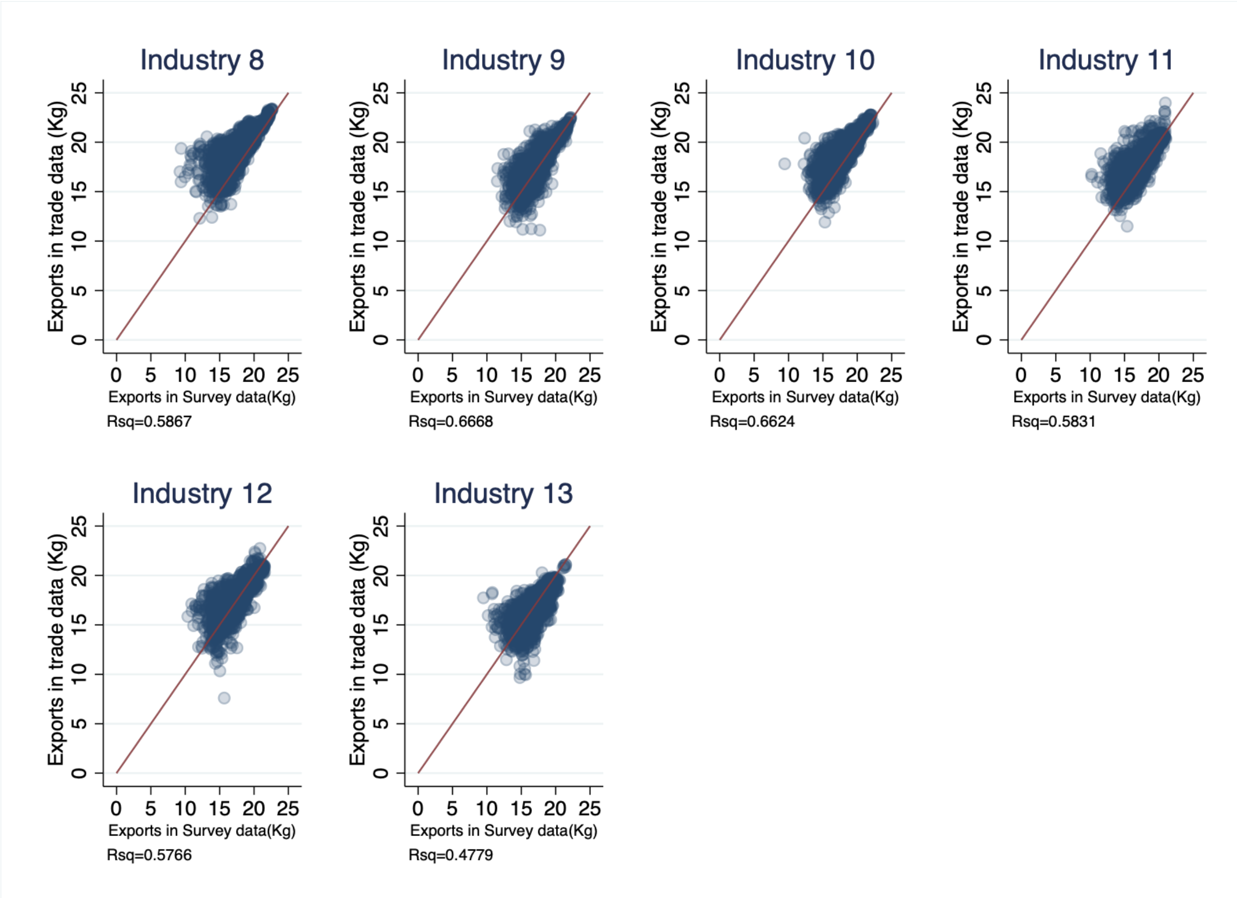
Notes: The figures show the correlation between exports and shipments in the ERFT survey in kilograms in each year. The Y-axis represents (log) bilateral trade (kg) by country-pair-industry-year using international trade data from Eurostat. The X-axis shows bilateral shipments by road (kg) aggregated by country-pair-industry-year obtained from the ERFT survey.

Figure A.2: Correlation with aggregate international trade data



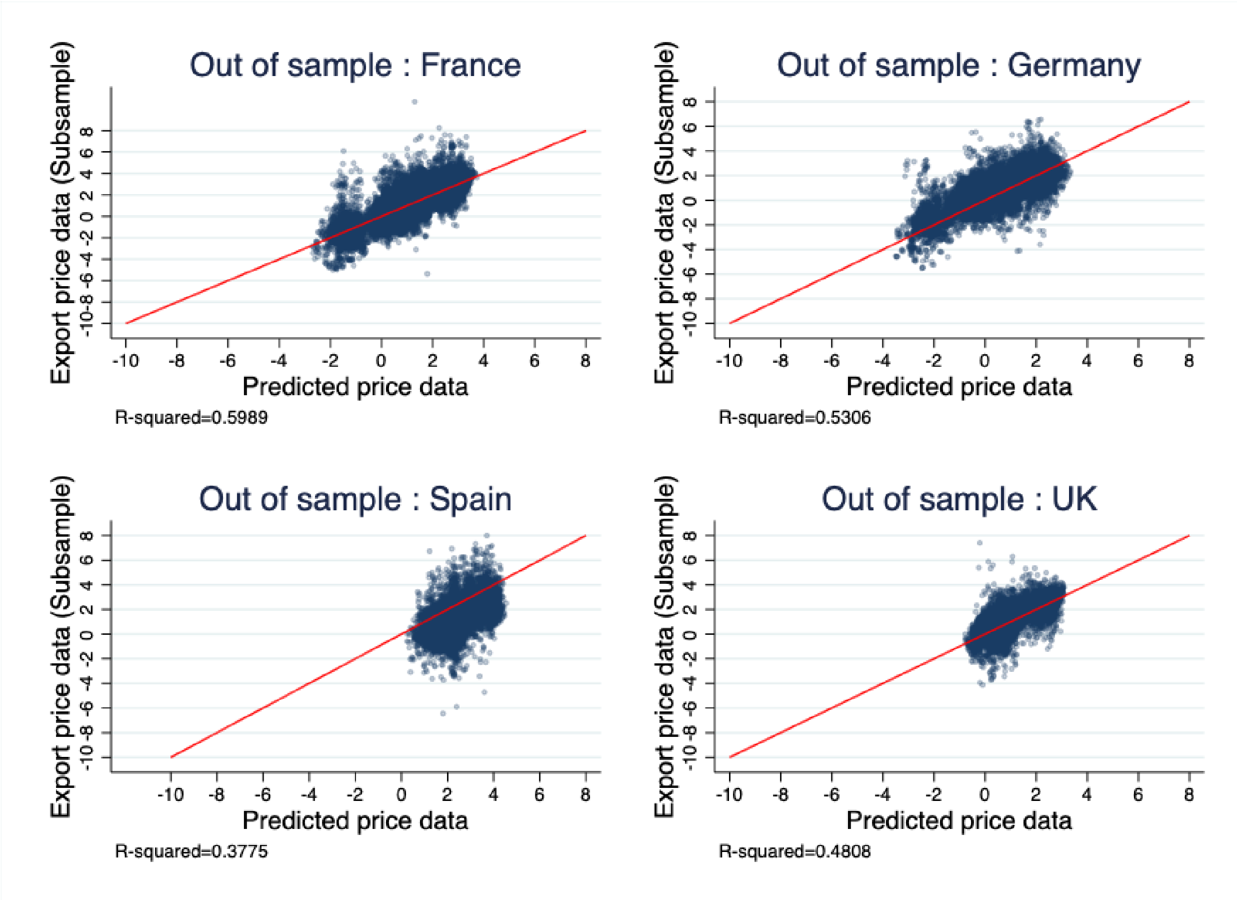
Notes: The figures show the correlation between exports and shipments in the ERFT survey in kilograms in each industry. The Y-axis represents (log) bilateral trade (kg) by country-pair-industry-year using international trade data from Eurostat. The X-axis shows bilateral shipments by road (kg) aggregated by country-pair-industry-year obtained from the ERFT survey.

Figure A.3: Correlation with aggregate international trade data



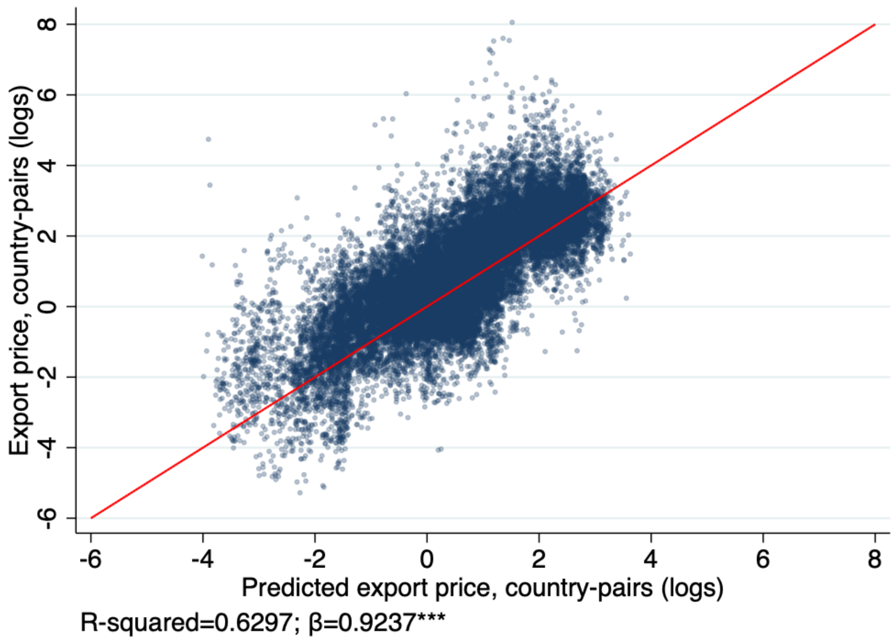
Notes: The figures show the correlation between exports and shipments in the ERFT survey in kilograms in each industry. The Y-axis represents (log) bilateral trade (kg) by country-pair-industry-year using international trade data from Eurostat. The X-axis shows bilateral shipments by road (kg) aggregated by country-pair-industry-year obtained from the ERFT survey.

Figure A.4: Out-of-sample Estimates



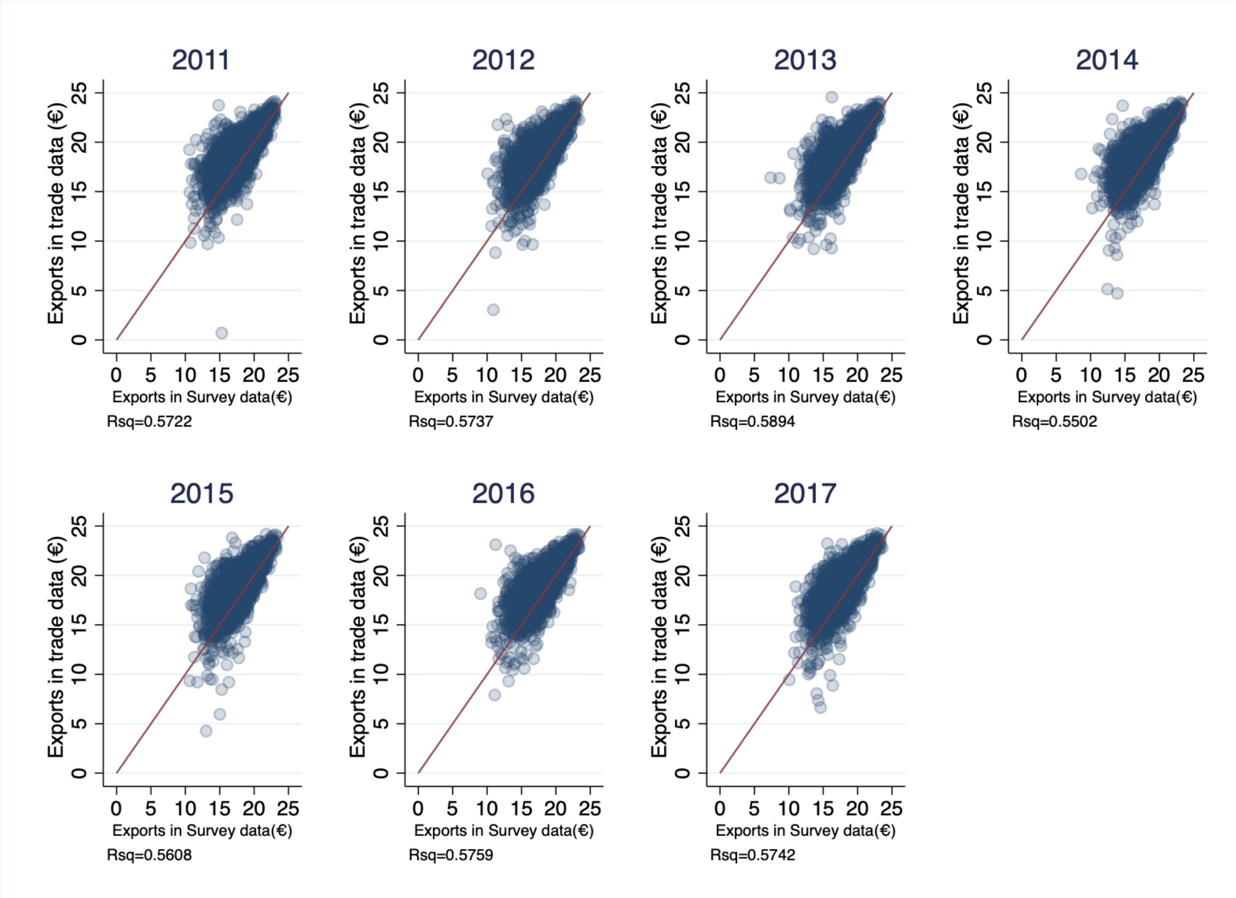
Notes: These figures show the out-of sample check to confirm the performance of the price imputation methodology. Each figure reports the (log) price per kg of exports of France, Germany, Spain and UK to all the countries in our sample by industry and year. The X-axis reports the estimated (log) price per kg of shipment in our regional trade dataset aggregated at the country-pair-industry-year level, predicted when we drop France, Germany, Spain and UK respectively.

Figure A.5: Country-to-Country Estimates



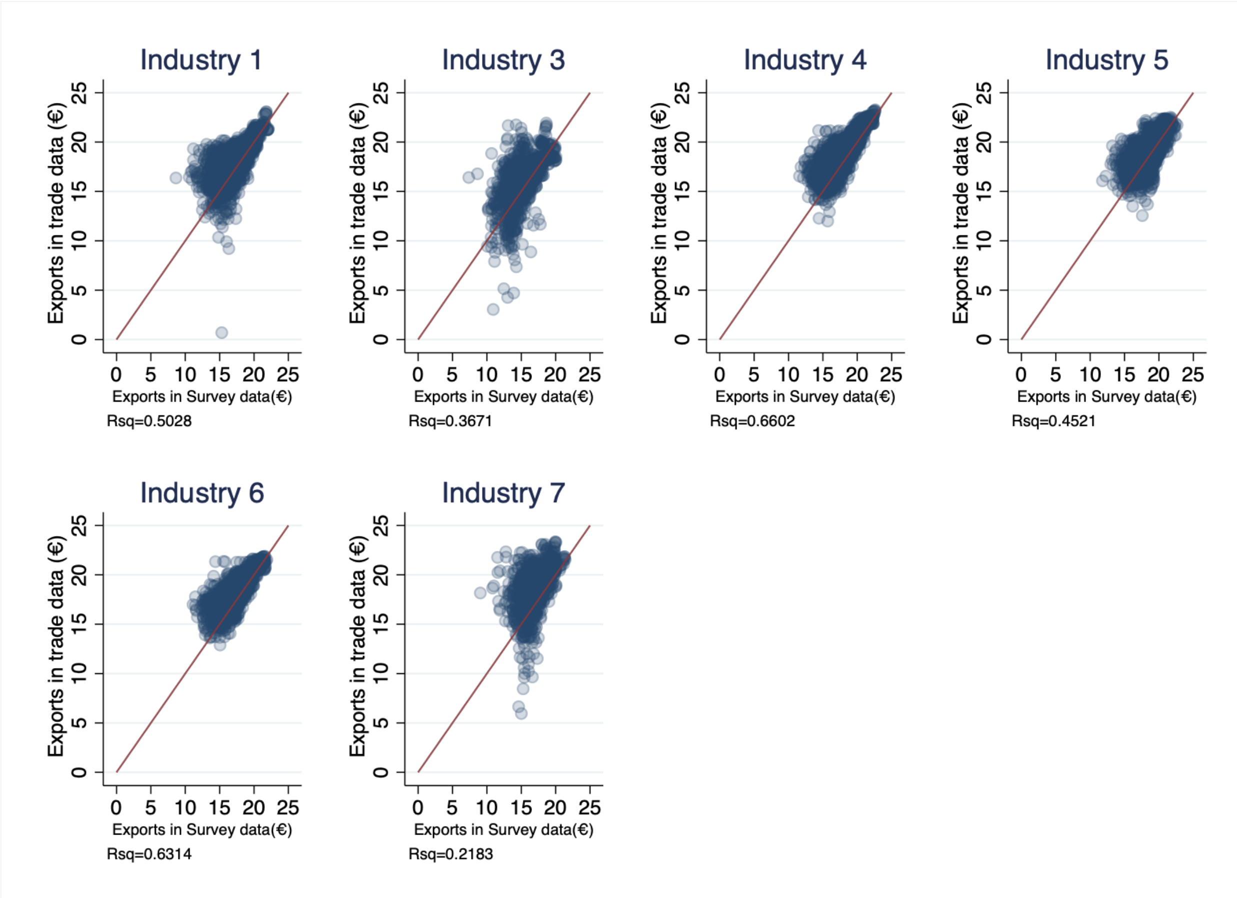
Notes: This figure shows the correlation between the price per kg of exports in international trade data and the imputed prices in our sample. The Y-axis reports the (log) price per kg of exports by country-pair, industry and year. The X-axis reports the estimated (log) price per kg of shipment in our regional trade dataset aggregated at the country-pair-industry-year level. In this figure we use all countries except France, Germany, Spain and United Kingdom.

Figure A.6: Correlation with aggregate international trade data



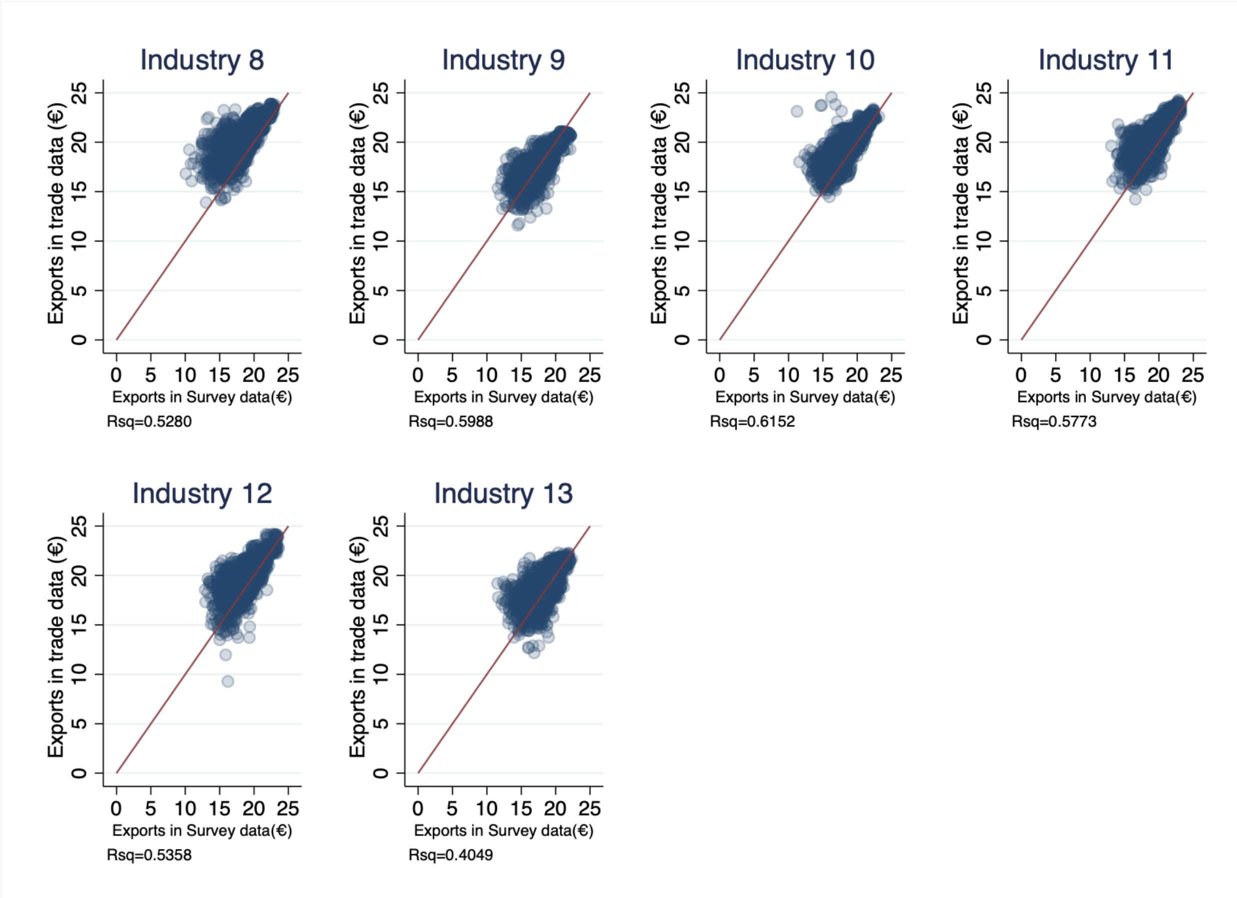
Notes: The figures show the correlation between exports and shipments in the ERFT survey in euros in each year. The Y-axis represents (log) bilateral trade (euros) by country-pair-industry-year using international trade data from Eurostat. The X-axis shows bilateral shipments by road (euros) aggregated by country-pair-industry-year obtained from the ERFT survey.

Figure A.7: Correlation with aggregate international trade data



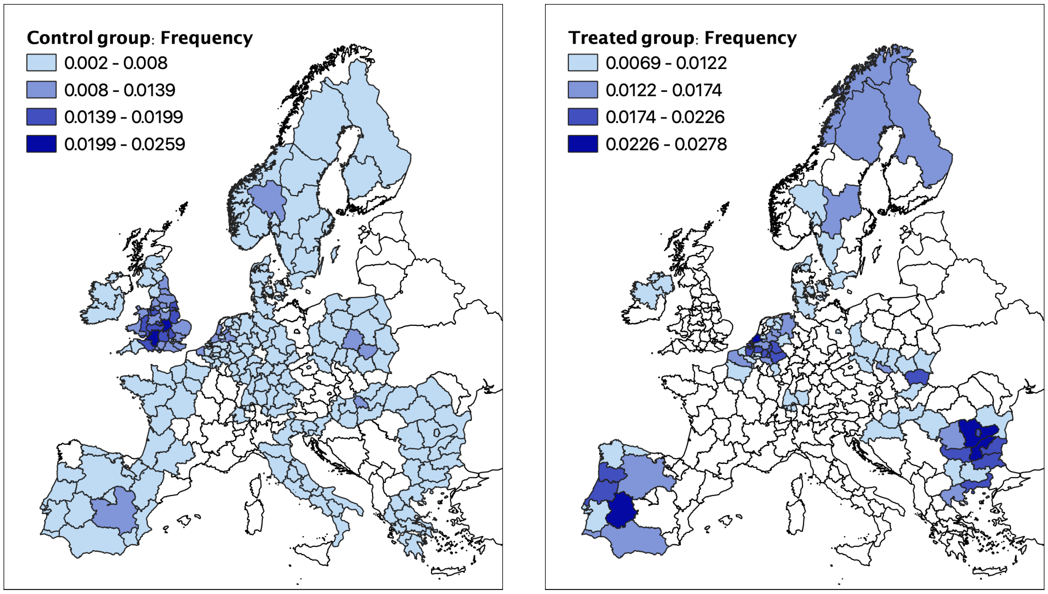
Notes: The figures show the correlation between exports and shipments in the ERFT survey in euros in each industry. The Y-axis represents (log) bilateral trade (euros) by country-pair-industry-year using international trade data from Eurostat. The X-axis shows bilateral shipments by road (euros) aggregated by country-pair-industry-year obtained from the ERFT survey.

Figure A.8: Correlation with aggregate international trade data



Notes: The figures show the correlation between exports and shipments in the ERFT survey in euros in each industry. The Y-axis represents (log) bilateral trade (euros) by country-pair-industry-year using international trade data from Eurostat. The X-axis shows bilateral shipments by road (euros) aggregated by country-pair-industry-year obtained from the ERFT survey.

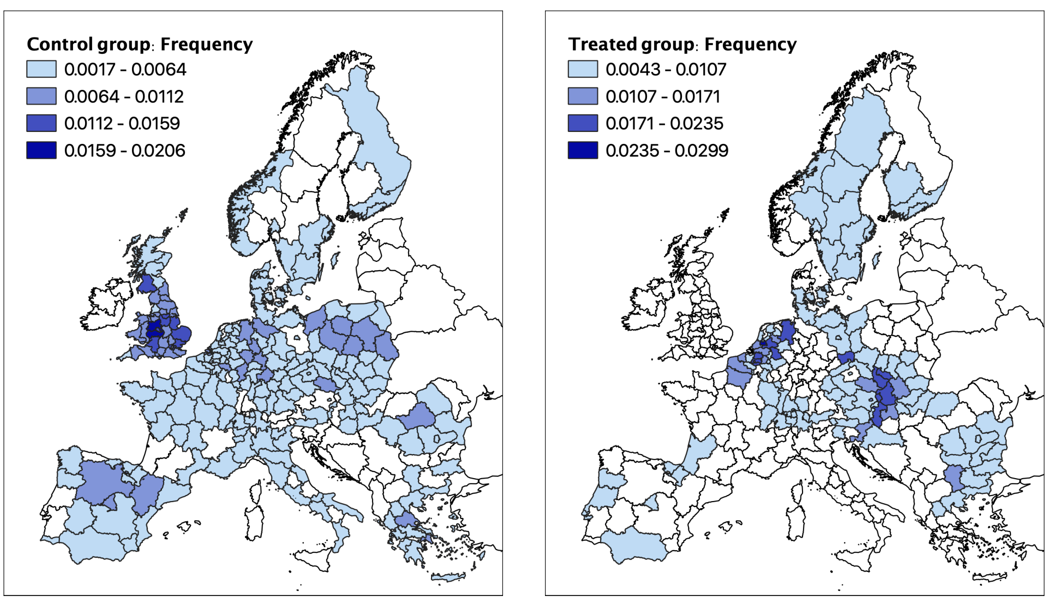
Figure A.9: Composition of regions in block 1



A) Control group **B) Treated group**

Notes: This figure shows the regions that are part of the block. The shading represents the frequency with which each region appears (as a part of a pair) in the control group (first panel) and treated group (second panel) in the block.

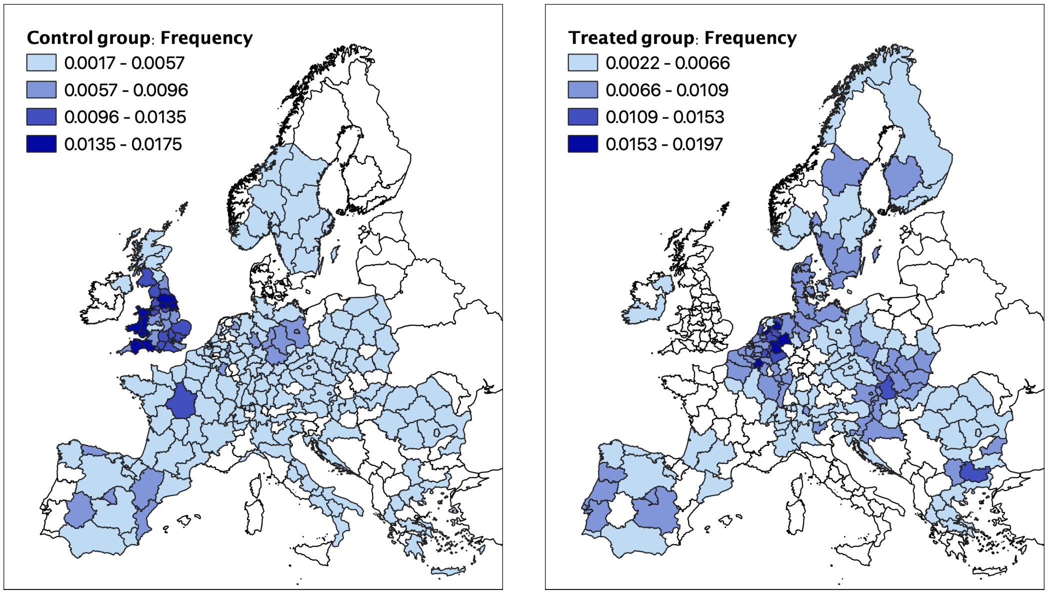
Figure A.10: Composition of regions in block 2



A) Control group **B) Treated group**

Notes: This figure shows the regions that are part of the block. The shading represents the frequency with which each region appears (as a part of a pair) in the control group (first panel) and treated group (second panel) in the block.

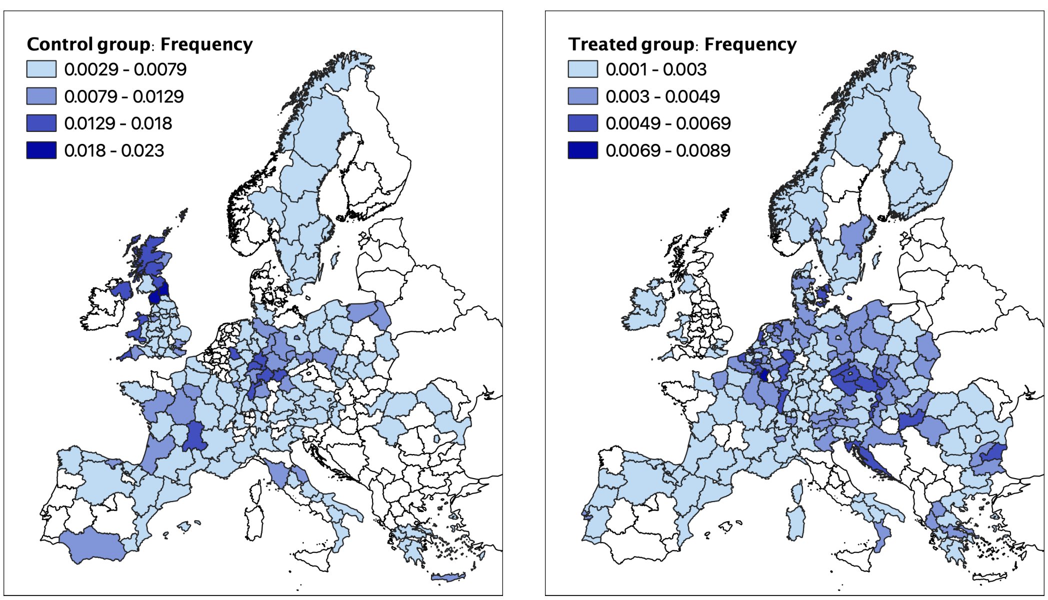
Figure A.11: Composition of regions in block 3



A) Control group **B) Treated group**

Notes: This figure shows the regions that are part of the block. The shading represents the frequency with which each region appears (as a part of a pair) in the control group (first panel) and treated group (second panel) in the block.

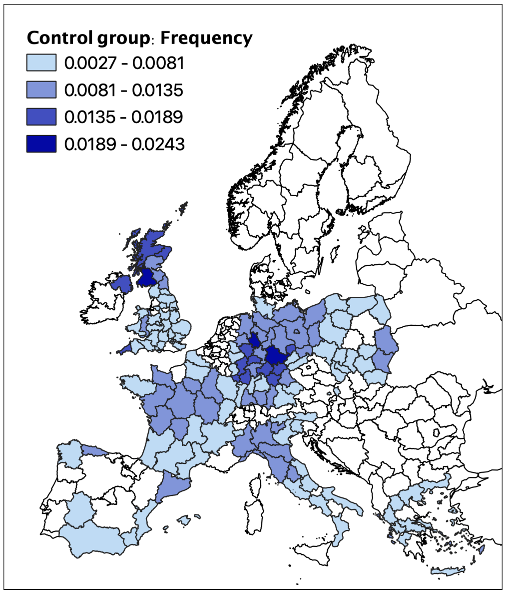
Figure A.12: Composition of regions in block 5



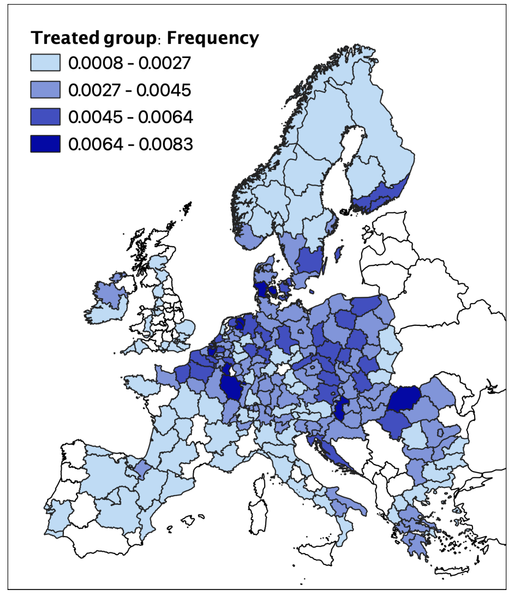
A) Control group **B) Treated group**

Notes: This figure shows the regions that are part of the block. The shading represents the frequency with which each region appears (as a part of a pair) in the control group (first panel) and treated group (second panel) in the block.

Figure A.13: Composition of regions in block 6



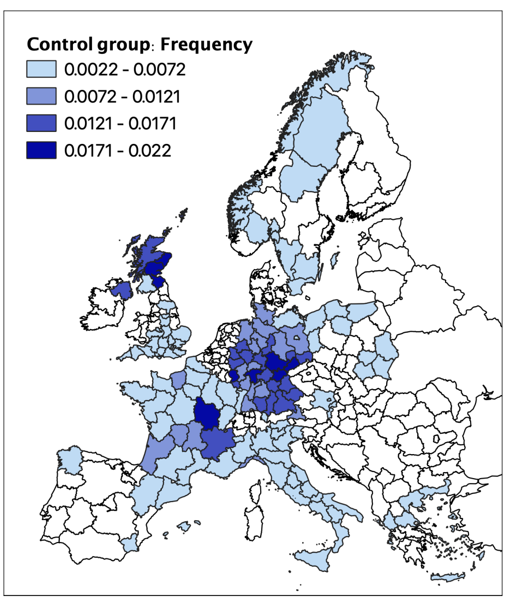
A) Control group



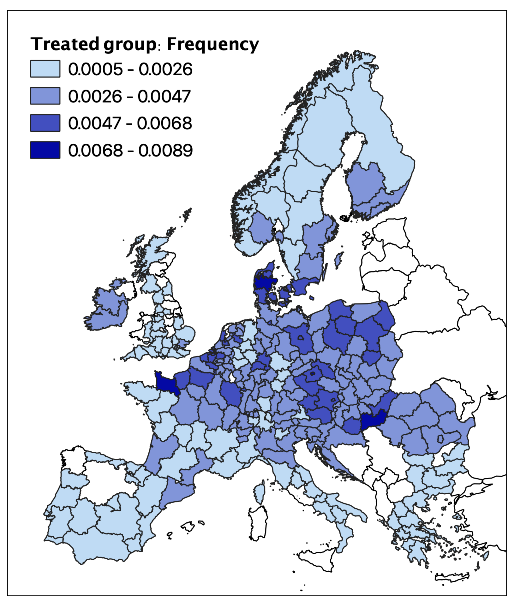
B) Treated group

Notes: This figure shows the regions that are part of the block. The shading represents the frequency with which each region appears (as a part of a pair) in the control group (first panel) and treated group (second panel) in the block.

Figure A.14: Composition of regions in block 7



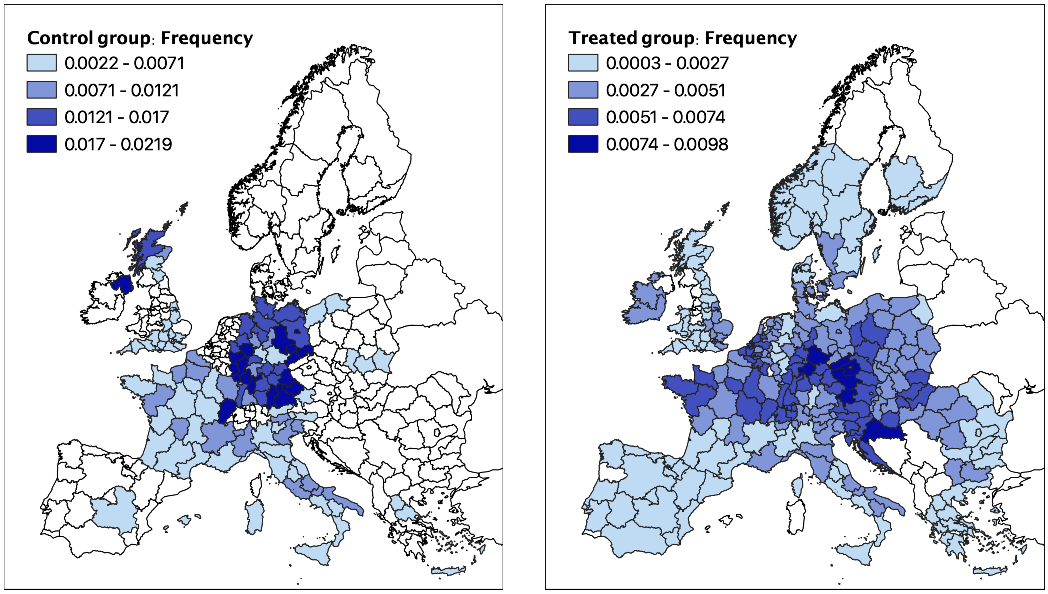
A) Control group



B) Treated group

Notes: This figure shows the regions that are part of the block. The shading represents the frequency with which each region appears (as a part of a pair) in the control group (first panel) and treated group (second panel) in the block.

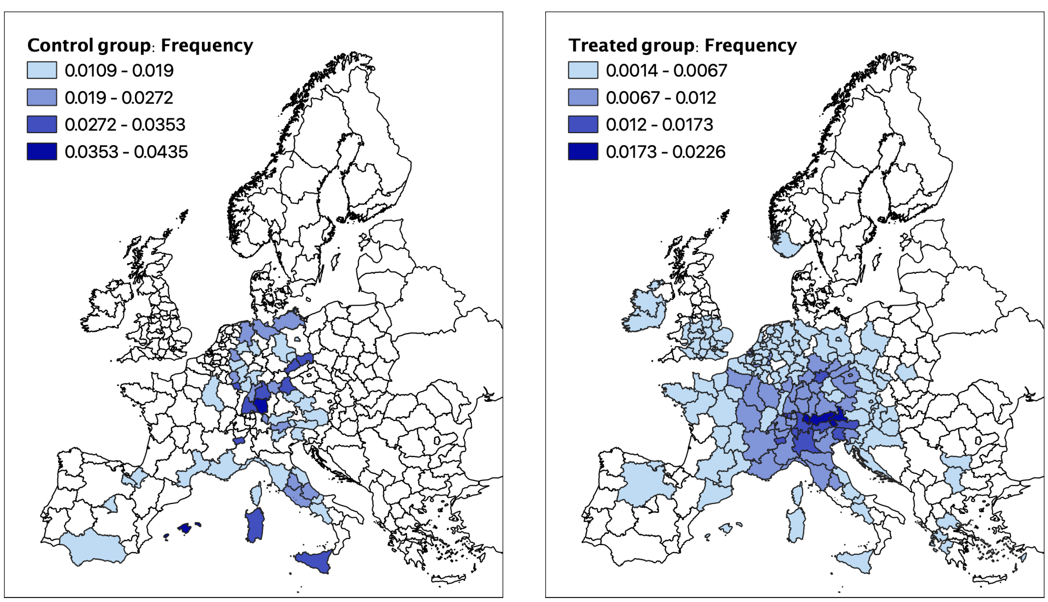
Figure A.15: Composition of regions in block 8



A) Control group **B) Treated group**

Notes: This figure shows the regions that are part of the block. The shading represents the frequency with which each region appears (as a part of a pair) in the control group (first panel) and treated group (second panel) in the block.

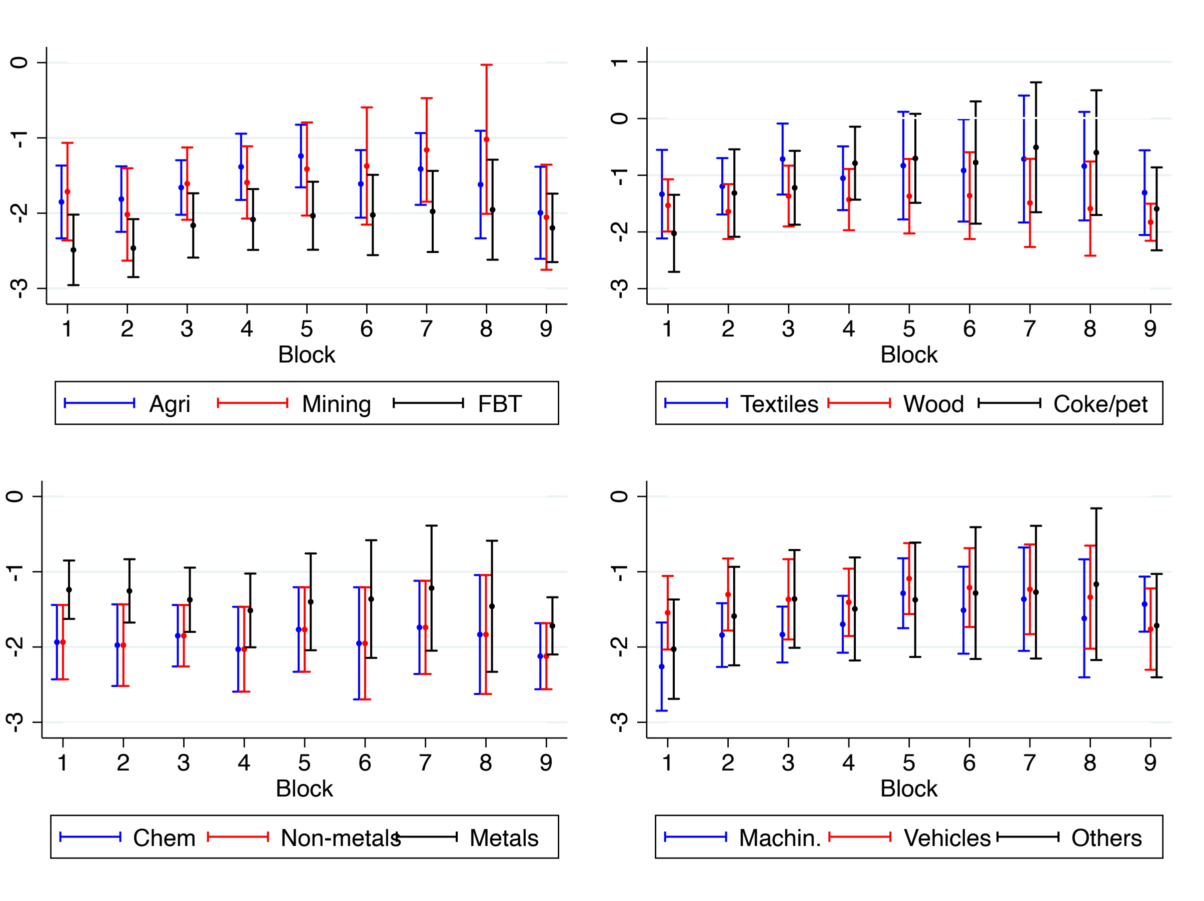
Figure A.16: Composition of regions in block 9



A) Control group **B) Treated group**

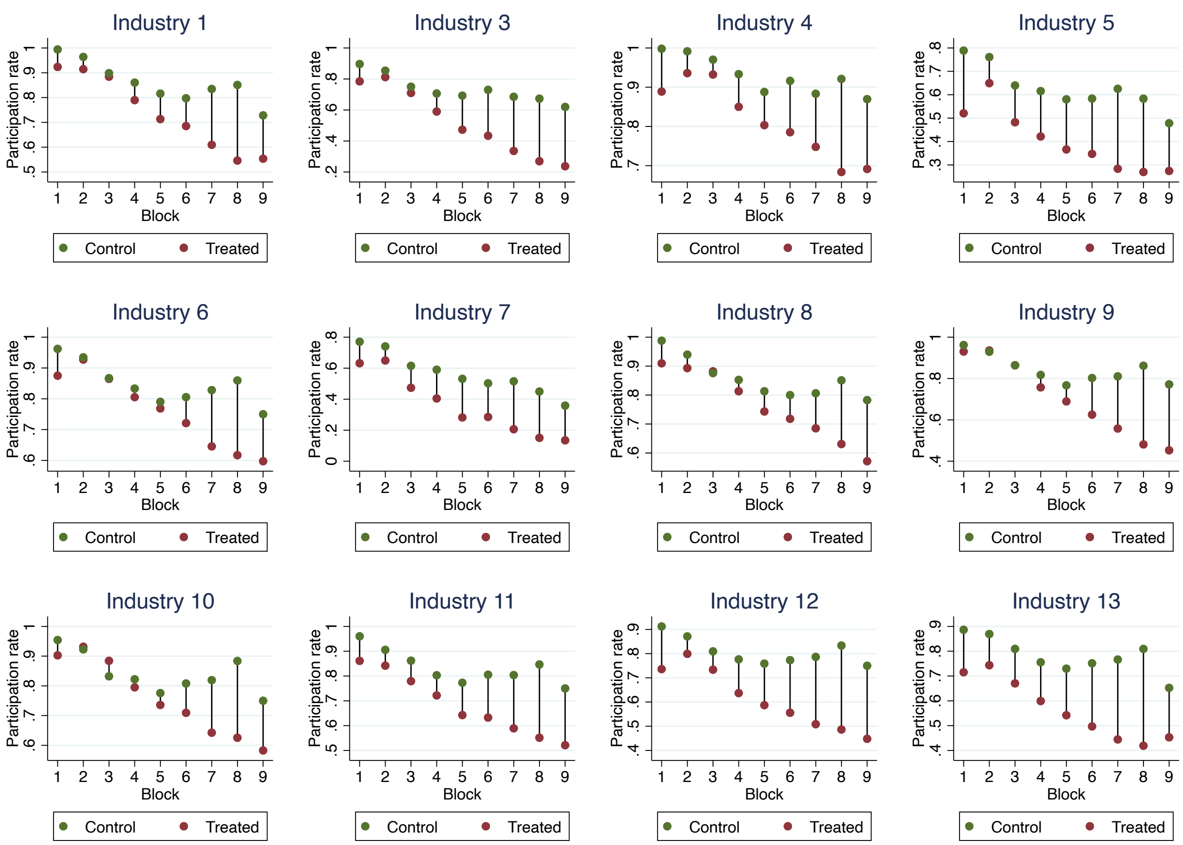
Notes: This figure shows the regions that are part of the block. The shading represents the frequency with which each region appears (as a part of a pair) in the control group (first panel) and treated group (second panel) in the block.

Figure A.17: Border effect - Industry level



Notes: These figures show the coefficient of the dummy Border estimated with specification (11) in each block and industry (dot). The confidence interval for the coefficient is represented by the vertical lines.

Figure A.18: Participation rates across industries



Notes: These figures show the participation rate (share of region pairs that display positive trade) in the control group, red circles, and in the treated group, green circles.

B Additional Tables

Table B.1: Industries in ERFT survey

Industry	Label	Sample
1	Products of agriculture, hunting, and forestry; fish and other fishing products	1
2	Coal and lignite; crude petroleum and natural gas	0
3	Metal ores and other mining and quarrying products; peat; uranium and thorium ores	1
4	Food products, beverages and tobacco	1
5	Textiles and textile products; leather and leather products	1
6	Wood and products of wood and cork (except furniture);pulp, paper and paper; printed matter and recorded media	1
7	Coke and refined petroleum products	1
8	Chemicals, chemical products, and man-made fibers; rubber and plasticproducts; nuclear fuel	1
9	Other non-metallic mineral products	1
10	Basic metals; fabricated metal products, except machinery and equipment	1
11	Machinery and equipment n.e.c.; communication equipment; medical, precision and optical instruments; watches and clocks	1
12	Transport equipment	1
13	Furniture; other manufactured goods n.e.c.	1
14	Secondary raw materials; municipal wastes and other wastes	0
15	Mail, parcels	0
16	Equipment and material utilized inthe transport of goods	0
17	Goods moved in the course of household and office removals	0
18	Grouped goods	0
19	Unidentifiable goods: goods which for any reason cannot be identified and therefore cannot be assigned to groups 01-16	0
20	Other goods n.e.c.	0

Table B.2: Price regressions

	DEP.VAR: Log Price	DEP.VAR: Log Price
	(1)	(2)
$\log(\text{Dist})_{o,d}$		0.451***
		(0.012)
Constant	10.550***	0.728
	(1.184)	(1.194)
Industry-Year FE	Yes	Yes
Origin Variables	Yes	Yes
Destination Variables	Yes	Yes
Obs.	48995	48995
R-squared	0.525	0.539

Notes: First column displays the results including only origin and destination level variables. The second column reports the results when adding the bilateral distance between origin and destination as a determinant of export prices.

Table B.3: Average border effect - Complete table

Dep. Var: $ln(S_{n,m})$	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8	Block 9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Border	-1.786***	-1.721***	-1.699***	-1.768***	-1.686***	-1.796***	-1.687***	-1.754***	-1.858***
	(0.182)	(0.178)	(0.175)	(0.175)	(0.238)	(0.289)	(0.268)	(0.290)	(0.201)
Distance	-0.899**	-1.378***	-1.643***	-0.618*	-1.949**	-0.532	-1.105**	-1.066***	-1.118
	(0.440)	(0.276)	(0.377)	(0.315)	(0.828)	(0.696)	(0.497)	(0.372)	(1.873)
Insularity	1.120	-0.861**	-0.157	-0.491	-1.777***	-0.913**	-1.596***	-1.554***	-1.024
	(0.754)	(0.376)	(0.430)	(0.412)	(0.534)	(0.418)	(0.351)	(0.319)	(0.862)
Mountain Ranges	0.014	-0.137*	-0.180**	-0.134	-0.322*	-0.088	-0.229**	-0.257***	-0.095
	(0.074)	(0.071)	(0.080)	(0.082)	(0.175)	(0.102)	(0.089)	(0.097)	(0.243)
River Basin	0.220	0.141	0.132	0.477***	0.155	0.514***	0.413**	0.348**	0.594
	(0.182)	(0.123)	(0.168)	(0.166)	(0.203)	(0.181)	(0.192)	(0.174)	(0.458)
Remoteness	2.236***	3.236***	3.339***	1.335**	3.412**	0.803	2.086**	2.167**	1.356
	(0.625)	(0.783)	(0.595)	(0.606)	(1.557)	(1.219)	(0.889)	(0.833)	(2.833)
Number of Borders	7.058***	6.695***	7.041***	10.779***	11.294***	11.833***	9.234***	8.091***	0.420
	(1.756)	(1.970)	(2.034)	(1.730)	(2.064)	(2.783)	(2.792)	(3.063)	(2.944)
Constant	-52.432***	-53.962***	-55.214***	-70.492***	-79.496***	-74.367***	-63.052***	-56.456***	-4.131
	(11.534)	(12.696)	(12.979)	(10.102)	(12.606)	(15.500)	(15.239)	(16.468)	(20.347)
N	645	813	1024	1364	968	1267	2011	2948	637
R^2	.572	.533	.501	.47	.375	.388	.31	.285	.299

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable is the (log) normalized market share of n in m. *Border* is a dummy for international border. *Number of Borders* is the average of the share of international borders that are faced by n and m. *Distance* is (log) bilateral distance between origin and destination in kilometers, *Insularity* takes value 1 if one of the regions is an island. *Mountain Ranges* is the highest difference in elevation between two regions in metres (difference between highest point and lowest point, in logs), *River Basin* takes value 1 if the region pair shares a river basin and *Remoteness* is the log of the average remoteness of the origin and the destination regions.

Table B.4: Average border effect - Dropping non-contiguous countries

Dep. Var: $ln(S_{n,m})$	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8	Block 9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Border	-1.776***	-1.723***	-1.677***	-1.754***	-1.615***	-1.735***	-1.695***	-1.743***	-1.725***
	(0.184)	(0.179)	(0.180)	(0.177)	(0.241)	(0.309)	(0.305)	(0.336)	(0.257)
Distance	-0.883*	-1.389***	-1.581***	-0.492	-2.118**	-0.770	-0.668	-1.162**	-1.953
	(0.439)	(0.278)	(0.378)	(0.337)	(0.858)	(0.813)	(0.552)	(0.430)	(2.862)
Insularity	1.122	-0.865**	-0.138	-0.495	-1.595***	-0.658	-1.041**	-1.029***	-0.523
	(0.752)	(0.375)	(0.439)	(0.485)	(0.544)	(0.497)	(0.414)	(0.298)	(1.341)
Mountain Ranges	0.015	-0.138*	-0.175**	-0.129	-0.393**	-0.201*	-0.274**	-0.458***	-0.514
	(0.074)	(0.071)	(0.080)	(0.085)	(0.193)	(0.115)	(0.106)	(0.106)	(0.392)
River Basin	0.219	0.140	0.153	0.481***	0.102	0.447**	0.502**	0.250	0.387
	(0.182)	(0.123)	(0.170)	(0.167)	(0.202)	(0.198)	(0.212)	(0.188)	(0.695)
Remoteness	2.205***	3.254***	3.253***	1.142*	3.750**	1.400	1.353	2.586**	2.323
	(0.621)	(0.781)	(0.607)	(0.657)	(1.622)	(1.422)	(1.119)	(1.047)	(4.534)
Number of Borders	7.090***	6.682***	7.205***	11.183***	11.693***	12.266***	10.665***	9.913**	2.894
	(1.764)	(1.972)	(2.084)	(1.744)	(2.123)	(3.139)	(3.528)	(4.027)	(5.014)
Constant	-52.497***	-53.949***	-55.993***	-72.388***	-82.934***	-79.117***	-69.208***	-68.731***	-18.153
	(11.553)	(12.716)	(13.244)	(10.113)	(12.587)	(17.205)	(18.771)	(22.167)	(34.654)
N	639	811	984	1270	833	1000	1451	1970	395
R^2	.566	.533	.494	.48	.376	.395	.316	.286	.359

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable is the (log) normalized market share of n in m. *Border* is a dummy for international border. *Distance* is (log) bilateral distance between origin and destination in kilometers, *Insularity* takes value 1 if one of the regions is an island. *Mountain Ranges* is the highest difference in elevation between two regions in metres (difference between highest point and lowest point, in logs), *River Basin* takes value 1 if the region pair shares a river basin and *Remoteness* is the log of the average remoteness of the origin and the destination regions.

Table B.5: Average border effect - No number of borders

Dep. Var: $ln(S_{n,m})$	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8	Block 9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Border	-1.478***	-1.382***	-1.336***	-1.166***	-1.092***	-1.162***	-1.215***	-1.350***	-1.841***
	(0.193)	(0.208)	(0.174)	(0.202)	(0.210)	(0.216)	(0.189)	(0.199)	(0.209)
Distance	-0.807*	-1.508***	-1.917***	-0.660*	-2.161**	-0.843	-1.289**	-1.190***	-1.098
	(0.458)	(0.327)	(0.439)	(0.349)	(0.898)	(0.771)	(0.523)	(0.361)	(1.843)
Insularity	1.298**	-0.833**	-0.259	-0.653	-2.202***	-1.419***	-1.969***	-1.836***	-1.020
	(0.520)	(0.379)	(0.380)	(0.417)	(0.588)	(0.490)	(0.398)	(0.303)	(0.853)
Mountain Ranges	-0.002	-0.110	-0.137	-0.079	-0.269	-0.079	-0.211**	-0.242**	-0.090
	(0.091)	(0.089)	(0.091)	(0.106)	(0.193)	(0.106)	(0.093)	(0.098)	(0.232)
River Basin	0.471**	0.300*	0.212	0.732***	0.315	0.635***	0.470**	0.409**	0.606
	(0.219)	(0.171)	(0.212)	(0.232)	(0.257)	(0.210)	(0.198)	(0.194)	(0.431)
Remoteness	2.795***	3.886***	4.087***	2.213***	4.674***	2.384*	3.327***	3.228***	1.368
	(0.735)	(1.005)	(0.797)	(0.776)	(1.691)	(1.211)	(0.913)	(0.720)	(2.838)
Constant	-13.122***	-16.590***	-15.672***	-10.163***	-17.628***	-10.528***	-13.645***	-13.189***	-1.784
	(3.456)	(5.477)	(3.780)	(3.555)	(5.732)	(3.785)	(3.384)	(3.466)	(8.180)
N	645	813	1024	1364	968	1267	2011	2948	637
R^2	.499	.473	.454	.384	.302	.314	.276	.262	.299

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable is the (log) normalized market share of n in m. *Border* is a dummy for international border. *Distance* is (log) bilateral distance between origin and destination in kilometers, *Insularity* takes value 1 if one of the regions is an island. *Mountain Ranges* is the highest difference in elevation between two regions in metres (difference between highest point and lowest point, in logs), *River Basin* takes value 1 if the region pair shares a river basin and *Remoteness* is the log of the average remoteness of the origin and the destination regions.

Table B.6: Summary statistics of covariates by block: Conditional on Border in 1910=0

	(1) mean/sd	(2) mean/sd	(3) mean/sd	(4) mean/sd	(5) mean/sd	(6) mean/sd
Distance	271.968	371.468	455.596	535.897	657.689	665.057
	97.15	113.65	131.67	170.34	217.44	274.37
Insularity	0.072	0.098	0.086	0.040	0.050	0.056
	0.26	0.30	0.28	0.20	0.22	0.23
Mountain Ranges	5.532	5.873	6.114	6.157	6.344	6.552
	0.98	0.90	0.81	0.78	0.81	0.80
River Basin	0.203	0.170	0.174	0.229	0.275	0.689
	0.40	0.38	0.38	0.42	0.45	0.46
Remoteness	1067.843	1028.199	1003.654	992.166	1001.298	1036.640
	235.34	206.65	183.78	164.26	142.07	144.14
Estimated propensity score	0.170	0.311	0.439	0.559	0.685	0.814
	0.04	0.04	0.04	0.04	0.04	0.04
N	775	552	466	375	298	161

Notes: This table reports the mean and standard deviation of each geographical covariate and the propensity score in each block. *Distance* is bilateral distance between origin and destination in kilometers, *Insularity* takes value 1 if one of the regions is an island. *Mountain Ranges* is the highest difference in elevation between two regions in metres (difference between highest point and lowest point), *River Basin* takes value 1 if the region pair shares a river basin and *Remoteness* is the average remoteness of the origin and the destination regions.

Table B.7: Balancing test of covariates by block: Conditional on Border in 1910=0

	(1)	(2)	(3)	(4)	(5)	(6)
Distance	0.0219	-0.0256	0.0362	-0.0197	-0.0510	-0.141
	(0.0388)	(0.0339)	(0.0321)	(0.0382)	(0.0513)	(0.0779)
Insularity	-0.0862	-0.114	0.000715	0.0277	0.211	0.333
	(0.0240)	(0.0273)	(0.0263)	(0.0202)	(0.0272)	(0.0410)
Mountain Ranges	-0.353	-0.227	-0.0558	0.162	0.701	0.692
	(0.0910)	(0.0836)	(0.0755)	(0.0806)	(0.103)	(0.160)
River Basin	-0.00557	0.0327	-0.0474	0.0168	0.0455	-0.0274
	(0.0376)	(0.0351)	(0.0355)	(0.0435)	(0.0609)	(0.0982)
Remoteness	0.0349	-0.0208	-0.00301	0.00436	0.0235	-0.104
	(0.0188)	(0.0172)	(0.0154)	(0.0150)	(0.0170)	(0.0252)
N	775	552	466	375	298	161

Notes: This table reports the difference in means between treated and control region pairs for each geographical covariate by block (defined as control minus treated). Standard errors in parenthesis, significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.. *Distance* is bilateral distance between origin and destination in kilometers, *Insularity* takes value 1 if one of the regions is an island. *Mountain Ranges* is the highest difference in elevation between two regions in metres (difference between highest point and lowest point), *River Basin* takes value 1 if the region pair shares a river basin and *Remoteness* is the average remoteness of the origin and the destination regions.

Table B.8: Border effect on the probability of having a common language

Dep. Var.: Language_{nm}	Block1 (1)	Block2 (2)	Block3 (3)	Block4 (4)	Block5 (5)	Block6 (6)	Block7 (7)	Block8 (8)	Block9 (9)
Border	-0.618***	-0.573***	-0.671***	-0.709***	-0.806***	-0.747***	-0.744***	-0.720***	-0.756***
	(0.0933)	(0.0997)	(0.0763)	(0.0568)	(0.0706)	(0.0625)	(0.0719)	(0.0841)	(0.148)
Distance (Geo)	0.0920	-0.184	-0.213	-0.550***	-0.0665	-0.205	-0.325**	-0.313**	-0.0226
	(0.103)	(0.166)	(0.174)	(0.188)	(0.395)	(0.300)	(0.153)	(0.128)	(0.665)
Insularity	0.435*	0.00897	0.136	-0.192	0.110	0.0343	-0.0629	-0.0102	0.174
	(0.252)	(0.104)	(0.126)	(0.129)	(0.247)	(0.267)	(0.175)	(0.134)	(0.305)
Mountain Ranges	0.0300	-0.0398	-0.0684	-0.0818	0.0399	0.0428	0.0190	0.0755**	0.123
	(0.0189)	(0.0371)	(0.0538)	(0.0552)	(0.0776)	(0.0479)	(0.0300)	(0.0369)	(0.0991)
River Basin	-0.0402	-0.0166	-0.0889	-0.0291	0.0119	0.102	0.0761	0.0719	0.181
	(0.0560)	(0.0740)	(0.0949)	(0.0901)	(0.139)	(0.0874)	(0.0623)	(0.0748)	(0.276)
Remoteness	-0.621**	0.236	0.112	0.630*	-0.224	0.0502	0.397	-0.00337	-1.351
	(0.262)	(0.331)	(0.331)	(0.343)	(0.687)	(0.537)	(0.337)	(0.351)	(1.199)
No. Borders	0.571	-0.220	0.444	0.297	-0.287	-0.952	-1.512*	-1.351	-0.795
	(0.532)	(0.728)	(1.053)	(1.059)	(1.075)	(0.772)	(0.858)	(0.926)	(2.268)
Constant	1.232	1.861	-0.995	-1.655	4.484	7.461*	9.413**	10.77*	14.43
	(2.461)	(3.670)	(5.911)	(5.969)	(7.372)	(4.457)	(4.593)	(5.641)	(14.87)
N	645	813	1024	1364	968	1267	2011	2948	637
R^2	.624	.469	.513	.542	.611	.572	.556	.45	.36

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable, $Language_{n,m}$ is a language dummy, that takes value 1 if regions n and m share a language. *Border* is a dummy for international border. All regressions include as control the *Number of borders*, the (log) sum of the number of borders that are faced by n and m.

Table B.9: Border effect on the share of common languages

Dep. Var.: Common Language Share_{nm}	Block1 (1)	Block2 (2)	Block3 (3)	Block4 (4)	Block5 (5)	Block6 (6)	Block7 (7)	Block8 (8)	Block9 (9)
Border	-0.668***	-0.561***	-0.664***	-0.662***	-0.753***	-0.768***	-0.740***	-0.719***	-0.676***
	(0.0873)	(0.0965)	(0.0733)	(0.0683)	(0.0625)	(0.0557)	(0.0664)	(0.0813)	(0.118)
Distance (Geo)	0.0332	-0.123	-0.131	-0.431***	-0.102	-0.420*	-0.281**	-0.168**	0.479
	(0.0943)	(0.146)	(0.118)	(0.155)	(0.316)	(0.238)	(0.136)	(0.0786)	(0.629)
Insularity	0.223**	0.151	0.190***	-0.0230	0.0566	-0.129	-0.0953	-0.0168	0.269
	(0.0829)	(0.0978)	(0.0678)	(0.112)	(0.193)	(0.171)	(0.119)	(0.0746)	(0.330)
Mountain Ranges	0.0213	-0.0548	-0.0849*	-0.0780*	0.00902	-0.0141	0.00500	0.0531*	0.111
	(0.0170)	(0.0421)	(0.0456)	(0.0450)	(0.0606)	(0.0337)	(0.0252)	(0.0273)	(0.113)
River Basin	-0.0922	-0.0619	-0.0653	-0.0448	-0.00528	0.0247	0.0545	0.102	0.278
	(0.0617)	(0.0859)	(0.0729)	(0.0629)	(0.0950)	(0.0514)	(0.0506)	(0.0753)	(0.222)
Remoteness	-0.367	0.236	0.145	0.474*	-0.102	0.473	0.340	-0.105	-1.888
	(0.260)	(0.291)	(0.251)	(0.284)	(0.544)	(0.399)	(0.267)	(0.247)	(1.266)
No. Borders	-0.162	-1.289	-0.425	-0.764	-0.665	-1.035*	-1.542**	-1.454*	-1.355
	(0.603)	(0.823)	(0.934)	(0.923)	(0.728)	(0.540)	(0.609)	(0.794)	(1.822)
Constant	4.288	8.163*	3.677	5.184	6.235	6.594**	9.749***	11.28**	18.41
	(2.728)	(4.449)	(5.151)	(5.072)	(5.074)	(3.062)	(3.570)	(5.098)	(13.45)
N	645	813	1024	1364	968	1267	2011	2948	637
R^2	.663	.556	.646	.66	.733	.775	.716	.61	.453

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable, $CommonLanguageShare_{n,m}$ is a language variable that measures the share of common languages between regions n and m. *Border* is a dummy for international border.

Table B.10: Border effect on distance traveled

Dep. Var.: Distance Traveled$_{nm}$	Block1 (1)	Block2 (2)	Block3 (3)	Block4 (4)	Block5 (5)	Block6 (6)	Block7 (7)	Block8 (8)	Block9 (9)
Border Effect	0.158**	0.0988***	0.101***	0.0717***	0.0612*	0.0168	0.0478	0.0398	0.0545**
	(0.0604)	(0.0292)	(0.0329)	(0.0206)	(0.0347)	(0.0429)	(0.0296)	(0.0344)	(0.0265)
Distance (Geo)	0.710***	0.952***	1.015***	1.039***	0.720***	0.987***	0.890***	0.865***	0.521**
	(0.122)	(0.104)	(0.110)	(0.0686)	(0.176)	(0.131)	(0.0917)	(0.0657)	(0.244)
Mountain Ranges	0.00368	0.0298	0.0594***	0.0656***	0.0168	0.0343	0.0437**	0.0472***	0.0523*
	(0.0205)	(0.0180)	(0.0196)	(0.0147)	(0.0295)	(0.0212)	(0.0174)	(0.0170)	(0.0268)
River Basin	-0.0830	-0.0763**	-0.0600	-0.0508**	-0.0865*	-0.0391	-0.0440	-0.0422**	-0.138*
	(0.0531)	(0.0310)	(0.0407)	(0.0223)	(0.0436)	(0.0372)	(0.0299)	(0.0198)	(0.0804)
Remoteness	0.307	-0.0811	-0.163	-0.251*	0.226	-0.295	-0.160	-0.0562	0.497
	(0.227)	(0.176)	(0.195)	(0.144)	(0.305)	(0.290)	(0.210)	(0.182)	(0.380)
No. Borders	0.536	0.563**	0.309	0.212	0.323	0.638	0.512	0.506	-0.401
	(0.377)	(0.251)	(0.310)	(0.223)	(0.398)	(0.548)	(0.470)	(0.484)	(0.561)
Constant	-3.795	-2.575	-0.941	0.108	-1.780	-1.756	-1.409	-1.955	1.982
	(2.519)	(1.569)	(1.983)	(1.435)	(2.370)	(2.650)	(2.493)	(2.289)	(3.381)
N	641	803	1010	1340	932	1190	1858	2606	505
R^2	.733	.822	.803	.838	.76	.711	.724	.584	.372

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable, Shipping distance$_{n,m}$ is a average distance travelled by a truck taking a shipment from n to m. *Border* is a dummy for international border.

Table B.11: Border effect on disagreement in values

D. V.: Disagreement$_{nm}$	Block1	Block2	Block3	Block4	Block5	Block6	Block7	Block8	Block9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Border	0.489	0.671**	0.637**	0.469	1.102***	0.870***	0.796***	0.348**	0.188
	(0.364)	(0.283)	(0.292)	(0.429)	(0.303)	(0.279)	(0.251)	(0.172)	(0.342)
Distance	-0.191	-0.821**	-0.187	0.049	-1.902	0.478	0.177	-0.359	2.159
	(0.341)	(0.346)	(0.764)	(0.568)	(1.467)	(0.677)	(0.570)	(0.613)	(2.321)
Insularity	0.000	-0.353	-0.553	-0.026	-0.753	0.651	0.345	-0.507	0.074
	(.)	(0.368)	(0.420)	(0.535)	(0.866)	(0.580)	(0.479)	(0.333)	(1.143)
Mountain Ranges	0.319**	0.136	0.266*	0.183	-0.112	0.021	0.065	-0.071	0.241
	(0.139)	(0.134)	(0.152)	(0.157)	(0.230)	(0.183)	(0.129)	(0.114)	(0.310)
River Basin	0.167	0.096	0.007	-0.095	-0.187	0.132	0.337	-0.075	0.408
	(0.217)	(0.196)	(0.233)	(0.232)	(0.338)	(0.229)	(0.229)	(0.247)	(0.644)
Remoteness	1.266**	3.035***	2.016	1.327	5.017*	0.632	0.513	1.856**	-1.581
	(0.610)	(0.693)	(1.319)	(1.057)	(2.636)	(1.368)	(0.964)	(0.922)	(3.916)
Number of Borders	-3.614	-4.662***	-4.822**	-4.850	-12.546***	-10.866***	-11.371***	-3.865	-4.152
	(2.204)	(1.704)	(2.044)	(3.104)	(2.754)	(3.268)	(3.448)	(2.975)	(9.155)
Constant	14.488	12.850	17.058	21.448	56.457***	62.021***	67.432***	16.010	24.023
	(15.188)	(11.161)	(14.693)	(21.324)	(14.927)	(21.405)	(21.939)	(19.312)	(58.130)
N	442	476	546	646	452	522	750	1010	216
R^2	.333	.315	.344	.192	.286	.215	.143	.0913	.135

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the country-pair level, are in parentheses. Dependent variable is the value disagreement between n and m. *Border* is a dummy for international border.

Table B.12: European Values Survey: Variables included in the Disagreement Index

Name	Description
a124_02	neighbours: people of a different race
a124_06	neighbours: immigrants/foreign workers
b008	protecting environment vs. economic growth
c002	jobs scarce: employers should give priority to (nation) people than immigrants
c002_01	jobs scarce: employers should give priority to (nation) people than immigrants
c038	people who don't work turn lazy
c041	work should come first even if it means less spare time
d001_b	trust: your family
g007_18_b	trust: your neighborhood
g007_33_b	trust: people you know personally
g007_34_b	trust: people you meet for the first time
g007_35_b	trust: people of another religion
g007_36_b	trust: people of another nationality
e035	income equality
e036	private vs state ownership of business
e039	competition good or harmful
e069_18	confidence: the european union
e224	democracy: governments tax the rich and subsidize the poor.
e225	democracy: religious authorities interpret the laws.
e226	democracy: people choose their leaders in free elections.
e227	democracy: people receive state aid for unemployment.
e228	democracy: the army takes over when government is incompetent.
e229	democracy: civil rights protect people's liberty against oppression.
e233	democracy: women have the same rights as men.
e233a	democracy: the state makes people's incomes equal
e233b	democracy: people obey their rulers
e235	importance of democracy
f114a	justifiable: claiming government benefits to which you are not entitled
f115	justifiable: avoiding a fare on public transport
f116	justifiable: cheating on taxes
f117	justifiable: someone accepting a bribe
g255	how close you feel: your village, town or city

C Construction of European regional trade dataset

in this section we explain the methodology we follow to construct the matrix of regional trade flows in Europe. First, we explain the data sets used for the price imputation procedure. Second, we provide additional details about how we clean and use the European Road Freight Transport dataset (ERFT).

C.1 Regional price data

The subsample of region to country level trade data is collected individually for our subset of four countries:

France The French Douane administration provides international trade data for the different Regions and Departements in France. The data is available quarterly for the years 2011 and 2014 at the industry level (4 digits of disaggregation of CPA4) for the different origin/destination countries. The trade flows are collected in value and weight, both imports and exports.[38] We use a 2 digits industrial disaggregation (22 industries).

Germany The German agency of statistics, Destatis, provides Foreign trade data for the 16 German states (Bundeslander). The data is available monthly for the years 2008 to November 2016 at the industry level (1, 2 or 3 digits of aggregation) for the different origin/destination countries. The trade flows are collected in value and weight (Tons). For this paper we use annual data for the years 2011 to 2014, at a 2 digits level of disaggregation (30 industries).[39]

Spain The Spanish secretary of commerce provides Foreign trade data for the 17 Spanish regions (Comunidades Autonomas). The data is available monthly for the years 1995 to 2015 at different industry levels for the different origin/destination countries. The trade flows are collected in value, not weight. For this paper we use annual data for the years 2011 to 2014, at a 2 digits level of disaggregation (22 industries).[40]

United Kingdom The UK Customs department provides Foreign trade data for the 12 regions in the UK. The data is available monthly for the years 2009 to 2016 at different industry levels (several digits available) for the different origin/destination countries. The trade flows are collected in value and weight. For this paper we use annual data for the years 2011 to 2014, at a 2 digit level of disaggregation.[41]

We aggregate each dataset to a 20 industry NST 2007 classification (European classification system for transport statistics), which is the classification used in the European Road Freight Transport Survey. This subsample of 58 regions allows us to observe 2,688 region to country trade flows (region-country pairs) each year.

[38]The data can be accessed at `http://lekiosque.finances.gouv.fr/portail_default.asp`.
[39]The data can be accessed at `https://www-genesis.destatis.de/genesis/online/data`.
[40]The data can be accessed at `http://datacomex.comercio.es/principal_comex_es.aspx`.
[41]The data can be access at: Statistical department of the United Kingdom government.

Table C.1: Foreign Trade Sample

Country	**Unit**	**Freq**	**Year**	**Industries**	**Unit**
Spain	NUTS2	Monthly	2011-2014	22, 99	€, kg
Germany	NUTS1	Monthly	2011-2014	30, 211	€, kg
France	NUTS3	Trimester	2011-2014	22 ,>200	€, kg
UK	NUTS1	Quarterly	2011-2014	67	£, kg

C.2 Variables for price imputation and robustness checks

We put together an extensive database of economic and geographic characteristics at the regional and country to use as determinants of price levels across regions. Our preferred specification is to pool all time periods and industries in the following regression:

$$\ln P_{nm}^{it} = \eta_n^t X_n^{t-1} + \pi_m^t Z_m^{t-1} + \beta d_{nm} + \phi^{it} + e_{nm}^{it},$$

where P_{nm}^{it} is the unit price of exports of industry i shipped from origin n to destination m in year t. The price of exports is calculated as the ratio between the value of exports and the weight of exports for each industry, origin, destination and year. Table C.2 reports the complete list of variables that we include as controls.[42] In addition, we also compute the geodesic distance between the centroid of the origin and the destination region, and we use it as a proxy for bilateral distance d_{nm}.

To test the accuracy of our predicted prices, we collect data of country to country trade flows at the year-industry level from Eurostat dataset COMEXT. Comext is Eurostat's reference database for detailed statistics on international trade in goods. It provides information about the value and quantity of the trade transaction, allowing us to compute the price per kilo of exports. We download the data for the years in our sample, 2011-2017, from the website: `http://epp.eurostat.ec.europa.eu/newxtweb/`.

C.3 European Road Freight Transport survey

The European Road Freight Transport survey microdata is a database collected by Eurostat in order to understand the magnitude of the shipment of goods across Europe. The ERFT survey covers 27 EU countries (except Malta) and EFTA countries (except Iceland). Each member state collects statistics on the carriage of goods by road by means of any road

[42] EuroRegional Map: https://eurogeographics.org/products-and-services/euroregionalmap/

Table C.2: Explanatory Variables for Price regressions

Label	Included	Level	Source
log(Pop Dens)	or/dest	NUTS2, year	Eurostat
log(GDP pc)	or/dest	NUTS2, year	Eurostat
log(Life Exp.)	or/dest	NUTS2, year	Eurostat
log(Total Emp.)	or/dest	NUTS2, year	Eurostat
Manuf. Sh. of Emp.	or/dest	NUTS2, year	Eurostat
Low Tech. Sh. of Emp.	or/dest	NUTS2, year	Eurostat
Edu (None) Sh.	or/dest	NUTS2, year 2011	2011 census
Edu (ISEC3) Sh.	or/dest	NUTS2, year 2011	2011 census
Edu (ISEC6) Sh.	or/dest	NUTS2, year 2011	2011 census
Ind Agri. Sh.	or/dest	NUTS2, year 2011	2011 census
Ind Manu. Sh.	or/dest	NUTS2, year 2011	2011 census
Ind. Prof/Science Sh.	or/dest	NUTS2, year 2011	2011 census
Ind. Fin. Sh.	or/dest	NUTS2, year 2011	2011 census
Ind. Pub. Sh	or/dest	NUTS2, year 2011	2011 census
Birth (Other EU) Sh.	or/dest	NUTS2, year 2011	2011 census
Birth (Non-EU) Sh.	or/dest	NUTS2, year 2011	2011 census
log(Heating h)	or/dest	NUTS2, year	Eurostat
log(av sun h)	or/dest	NUTS2	PVGIS 5 solar irradiation
log(max_sun h)	or/dest	NUTS2	PVGIS 5 solar irradiation
log(distRiver)	or/dest	NUTS2	EuroRegional map
log(distCoast)	or/dest	NUTS2	EuroRegional map

freight vehicle from a representative sample of road vehicles collected from the national vehicle registry. In case such a registry is not available, the sample will be selected either from the registry of licensed road haulage operators or the registry of persons licensed to operate such vehicles. In particular, Eurostat provides three interlinked datasets that contain the micro data at the vehicle, journey and goods level.

The Vehicle dataset (Dataset A1) records characteristics of each individual road vehicle and besides identifying each respondent vehicle contains information such as the age, axle configuration, unladen weight, total permissible weight and total kilometers performed during the survey.

The Journey dataset (Dataset A2) contains information about specific journeys performed by a vehicle identified in the A1 dataset. Each journey is assigned a journey identifier and can be linked to the corresponding vehicle in the A1 dataset that performs it. Journey related variables include gross weight of goods transported, place of loading and unloading (reported at a NUTS 2 level of disaggregation), actual distance traveled, tonne-km effected, degree of loading in terms of total volume and countries crossed in transit during each journey. Notably, survey distinguishes different journey types based on their laden/unladen status and the number of distinct transport operations involved. As a result four main journey types are identified: Laden-Involving one single transport operation, laden-Involving multiple transport operations, laden-collection/distribution and unladen. Journeys that involve 5 or more distinct locations are considered to be of collection/distribution type.

The goods dataset (Dataset A3) each journey is broken down to represent specific shipments of goods between two geographical units. Each goods' transfer between any two geographical units is identified and linked to the specific journey it is part of. Journeys that involve either multiple destinations for loading/unloading and/or different types of goods are further broken down in the goods dataset (Dataset A3). Each observation in Dataset A3 represents a flow of one type of good between two specific geographical units.

Region border changes Throughout the paper we use the classification NUTS2013 for most regions for consistency. In cases for which there was a change, a region split in more regions, from NUTS2010 to NUTS2013 we use the aggregated NUTS2010. This is the case for regions FI1B and FI1C (NUTS2013) in Finland, which we aggregate for all years in our data and corresponds to FI18 (NUTS2010). For London area regions UKI3, UKI4, UKI5, UKI6 and UKI7 (NUTS2013) we use the aggregated UKI1 (UKI3 + UKI4) and UKI2 (UKI5 + UKI6 + UKI7) NUTS2010 regions.

Cleaning data To create our matrix of weights of goods shipped between each region-pair we merge the good-level dataset (A3) for the years 2011 to 2017. We drop "unladen" journeys. We also drop "distribution" journeys, since these are journeys that involve five or more stops in distinct locations considered to be of collection/distribution nature. These are more likely associated with distribution or logistics than with trade. We then normalise the region identifiers to the 2013 NUTS version, since there are some regions that change name between 2011 and 2017. Finally, we apply the weights provided by Eurostat to each shipment to account for under-sampling of some journeys.

We then aggregate the value traded across all industries by each region pair by adding up the value traded in all industries for each region-pair in each year. Finally, to construct our region-pair level dataset we take the average of the value traded by each region pair (n,m) across all years 2011-2017.

D Additional data sources

D.1 Construction of geographical variables

1. *Distance*: We construct bilateral distance by calculating the length of the curve linking the central point of the origin region (centroid) and the central point of the destination region, in kilometers. We use a curve since we take into account the curvature of earth's surface. We compute the centroid as the center point of the polygon of the area of the region, using the software ArcGIS.

2. *Insularity*: Dummy variable taking value one if there is the need to cross a sea to reach from one region to the other, and zero otherwise.

3. *Mountain ranges*: Largest altitude difference between two regions, computed as the difference between the highest altitude point and the lowest altitude point along the straight line that joins the centre the origin region (centroid) and the centre of the destination region. To compute this maximum difference in altitude we use a topographic layer of Europe. We compute the straight line segment that links each possible region-pair (centroid to centroid). We then compute the altitude at different intervals along the line (computed using the cells of the altitude raster) and keep the highest and the lowest points. Finally, we take the difference between the highest and the lowest point.

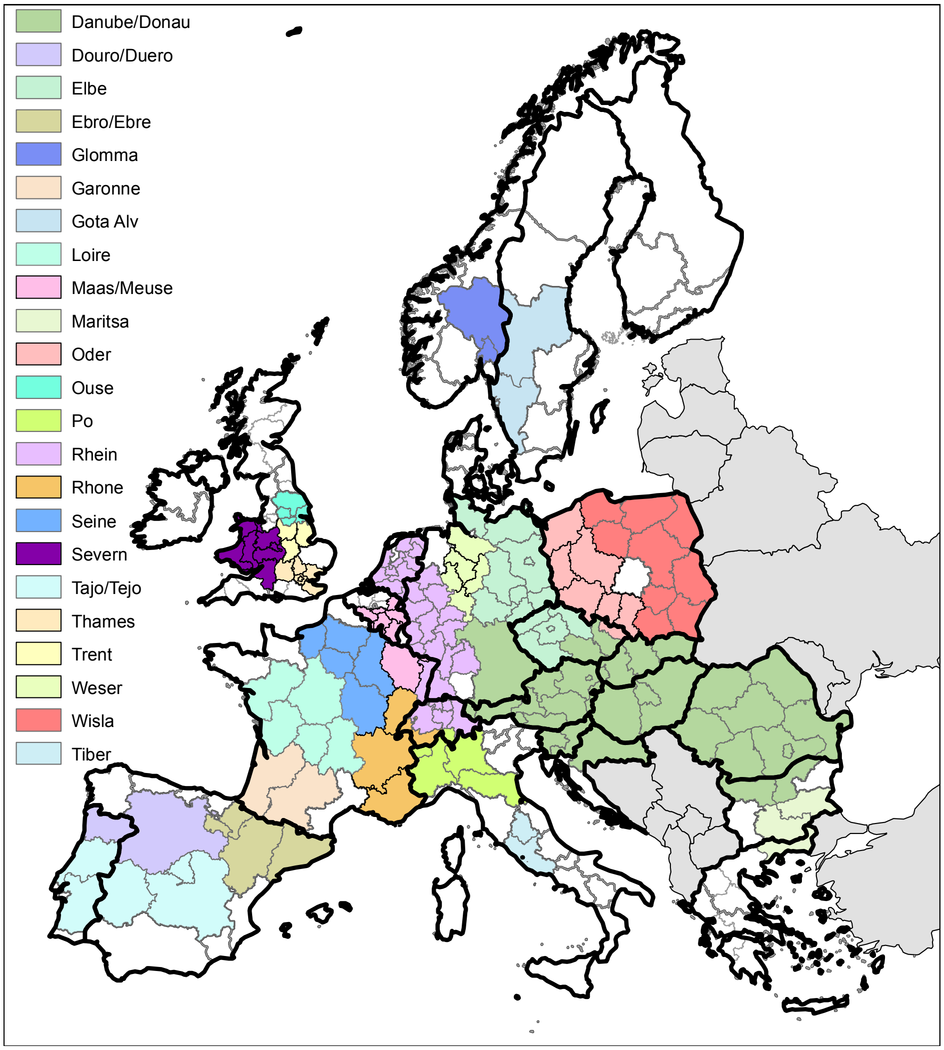
4. *River basin.* Dummy variable taking value 1 if both regions belong to the same river basin. We consider the largest rivers in Europe. A map of the areas covered by each

river basin is shown in figure D.1. We consider the major European rivers: Danube, Douro/Duero, Elbe, Ebro, Glomma, Garonne, Gota Alv, Loire, Meuse/Maas, Maritsa, Oder, Ouse, Po, Rhein, Rhone, Seine, Severn, Tejo/Tajo, Thames, Tiber, Trent, Weser, Vistula.

5. *Remoteness.* We calculate the remoteness of a region as the sum of the bilateral distance from that region to every other region in the sample. Then, we calculate the remoteness of a pair as the average remoteness of both regions.

6. *Number of borders* We sum the number of borders of the origin region and the number of border of the destination region, and we take the log of the sum. We compute the number of borders of the origin as the number of regions in the sample minus one (the border of the origin with itself) minus the number of regions in the country that the origin region belongs to (regions with which the origin region does not have a border). We do the same for the destination.

Figure D.1: Regions that share a river basin



Notes: This figure shows the different river basins that we consider, represented by different colors.

D.2 Collection of historical borders

We thank Matteo Cervellati, Sara Lazzaroni, Giovanni Prarolo and Paolo Vanin for kindly sharing their digitised data of historical borders in Europe from their paper Cervellati et al. (2019). We use the shapefile provided by the authors to identify borders in 1910 between our 269 regions.