Biased Gravity Estimates: Heteroskedasticity or Misspecification?

Scott French*

August, 2014

Abstract

Gravity estimations based on sector-level data implicitly assume that the effect of trade barriers on aggregated trade flows is independent of the patterns of comparative advantage that exist in the data. However, using a model that nests widely used quantitative trade models but allows for non-trivial patterns of comparative advantage across products, I show that, in general, sector-level trade flows follow a modified gravity equation that contains an unobservable, bilateral term that is ignored by traditional structural gravity estimations, which implies that their estimates suffer from omitted variable bias. I find that using product-level data to account for these patterns, leads to coefficient estimates that differ from traditional estimates in the ways predicted by the theory and that the product-level estimates are much more robust to distributional assumptions, implying that this bias is important and that, once it has been corrected, the remaining biases due to heteroskedasticity and sample selection are less severe.

JEL Classification: F10, F14, C13, C21

Keywords: gravity, omitted variable bias, international trade, pseudo-maximum likelihood, comparative advantage, heteroskedasticity

*School of Economics, University of New South Wales. scott.french@unsw.edu.au.

1 Introduction

The gravity model – which relates bilateral trade flows to the sizes of a pair of countries and the barriers to trade that exist between them – has long been celebrated as a parsimonious yet empirically successful way to describe bilateral trade flows. It is extremely useful as a framework within which to estimate the effect of factors that determine barriers to trade and to predict the effects of altering these factors. Since Anderson (1979), who showed that this empirical relationship is theoretically founded, it has also been useful as a method to quantify trade models, allowing for serious, general equilibrium analysis of the effects of such factors on economic outcomes and welfare. Quite often, the variables of interest are aggregate country-level or bilateral quantities, and the data that is most readily available are also quite aggregated, leading researchers to estimate the parameters of gravity models using aggregate data or, at least, industry-level data.

In this paper, I show that, in the presence of non-trivial patterns of comparative advantage across products, such estimations are misspecified, as they ignore the role of comparative advantage in shaping the effect of trade costs on trade barriers. To show this, I first develop a simple model – which is consistent with generalizations of a wide class of quantitative trade models – in which arbitrary patterns of comparative advantage can be taken into account. The presence of these patterns implies that the elasticity of aggregated trade flows with respect to trade costs is non-constant. I show that this effect is embodied in a bilateral term that appears in an otherwise standard sector-level gravity equation. Because this term is unobservable and a function of trade costs, it is a source of omitted variable bias in traditional gravity estimations.

This implies that bilateral trade barriers cannot be inferred from sector-level trade data. As a result, I propose a method to estimate trade barriers using pooled product-level trade data. This estimation procedure utilizes the models' product-level gravity structure to overcome practical issues that arise in the use of such data, such as the lack of available data on domestic trade flows and the computational infeasibility of fixed effects estimators with datasets consisting of a large number of countries and products.

I implement this procedure using data on bilateral product-level trade flows for 132 countries and 4,608 manufactured products from the UN Comtrade database and compare the coefficient estimates to those based on aggregated data. I find that the product-level estimates are much more robust to distributional assumptions regarding the error term than the sector-level estimates, and the coefficients from the sector-level and product-level estimations generally differ in the direction predicted by the theory. These results indicate that the omitted variable bias due to ignoring the effects of patterns of comparative advantage is at least as significant as biases due to heteroskedasticity and sample selection, which were highlighted by Santos Silva and Tenreyro (2006). Moreover, the bias is quantitatively important. For example the distance elasticity estimated via poisson pseudo-maximum-likelihood (PML) increases in absolute value by 29% from -0.73 to -0.94, which implies an ad-valorem equivalent trade cost that is 60% larger for a pair of countries 1,000 kilometers apart.[1]

The next section develops the model and derives the effect of patterns of comparative advantage on aggregate trade flows. Section 3 develops the product-level estimation procedure and presents the empirical results. The final section concludes.

2 Theoretical Framework

The world economy is comprised of N countries, each of which is characterized by a representative consumer who owns the factors of production and consumes goods from J sectors. I assume that the allocation of expenditure across products and producers within a sector can be analyzed separately from the allocation of production and consumption across sectors.[2] Each sector is made of a finite number of product categories, $k = 1, ..., K^j$, which each contain a continuum of product varieties $\omega \in [0, 1]$.[3] Thus, a particular variety is identified by the triple (j, k, ω). This structure allows the model to be mapped directly into data on product-level trade flows, such as the U.N. Comtrade database, which contains bilateral trade flow data for hundreds of countries classified into thousands of 6-digit Harmonized System product categories.

In line with the discussions in Anderson and van Wincoop (2004) and Arkolakis et al. (2012), the gravity structure derived below is consistent with a number of underlying models. I choose a Ricardian framework, based on Eaton and Kortum (2002), because it allows for a straightforward interpretation of the way in which patterns of comparative advantage confound the effect of trade barriers on aggregate trade flows. But, the results that follow do not depend crucially on this choice.

[1] This assumes a trade cost elasticity of 4.1, the value estimated by Simonovska and Waugh (2013).

[2] See Anderson and van Wincoop (2004) for a discussion of the conditions under which such separability arises.

[3] The assumption of a continuum of varieties is purely for analytical convenience. If the number of varieties per product category were finite, the results below would hold in expectation.

2.1 Preferences and Demand

A representative consumer in country n maximizes a nested Spence-Dixit-Stiglitz utility function over all varieties of all products, which implies the following:

Assumption 1. *Given total expenditure on all products in sector j, X_n^j, expenditure by country n on variety (j, k, ω) is given by*

$$X_n^{jk}(\omega) = \left(\frac{p_n^{jk}(\omega)}{P_n^{jk}}\right)^{1-\eta^{jk}} X_n^{jk},$$

where $\eta^{jk} > 1$, p_n^{jk} is the price of variety (j, k, ω) in n, $P_n^{jk} = \left(\int_0^1 p_n^{jk}(\omega)^{1-\eta^{jk}}\right)^{\frac{1}{1-\eta^{jk}}}$, and

$$X_n^{jk} = \left(\frac{P_n^{jk}}{P_n^j}\right)^{1-\sigma^j} X_n^j, \tag{1}$$

where $\sigma^j > 1$ and $P_n^j = \left(\sum_{k=1}^{K^j} (P_n^{jk})^{1-\sigma^j}\right)^{\frac{1}{1-\sigma^j}}$.

2.2 Technology and Prices

Every variety can be produced in any country and shipped anywhere in the world. Doing so, however, incurs trade costs, which are assumed to take the "iceberg" form, as in Samuelson (1954). Specifically, I assume the following:

Assumption 2. *Delivering one unit of variety (j, k, ω) to n from i requires shipping $d_{ni}^{jk} \geq 1$ units, where $d_{nn}^{jk} = 1$, for all n, and $d_{ni}^{jk} = d_{ni}^j d_n^{jk}$, for all $n \neq i$.*

The last restriction of Assumption 2 implies that product-level trade costs can be decomposed into a bilateral sector-specific component and an importer product-specific component. This restriction, which greatly simplifies the analysis that follows, is satisfied trivially by the assumptions of most sector-level gravity models and is consistent with import tariffs and non-tariff barriers that obey the Most Favored Nation principle of the WTO.[4]

I assume perfect competition and constant returns to scale in production, which together imply the following:

Assumption 3. *The price of variety (j, k, ω) in n is given by $p_n^{jk}(\omega) = \min_i\{p_{ni}^{jk}(\omega)\}$, where $p_{ni}^{jk}(\omega) = c_i^j d_{ni}^j / Z_i^{jk}(\omega)$, c_i^j is the cost of a bundle of inputs used in production in sector j and country i, and $Z_i^{jk}(\omega)$ is the productivity with which (j, k, ω) is produced in i.*

[4] This specification also does not rule out the possibility that trade barriers are asymmetric, i.e. $d_{ni}^j \neq d_{in}^j$ in general.

Finally, similar to Eaton and Kortum (2002), I assume that the productivity level for a given variety is drawn from a Fréchet distribution.

Assumption 4. *The productivity with which (j, k, ω) is produced in i is an independent realization of a random variable with the following cdf:*

$$F_i^{jk}(z) = e^{-(\gamma^{jk} Z_i^{jk}/z)^{\theta^j}}, \tag{2}$$

where $\theta^j > 1$.[5]

Idiosyncratic differences in productivity across varieties give rise to intra-product trade. The parameter θ^j governs the strength of comparative advantage within product categories, with a larger value of θ^j implying less variance in productivity across varieties and less scope for within-product comparative advantage. Inter-product trade depends on countries' patterns of comparative advantage *across* products, which are governed by relative values of Z_i^{jk}, which determines a country's average productivity across all varieties of product (j, k). Finally, a country's average overall level Z_i^{jk} determines its absolute advantage in sector j.

2.3 International Trade

To avoid excessive notation, in the remainder of the paper, I omit the sector superscript wherever there is no ambiguity. Following the analysis of Eaton and Kortum (2002), product-level imports by n from i are given by

$$X_{ni}^k = \left(\frac{c_i d_{ni}^k / Z_i^k}{P_n^k}\right)^{-\theta} X_n^k, \tag{3}$$

where $(P_n^k)^{-\theta} = \sum_i \left(c_i d_{ni}^k / Z_i^k\right)^{-\theta}$.

Using this result, it is possible to derive an expression relating total sector-level trade flows to countries' total output and expenditure and bilateral trade costs.

Proposition 1 (Sector-Level Gravity)**.** *Given assumptions 1 - 4, sectoral trade flows are given by the following system of equations:*

$$X_{ni} = \frac{Y_i X_n}{Y} \left(\frac{d_{ni}}{P_n \Pi_i}\right)^{-\theta} \tilde{Z}_{ni} \tag{4}$$

[5] The condition $\eta^{jk} - 1 < \theta^j$ is required for P_n^{jk} to be well defined. The constant $\gamma^{jk} = \Gamma(1 - (\eta^{jk} - 1)/\theta^j)^{\frac{1}{\eta^{jk}-1}}$, where $\Gamma(\cdot)$ is the gamma function. This constant is included in (2) purely for notational convenience, as it eliminates constants in the expressions for price indexes and relative expenditure across products that would appear otherwise. The only role that γ^k plays is in the mapping between relative productivity across products and relative sales. Otherwise, both η^{jk} and γ^{jk} are irrelevant to the analysis of this paper.

$$\Pi_i^{-\theta} = \sum_n \left(\frac{d_{ni}}{P_n}\right)^{-\theta} \frac{X_n}{Y} \tilde{Z}_{ni} \quad (5)$$

$$P_n^{-\theta} = \sum_i \left(\frac{d_{ni}}{\Pi_i}\right)^{-\theta} \frac{Y_i}{Y} \tilde{Z}_{ni}, \quad (6)$$

where $\tilde{Z}_{ni} = \sum_k (d_n^k)^{-\theta} \left(\frac{P_n^k}{P_n}\right)^{\theta-(\sigma-1)} \left(\frac{Z_i^k}{Z_i}\right)^{\theta}$, *and* $Z_i = \left(\sum_k (Z_i^k)^{\sigma-1}\right)^{\frac{1}{\sigma-1}}$.

Equation (4) is very nearly a standard gravity equation, as in Anderson and van Wincoop (2004), except for the presence of the term $\tilde{Z}_{ni}$. This term summarizes the effect of the interaction among countries' patterns of across-product comparative advantage on sector-level trade flows.

To understand this effect, note that P_n^k summarizes n's ease of access to efficiently produced varieties of k from anywhere in the world. A high price of k in n has two effects. From (3), given $c_i d_{ni}/Z_i^k$, a high value of P_n^k implies that sales in n of producers of k from i will be relatively high because there is relatively little competition from other locations. And, from (1), n's overall expenditure on k will be relatively low. The strength of the first effect is governed by θ, and that of the second is governed by $\sigma - 1$. If $\theta > \sigma - 1$ – which implies that the elasticity of substitution across source countries for a particular product is greater than the elasticity of substitution across products – then i will export relatively more to n if it is relatively productive for the products which are relatively difficult for n to obtain elsewhere.[6] In addition, i will export relatively more to n if the importer product-specific component of trade costs is relatively low for the products for which i is relatively productive.

The effect of countries' patterns of across-product comparative advantage on sector-level trade flows, which is summarized by $\tilde{Z}_{ni}^k$, is omitted from the most widely used quantitative trade models that imply a gravity equation. For example, Anderson and van Wincoop (2003) assumes that each country produces only a single good in each sector, which immediately precludes a role for comparative advantage across products. Krugman (1980) assumes that each country produces an entirely unique set of products, which implies that the margin of substitution across sources for a given product, described by (3), is degenerate. And, Eaton and Kortum (2002), implicitly assumes that relative productivity across products is identical for all countries – i.e. $Z_i^k/Z_i^{k'} = Z_{i'}^k/Z_{i'}^{k'}$ for any two countries and products – so that the only form of comparative advantage is the idiosyncratic within-product form. In all of these

[6] While the assumption that $\theta > \sigma - 1$ is not strictly necessary, it is commonly made with regard to nested CES models. If it were not the case, it would have the counterintuitive implication that an increase in $Z_{i'}^k$, for some $i' \neq i$, would lead to an increase in X_{ni}^k, for all n. As a result, I maintain this assumption for the discussion that follows.

cases, Z_{ni} drops out of (4).[7] However, in general, the presence of any non-trivial patterns of across-product comparative advantage in the data implies that this effect must be taken into account when estimating trade costs.

2.4 Biased Gravity

Because trade costs imply that prices differ a across markets, $\tilde{Z}_{ni}$ is bilateral term. This has important implications for the estimation of trade costs. The standard practice in the structural gravity literature is to estimate trade costs using aggregate or sectoral data, controlling for the endogenous variables in (4) with source and destination country fixed effects. However, because $\tilde{Z}_{ni}$ varies across country pairs, its effect on trade flows is not captured by the fixed effects, which implies that the coefficient estimates will suffer from omitted variable bias.

To gain some insight into the nature of this bias, consider the (partial) elasticity of $\tilde{Z}_{ni}$ with respect to d_{ni}.[8]

$$\frac{\partial \ln(\tilde{Z}_{ni})}{\partial \ln d_{ni}} = [\theta - (\sigma - 1)] \sum_k \frac{X_{ni}^k}{X_{ni}} \left(\frac{X_{ni}^k}{X_n^k} - \frac{X_{ni}}{X_n} \right). \tag{7}$$

The summation term lies in the interval $[0, 1 - X_{ni}/X_n)$. Thus it is always positive, and it is weakly increasing in both i's overall market share in n and the degree to which i's exports to n are concentrated in products for which it has a relatively strong comparative advantage. If we suppose that $\theta > \sigma - 1$, then this – together with the fact that X_{ni} is increasing in $\tilde{Z}_{ni}$ – implies that estimates based on the standard approach will be biased toward zero, in general, and the bias will be more severe in samples of relatively large exporters and ones whose exports are concentrated in a relatively unique set of products.

3 Estimating Trade Costs

In order to evaluate the degree to which trade cost estimates based on aggregated data are biased due to ignoring the effect of countries' patterns of comparative advantage, I propose and implement a method for estimating trade costs using product-level trade data, which does not suffer from omitted variable bias. To this end, following the gravity literature, I

[7] See French (2014) for a more detailed analysis of the mapping between these models and the one of this paper.

[8] This elasticity holds constant total sectoral expenditure and input costs in every country.

parameterize trade costs in the following way:

$$\ln(d_{ni}^k) = \ln(d_n^k) + \ln(d_i) + \beta \ln(dist_{ni}) + bord_{ni} + lang_{ni} + col_{ni} + rta_{ni}, \tag{8}$$

for $n \neq i$, where d_n^k is an importer product-specific trade cost, d_i is an exporter-specific border cost; $dist_{ni}^m$ is the geographical distance between n and i; $bord_{ni}$ is the effect of countries n and i sharing a common border; $lang_{ni}$ is the effect of sharing a common language; col_{ni} is the effect of having a colonial relationship; and rta_{ni} is the effect of n and i being part of a regional trade agreement[9]

3.1 Standard Estimation Methods

The typical strategy employed to identify the parameters of a trade cost function such as (8) is to take advantage of the log-linear form of (4) to estimate the parameters via OLS, controlling for the endogenous variables using importer and exporter fixed effects. However, this can potentially produce biased estimates for three reasons. First, as discussed above, it suffers from omitted variable bias by ignoring the effect of $\tilde{Z}_{ni}$. Second, using $\ln(X_{ni})$ as the dependent variable means that country pairs with zero trade flows are dropped from the estimation, resulting in sample selection bias.[10] And, third, as is pointed out by Santos Silva and Tenreyro (2006), due to Jensen's inequality, estimates based on the log-linear specification are biased in the presence of heteroskedasticity.

In order to correct for the last two sources of bias, Santos Silva and Tenreyro (2006) propose estimating (4) in its multiplicative form using the pseudo-maximum-likelihood (PML) techniques first described by Gourieroux et al. (1984). These estimators are unbiased as long as the conditional mean is correctly specified and allow the inclusion of zero valued trade flows in the estimation. Santos Silva and Tenreyro (2006) advocate the use of Poisson PML; however, other distributional assumptions within the linear-exponential family – such as the Gaussian and the gamma distributions – also imply valid PML estimators. The primary difference among the set of PML estimators is the form of heteroskedasticity implied by the underlying distributions and thus the weighting of different observations in the likelihood function. This point will prove to be important for interpreting the estimation results below.

Despite addressing two potential sources of bias in traditional structural gravity estimations, PML estimators, alone, cannot correct for the omitted variable bias that is present when the effect of $\tilde{Z}_{ni}$ is ignored, since this implies that the conditional expectation of X_{ni} is

[9] I assume that the border effect is exporter-specific following Waugh (2010), which argues that this specification is more consistent with data on the prices of tradable goods.

[10] See Eaton and Tamura (1994), Helpman et al. (2008), and Hallak (2006) for attempts to deal with this form of bias while maintaining the log-linear formulation of the regression equation.

misspecified. Further, because, $\tilde{Z}_{ni}$ is generally unobservable from sector-level data, the only way to consistently estimate the parameters of (8) is to use product-level data to account for the patterns of comparative advantage that exist in the data.

3.2 Issues With Product-Level Estimation

Before outlining the estimation procedure based on product-level data in detail, there are two issues related to using product-level data in a gravity-type estimation that warrant discussion. The first is that the large amount of data necessarily utilized makes standard techniques involving fixed effects infeasible. The second arises due to the general lack of data on domestic trade flows at the product level.

3.2.1 Fixed Effects Estimation

In principle, based on (3), trade costs can be estimated using product-level data and controlling for the unobservable variables with a full set of importer-product and exporter-product fixed effects. Under the restriction on the form of trade costs of Assumption 2, the d_n^k component is absorbed by the importer-product fixed effect, and this provides a consistent estimate of d_{ni}. To obtain the most efficient estimates, the estimation should use the entire sample of product-level bilateral trade flows, pooled across all products within the sector. However, this quickly becomes computationally infeasible as the sample size gets large. For example, there are 132 countries and 4,608 product categories in the dataset employed below. This method would require the computation of $2K(N-1) = 1,207,296$ fixed effects, which is well beyond the capabilities of most computers.

Alternatively, the estimation could be done product-by-product. This is technically feasible, as it requires on $2(N-1)$ fixed effects per product-level estimation. It also has the advantage of relaxing the restriction on the form of trade costs of Assumption 2 by allowing coefficient estimates to differ across products. However, this is not only an inefficient way to estimate d_{ni}, it results in potentially thousands of sets of coefficient estimates, making interpretation of the results very difficult. Thus, I do not consider this to be a particularly useful estimation procedure.[11]

[11] As a test, I performed such a product-by-product estimation and found that, while there was a significant amount of variance in the coefficient estimates across products, for fewer than one quarter of the estimates could the null hypothesis that the value was equal to its corresponding value from the pooled estimation (reported in the bottom section of Table 1) be rejected at the 10% level of significance. Thus, it is very difficult to conclude whether there is significant deviation in the data from trade costs of the form of Assumption 2 or whether the variance in estimates is simply due to noise in the data.

3.2.2 Domestic Trade Flows

The second major issue with a gravity estimation using product-level data is that data on domestic trade flows, X_{nn}^k, is typically not available at anywhere near the level of disaggregation of the international trade data. This is important because such data is require to identify the country-specific components of trade costs, d_i and d_n^k. This is because these costs are only incurred when a product crosses an international border and thus data on trade flows that do not cross borders is required to identify their effects.

3.3 Product-Level Gravity

To address these issues, Proposition 2 shows that product-level trade flows can be expressed as a function of only countries' total product-level exports and imports and bilateral trade costs.

Proposition 2 (Product-Level Gravity)**.** *Given assumptions 1 - 4, product-level trade flows are given by the following system of equations:*

$$X_{ni}^k = \frac{E_i^k M_n^k}{E^k} \left(\frac{\tilde{d}_{ni}}{\tilde{P}_n^k \tilde{\Pi}_i^k} \right)^{-\theta} \tag{9}$$

$$(\tilde{\Pi}_i^k)^{-\theta} = \sum_{n \neq i} \left(\frac{\tilde{d}_{ni}}{\tilde{P}_n^k} \right)^{-\theta} \frac{M_n^k}{E^k} \tag{10}$$

$$(\tilde{P}_n)^{-\theta} = \sum_{i \neq n} \left(\frac{\tilde{d}_{ni}}{\tilde{\Pi}_i} \right)^{-\theta} \frac{E_i^k}{E^k}, \tag{11}$$

where E_i^k is total exports of k by i, M_n^k is total imports of k by n, E^k is total world trade flows of k, and $\tilde{d}_{ni} = d_{ni}^k/(d_n^k d_i)$.

This proposition is useful for two reasons. First, it shows that, as in the sector-level model of Anderson and van Wincoop (2003), the endogenous variables, $\tilde{P}_n^k$ and $\tilde{\Pi}_i^k$ can be computed from data on M_n^k and E_i^k, given values of $\tilde{d}_{ni}$, so fixed effects are not required in an estimation. Second, because trade flows are expressed as a function of total product-level imports and exports, and not expenditure and output, no data on X_{nn}^k is required.

However, the issue of identifying d_n^k and d_i remains. For this, I use the fact that sector-level data on domestic trade flows generally *is* available. For the moment, assume that $d_n^k = 1$, for all n and k. Then, given the estimated value of $\tilde{d}_{ni}$, denoted $\hat{d}_{ni}$, the predicted value of X_{nn}^k is given by

$$\hat{X}_{nn}^k = d_i^\theta \frac{E_i^k M_n^k / E^k}{(\hat{P}_n^k \hat{\Pi}_i^k)^{-\theta}},$$

where $\hat{P}_n^k$ and $\hat{\Pi}_i^k$ are the respective values of $\tilde{P}_n^k$ and $\tilde{\Pi}_i^k$ evaluated at $\tilde{d}_{ni} = \hat{d}_{ni}$. Thus, I take the value of d_i to be the one for which $X_{nn} = \sum_k \hat{X}_{nn}^k$.[12]

3.4 Estimation Procedure

With this specification of trade costs, the stochastic form of (9) is

$$X_{ni}^k = \frac{M_n^k E_i^k}{E^k} \left(\frac{\tilde{d}_{ni}}{\tilde{P}_n^k \tilde{\Pi}_i^k} \right)^{-\theta} + \epsilon_{ni}^k, \tag{12}$$

where $\tilde{P}_n^k$ and $\tilde{\Pi}_i^k$ are given by (10) and (11), respectively. The error term can be thought of as measurement error.

As with sector-level estimations, estimates based on the log-linear form of (12) will suffer from sample selection bias and bias in the presence of heteroskedasticity. Thus, I employ Poisson PML as suggested by Santos Silva and Tenreyro (2006) as well as, for comparison, gamma and Gaussian PML, where the later is equivalent to non-linear least squares (LS) based on the multiplicative form of (12). For comparison with more traditional estimations in the literature, I also employ a least squares estimator based on the log-linear form of (9). This last estimator reduces to OLS when fixed effects are used to control for $\tilde{P}_n^j$ and $\tilde{\Pi}_i^j$, and it is very similar to the nonlinear LS estimator of Anderson and van Wincoop (2003), when these terms are computed using (10) and (11).

The estimation proceeds as follows. Given a set of parameters of (8) and data on M_n^k and E_i^k, I compute $\tilde{P}_n^j$ and $\tilde{\Pi}_i^j$ using (10) and (11). Then, using all of these values, I calculate the predicted values of X_{ni}^k according to (9). I then update the set of parameters of (8) until reaching an optimum of the objective function implied the assumed distribution of ϵ_{ni}^k. Because this procedure predicts the value of trade flows conditional on data on total imports and exports, I refer this as the "conditional" estimation in what follows.

To assess the degree of bias in sectoral estimates that ignore the effect of $\tilde{Z}_{ni}$, I compare the estimates based on product-level data and this procedure, with two sets of estimates based on sector-level data. The first uses importer and exporter fixed effects, as is most commonly done in the literature. The second uses the same procedure as the product-level conditional estimation but uses only sector-level data, which implicitly assumes that $\tilde{Z}_{ni} = 1$

[12]The discussion of the identification of d_n^k is coming soon!

for all country pairs in (4).

3.5 Data

Product-level trade flow data are from the U.N. Comtrade database. The data used is for manufactured goods for the year 2003, and bilateral trade flows are categorized at the 6-digit level according to the 1996 revision of the Harmonized System. Manufacturing output data, which is used to calculate X_{nn}^{j} is taken from the OECD STAN database, where available, or the UNIDO INDSTAT database. Where not available from either source, it is imputed based on manufacturing value added from the World Bank's WDI database. Data on bilateral relationships are taken from CEPII's Gravity dataset.

When manufacturing is taken to consist of a single sector, the sample consists of trade flows among 132 countries classified into 4,608 product categories. When a sector is defined as a 2-digit ISIC industry, the sample size is reduced to 60 countries due to lack of disaggregated manufacturing output data. Table A1 lists the countries in the sample and the source of output data for each, and Table A2 lists the set of industries. Further details are in the Appendix.

3.6 Estimation Results

Table 1 presents the coefficient estimates from the three estimation procedures and four distributional assumptions described above for the entire manufacturing sector. The results from the aggregate estimations are roughly in line with the literature. Bilateral trade is generally decreasing in distance and higher if countries share a border, language, colonial ties, or a regional trade agreement.

As in Santos Silva and Tenreyro (2006), in the sector-level fixed effects estimation, the coefficient estimates differ greatly between the log LS and Poisson PML estimations, which they take as evidence of bias due to heteroskedasticity and sample selection. Interestingly, the gamma PML estimates, which should not suffer from these biases, differ to a greater extent from the Poisson PML estimates in many cases, while the multiplicative LS estimates are somewhat closer to those from Poisson PML and, in every case, differ in the opposite direction than the gamma PML estimates.

Moving to the aggregate conditional estimations, we see that the results are generally similar to those of the aggregate fixed effects estimations, with the exception of the coefficient on shared border, which switches signs for both log LS and gamma PML. The Poisson PML estimates are identical in both cases. This is because, when country-level fixed effects are included, the Poisson likelihood function is maximized at the point where each country's

Table 1: Trade Cost Coefficient Estimates

	Log LS		Gamma PML		Poisson PML		Mult. LS	
Variable	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
			Sector-Level Fixed Effects Estimation					
mean($\ln(d_i)$)	−0.48	(0.34)	1.88	(0.7)	−3.53	(0.31)	−4.71	(0.38)
Distance	−1.73	(0.07)	−2.09	(0.1)	−0.73	(0.05)	−0.61	(0.06)
Shared Border	0.76	(0.15)	0.72	(0.25)	0.42	(0.09)	0.29	(0.05)
Common Language	0.94	(0.1)	1.16	(0.14)	0.38	(0.09)	0.60	(0.11)
Colonial Ties	0.94	(0.12)	1.20	(0.29)	0.02	(0.12)	−0.28	(0.14)
RTA	0.53	(0.12)	0.29	(0.21)	0.82	(0.08)	1.02	(0.12)
			Sector-Level Conditional Estimation					
mean($\ln(d_i)$)	0.45		−1.64		−3.53		−4.03	
Distance	−1.71	(0.07)	−1.35	(0.07)	−0.73	(0.05)	−0.57	(0.05)
Shared Border	−0.54	(0.21)	−0.22	(0.15)	0.42	(0.09)	0.32	(0.09)
Common Language	1.36	(0.16)	0.63	(0.13)	0.38	(0.09)	0.35	(0.09)
Colonial Ties	1.43	(0.17)	0.77	(0.18)	0.02	(0.12)	−0.17	(0.12)
RTA	1.31	(0.16)	0.50	(0.14)	0.82	(0.08)	1.22	(0.08)
			Product-Level Conditional Estimation					
mean($\ln(d_i)$)	−1.41		−1.74		−2.77		−3.77	
Distance	−1.24	(0.03)	−1.16	(0.13)	−0.94	(0.06)	−0.71	(0.04)
Shared Border	0.58	(0.04)	0.39	(0.25)	0.39	(0.08)	0.33	(0.08)
Common Language	0.71	(0.04)	1.00	(0.21)	0.45	(0.08)	0.61	(0.07)
Colonial Ties	0.63	(0.06)	0.39	(0.16)	0.02	(0.09)	−0.52	(0.09)
RTA	0.32	(0.06)	0.43	(0.45)	0.80	(0.08)	0.98	(0.08)

Notes: Standard errors are clustered by exporter. Parameters reported represent $-\theta\hat{\beta}$. The implied percentage effect of each coefficient on the ad valorem tariff equivalent trade cost is $100 \times (e^{\hat{\beta}/\theta} - 1)$. Number of observations: 11,588 for aggregate log LS; 17,292 for aggregate PML; 3,600,740 for product-level log LS; 79,681,536 for product-level PML.

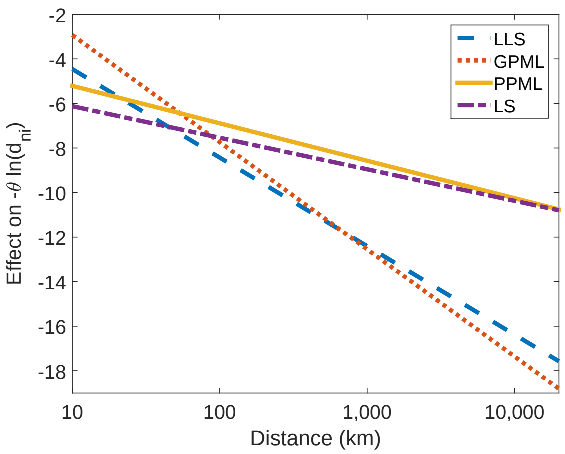
total trade flows exactly match the data. This is exactly the condition that is imposed by the conditional estimator, so the two are equivalent.

By contrast, the coefficient estimates based on the four different distributional assumptions are much more similar in the product-level estimations. This is consistent with the presence of omitted variable bias in the aggregate estimates. In the cases of Poisson PML and multiplicative LS, the coefficients all move away from zero – except for the RTA coefficient for multiplicative LS – which is what the model predicted. This is particularly the case for the effect of distance in trade flows. This is illustrated more clearly in Figure 1, which plots the estimated effect of distance for a country with the average exporter-specific border cost.

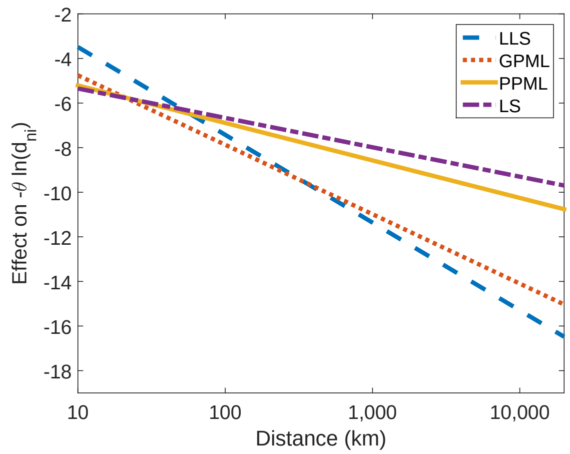
A different pattern, however, is evident for the cases of log LS and gamma PML, for which most of the coefficients are closer to zero in the product-level estimations. I will argue that this is to be expected given the properties of these estimators. First, log LS and gamma PML

Figure 1: Estimated Distance-Related Trade Costs

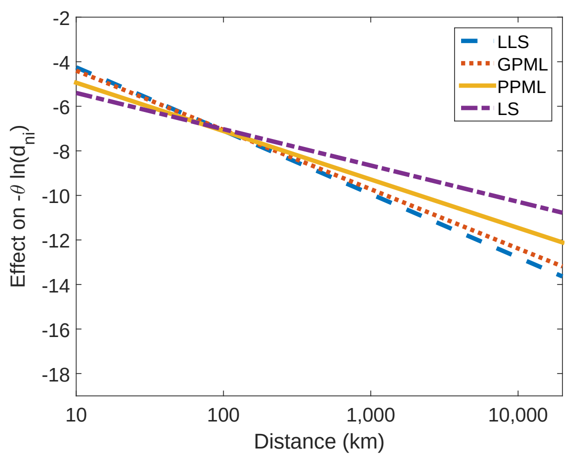
(a) Aggregate F.E. Estimation



(b) Aggregate Conditional Estimation



(c) Product-Level Conditional Estimation



assume that the conditional variance of the dependent variable is proportional the square of its expected value and thus places more weight on smaller observations. By contrast, Poisson PML assumes that the variance is proportional to the expected value and weights all observations equally, and multiplicative LS assumes that the variance is constant and places greater weight on larger observations. Further, in addition to pointing out that log LS is biased in the presence heteroskedasticity, Santos Silva and Tenreyro (2006) demonstrate in a Monte Carlo experiment that the gamma PML estimator tends to be very sensitive to a particular form of measurement error in the data, whereby small trade flows are rounded to zero.

It is intuitive that the gamma PML estimator, which heavily weights small observations, would be sensitive to such measurement error, and, combined with the omitted variable bias present in sector-level gravity estimations, this is likely to bias the coefficient estimates away from zero. To understand why, recall that, as equation (7) demonstrates, smaller trade flows are generally more sensitive to trade costs. Thus, the misspecified sector-level model – if fitted to match intermediate-sized trade flows – would under-predict the effect of trade barriers on smaller trade flows and thus over-predict these trade flows. If such small expected trade flows show up as zeros in the data, the estimator will require that trade flows, overall, be very sensitive to trade costs in order to make the predicted values of these small observations close to zero.

This intuition is also consistent with the flipping of the sign of the coefficient on shared border between the fixed effects and conditional estimations for these estimators. As is discussed above, the key difference between these two estimation techniques is that the latter imposes the adding-up constraint that a country's bilateral imports and exports sum to their total values in the data. In the misspecified sector-level estimation, the log LS and gamma PML estimators need large distance-related trade costs to keep the model from over-predicting the value of relatively small trade flows. This comes at the expense of seriously under-predicting the value of relatively large trade flows, which receive little weight in the objective function. In the conditional estimation, by contrast, because bilateral trade flows must add up to a fixed quantity, large distance-related trade costs *increase* the model's predicted values for trade flows among nearby countries, which tend to be relatively large. In this case, a negative effect of a shared border can partially correct for this over-prediction.

The fact that the coefficients are much more similar for the product-level estimation across all the distributional assumptions indicates that the model that accounts for the effects of patterns of comparative advantage is much more capable of predicting trade flows across all country pairs than is a sector-level gravity model. And, that the coefficients from the log LS and poisson PML estimators are much more similar in the product-level estimations indicates that the bias due to heteroskedasticity and sample selection in the case of log LS may not be as severe as previously thought.

However, the fact that the coefficients on colonial ties and regional trade agreements change monotonically as one moves from gamma PML to multiplicative LS – increasing the weight the placed on larger observations – indicates the true elasticity of trade flows with respect to these variables may be non-constant. In particular, it seems reasonable to speculate that colonial ties are more important for relatively small former colonies whose economies may have been greatly shaped by their colonizers and may still depend heavily on investment as well as political and military support from the former colonial power. In the case of trade agreements, this could reflect the fact that trade agreements among blocs of large countries, such as NAFTA and the EU customs union, go much farther in scope than other regional agreements.

3.7 Multiple Sectors

Though gravity estimations have almost always been conducted at the sector level, recent paper such as Anderson and Yotov (2011) and Levchenko and Zhang (2013) have defined sectors more narrowly, studying trade flows within manufacturing industries defined approximated at the 2-digit ISIC level. To evaluate the extent to which the omitted variable bias that is the subject of this paper is problematic for gravity estimations focused on somewhat more disaggregate sectors, I repeat the estimations from above industry-by-industry.

Tables A3 - A6 present the results of industry-level estimations comparable to those reported in Table 1, which were conducted separately for the 18 industries defined in Table A2.[13] The overall message from these tables is that, while the coefficient estimates differ substantially across industries, the patterns across estimations are generally similar to those of the estimation on manufacturing as a whole. In particular, 3/4 of the coefficients estimated by poisson PML are larger in absolute value in the product-level than the sector-level estimations. This indicates that, even at the industry-level, the patterns of cross-product comparative advantage are such that the omitted variable bias of sector-level gravity estimations remains large.

4 Conclusion

This paper has shown theoretically that presence of non-trivial patterns of comparative advantage across products implies that aggregated trade flows do not obey a standard structural gravity equation. In particular, the theoretically correct equation includes an unobservable bilateral term which leads to omitted variable bias in traditional gravity estimations. As a result, I have developed an approach to estimating trade costs using product-level trade data and using the structure of the model to overcome practical issues with fixed effects estimation using product-level data. Comparing coefficient estimates based on sector and product-level data indicates that this bias is significant. Thus, researches should use caution in interpreting the results of sector-level gravity estimations.

[13] Note: These results are preliminary

References

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

_ **and** _ , "Trade Costs," *Journal of Economic Literature*, 2004, *42* (3), 691–751.

_ **and Yoto V. Yotov**, "Specialisation: Pro and Anti-Globalizing 1990-2002," 2011, *Working Paper.*

Arkolakis, Costas, Arnaud Costinot, and Andres Rodriguez-Clare, "New Trade Models, Same Old Gains?," *American Economic Review*, 2012, *102* (1), 94–130.

Eaton, Jonathan and Akiko Tamura, "Bilateralism and Regionalism in Japanese and U.S. Trade and Direct Foreign Investment Patterns," *Journal of the Japanese and International Economies*, 1994, *8* (4), 478 – 510.

_ **and Samuel Kortum**, "Technology, Geography, and Trade," *Econometrica*, 2002, *70* (5), 1741–1779.

French, Scott, "The Composition of Trade Flows and the Aggregate Effects of Trade Barriers," 2014, *Working Paper.*

Gourieroux, C., A. Monfort, and A. Trognon, "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 1984, *52* (3), pp. 681–700.

Hallak, Juan Carlos, "Product quality and the direction of trade," *Journal of International Economics*, 2006, *68* (1), 238 – 265.

Helpman, Elhanan, Marc Melitz, and Yona Rubinstein, "Estimating Trade Flows: Trading Partners and Trading Volumes," *Quarterly Journal of Economics*, 2008, *123* (2), 441–487.

Krugman, Paul, "Scale Economies, Product Differentiation, and the Pattern of Trade," *The American Economic Review*, 1980, *70* (5), pp. 950–959.

Levchenko, Andrei A. and Jing Zhang, "The Evolution of Comparative Advantage: Measurement and Welfare Implications," 2013, *Working Paper.*

Samuelson, Paul A., "The Transfer Problem and Transport Costs, II: Analysis of Effects of Trade Impediments," *The Economic Journal*, 1954, *64* (254), pp. 264–289.

Silva, J. M. C. Santos and Silvana Tenreyro, "The Log of Gravity," *The Review of Economics and Statistics*, 2006, *88* (4), pp. 641–658.

Simonovska, Ina and Michael E. Waugh, "The Elasticity of Trade: Estimates and Evidence," 2013, *Working Paper.*

Waugh, Michael E., "International Trade and Income Differences," *American Economic Review*, 2010, *100* (5), 2093–2124.

A Additional Tables

Table A1: Countries and Sources of Manufacturing Output Data

Country	Source	Country	Source	Country	Source
Albania	INDSTAT	Gambia	WDI	Panama	INDSTAT(int.)
Argentina	WDI	Georgia*	INDSTAT	Papua New Guinea	WDI
Australia*	INDSTAT	Germany*	STAN	Paraguay	WDI
Austria*	STAN	Ghana	INDSTAT	Peru	INDSTAT
Azerbaijan*	INDSTAT	Greece*	STAN	Philippines	INDSTAT
Bahamas	WDI	Guatemala	WDI	Poland*	STAN
Bangladesh	WDI	Honduras	WDI	Portugal*	STAN
Barbados	WDI	Hungary*	STAN	Qatar	INDSTAT
Belarus	WDI	Iceland*	STAN	Rep. of Korea*	STAN
Belize	WDI	India*	INDSTAT	Rep. of Moldova*	INDSTAT
Benin	WDI	Indonesia*	INDSTAT	Romania	INDSTAT
Bolivia	WDI	Iran*	INDSTAT	Russian Federation*	INDSTAT
Bosnia Herzegovina	WDI	Ireland*	STAN	Rwanda	WDI
Botswana	INDSTAT	Israel*	STAN	St. Lucia	WDI
Brazil*	INDSTAT	Italy*	STAN	St. Vinc. and Gren.	WDI
Brunei Darussalam	WDI	Jamaica	WDI	Samoa	WDI
Bulgaria*	INDSTAT	Japan*	STAN	Sao Tome and Princ.	WDI
Burkina Faso	WDI	Jordan*	INDSTAT	Saudi Arabia	INDSTAT(int.)
Burundi	WDI	Kazakhstan*	INDSTAT	Senegal	WDI
Cambodia	WDI	Kenya*	INDSTAT	Slovakia*	STAN
Cameroon	WDI	Kyrgyzstan*	INDSTAT	Slovenia*	STAN
Canada*	STAN	Latvia*	INDSTAT	South Africa*	INDSTAT
Cape Verde	WDI	Lebanon	WDI	Spain*	STAN
Central African Rep.	WDI	Lithuania*	INDSTAT	Sri Lanka	INDSTAT(int.)
Chile*	INDSTAT	Madagascar*	INDSTAT	Sudan	WDI
China*	INDSTAT	Malawi	WDI	Swaziland	WDI
Colombia*	INDSTAT	Malaysia*	INDSTAT	Sweden*	STAN
Costa Rica	WDI	Maldives	WDI	Switzerland*	STAN
Croatia	WDI	Malta*	INDSTAT	Syria	INDSTAT
Cuba	WDI	Mauritania	WDI	TFYR of Macedonia	INDSTAT
Cyprus	INDSTAT	Mauritius	INDSTAT	Thailand*	INDSTAT(int.)
Czech Rep.*	STAN	Mexico*	STAN	Togo	WDI
Cte d'Ivoire	WDI	Morocco	INDSTAT	Trinidad and Tobago*	INDSTAT
Denmark*	STAN	Mozambique	WDI	Tunisia*	INDSTAT
Dominican Rep.	WDI	Namibia	WDI	Turkey*	INDSTAT
Ecuador*	INDSTAT	Nepal	WDI	USA*	STAN
El Salvador	WDI	Netherlands*	STAN	Uganda	WDI
Eritrea	INDSTAT	New Zealand*	STAN	Ukraine*	INDSTAT
Estonia*	STAN	Nicaragua	WDI	United Kingdom*	STAN
Ethiopia	INDSTAT	Niger	WDI	Utd. Rep. of Tanzania	INDSTAT
Fiji	INDSTAT	Nigeria*	INDSTAT	Uruguay*	INDSTAT
Finland*	STAN	Norway*	STAN	Venezuela	WDI
France*	STAN	Oman	INDSTAT	Viet Nam	INDSTAT
Gabon	WDI	Pakistan	INDSTAT(int.)	Zambia	WDI

* Sector-level manufacturing output data available.
Notes: INDSTAT(int.) indicates that output data was interpolated based on INDSTAT data for years before and after 2003.

Table A2: ISIC Rev. 3 Sectors

ISIC code	Sector Description	HS-6 Products	δ^j
15A	Food, beverages, and tobacco	427	0.145
17	Textiles	541	0.023
18	Wearing apparel, fur	241	0.017
19	Leather, leather products, and footwear	67	0.007
20	Wood products (excluding furniture)	69	0.019
21	Paper and paper products	119	0.030
22	Printing and publishing	36	0.047
23	Coke, refined petroleum products, nuclear fuel	20	0.053
24	Chemicals and chemical products	877	0.102
25	Rubber and plastics products	121	0.039
26	Non-metallic mineral products	170	0.032
27	Basic metals	359	0.061
28	Fabricated metal products	221	0.055
29C	Office, accounting, computing machinery; Other machinery	565	0.093
31A	Electrical machinery; Communication equipment	235	0.085
33	Medical, precision and optical instruments	211	0.020
34A	Transport equipment	135	0.133
36	Furniture, other manufacturing	189	0.036

Table A3: Multi-Sector Trade Costs Coefficient Estimates (Log Least Squres)

(a) Aggregate Model

	Food	Textiles	Apparel	Leather	Wood	Paper	Printing	Petr/Coal	Chemicals	Rub/Plstc	Minerals	Bas. Metal	Fab. Metal	Computing	Electrical	Medical	Transport	Furniture
mean($\ln(d_i)$)	−0.81	1.54	1.15	1.50	−0.54	0.13	−0.95	1.42	0.55	0.30	−0.65	1.07	−0.28	0.70	0.43	−0.61	0.02	−0.26
Distance	−1.35	−1.48	−1.52	−1.43	−1.64	−1.74	−1.74	−3.04	−1.48	−1.50	−1.63	−1.81	−1.44	−1.31	−1.34	−1.01	−1.30	−1.54
	(0.08)	(0.10)	(0.10)	(0.07)	(0.08)	(0.10)	(0.13)	(0.10)	(0.10)	(0.08)	(0.10)	(0.09)	(0.07)	(0.06)	(0.09)	(0.09)	(0.08)	(0.12)
Shared Border	0.37	−0.12	−0.82	−0.03	−0.19	−0.08	0.07	0.15	−0.37	0.11	0.28	0.63	−0.17	−0.46	−0.38	0.05	−0.10	−0.27
	(0.16)	(0.18)	(0.19)	(0.13)	(0.20)	(0.17)	(0.32)	(0.23)	(0.12)	(0.14)	(0.11)	(0.23)	(0.17)	(0.17)	(0.25)	(0.28)	(0.16)	(0.19)
Common Language	0.41	0.25	0.98	0.64	0.76	0.40	0.74	0.31	0.91	0.64	0.44	0.01	0.63	0.78	0.56	0.27	0.67	0.56
	(0.16)	(0.21)	(0.20)	(0.17)	(0.23)	(0.13)	(0.22)	(0.26)	(0.11)	(0.16)	(0.13)	(0.15)	(0.19)	(0.17)	(0.28)	(0.20)	(0.19)	(0.21)
Colonial Ties	1.75	2.32	2.98	2.88	1.51	1.12	1.85	1.33	1.34	1.39	1.57	1.26	1.84	1.59	1.62	1.38	1.17	1.33
	(0.15)	(0.20)	(0.16)	(0.14)	(0.17)	(0.17)	(0.24)	(0.20)	(0.15)	(0.20)	(0.18)	(0.17)	(0.16)	(0.16)	(0.21)	(0.27)	(0.14)	(0.27)
RTA	1.08	1.85	1.74	1.19	1.13	1.80	1.26	−0.38	0.99	1.70	1.19	0.61	1.69	1.43	1.54	1.43	1.93	1.24
	(0.09)	(0.15)	(0.19)	(0.11)	(0.14)	(0.19)	(0.19)	(0.13)	(0.15)	(0.12)	(0.20)	(0.12)	(0.09)	(0.11)	(0.14)	(0.16)	(0.10)	(0.27)
No. of Obs.	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540

(b) Product-Level Model

	Food	Textiles	Apparel	Leather	Wood	Paper	Printing	Petr/Coal	Chemicals	Rub/Plstc	Minerals	Bas. Metal	Fab. Metal	Computing	Electrical	Medical	Transport	Furniture
mean($\ln(d_i)$)	−2.20	−0.62	−1.31	−0.60	−2.05	−0.54	−1.85	−1.02	−0.60	−0.49	−2.35	−0.54	−1.40	−0.13	−0.06	−1.07	−0.61	−1.08
Distance	−1.24	−1.16	−1.16	−1.18	−1.30	−1.57	−1.44	−1.63	−1.13	−1.35	−1.20	−1.22	−1.24	−1.15	−1.28	−0.97	−1.17	−1.11
	(0.04)	(0.04)	(0.06)	(0.05)	(0.05)	(0.04)	(0.04)	(0.08)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)	(0.04)
Shared Border	0.89	0.51	0.46	0.82	0.96	0.85	0.95	0.90	0.64	0.97	0.94	0.69	0.83	0.50	0.74	0.56	0.66	0.82
	(0.06)	(0.05)	(0.09)	(0.06)	(0.08)	(0.06)	(0.08)	(0.11)	(0.06)	(0.05)	(0.04)	(0.05)	(0.05)	(0.04)	(0.08)	(0.06)	(0.05)	(0.06)
Common Language	0.61	0.74	0.61	0.40	0.85	0.71	0.89	0.23	0.59	0.60	0.59	0.60	0.68	0.56	0.39	0.40	0.47	0.66
	(0.06)	(0.07)	(0.14)	(0.10)	(0.06)	(0.06)	(0.12)	(0.14)	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.04)	(0.08)	(0.03)	(0.06)	(0.05)
Colonial Ties	0.40	0.27	0.50	0.78	0.48	0.24	0.68	0.78	0.55	0.51	0.35	0.31	0.52	0.53	0.69	0.57	0.39	0.42
	(0.11)	(0.08)	(0.12)	(0.12)	(0.09)	(0.10)	(0.12)	(0.13)	(0.09)	(0.13)	(0.08)	(0.09)	(0.09)	(0.07)	(0.10)	(0.08)	(0.08)	(0.07)
RTA	0.37	0.37	0.26	−0.09	−0.17	0.45	0.54	0.09	0.28	0.62	0.21	0.35	0.38	0.26	0.34	0.07	0.50	0.07
	(0.12)	(0.08)	(0.09)	(0.10)	(0.10)	(0.11)	(0.10)	(0.09)	(0.11)	(0.13)	(0.09)	(0.09)	(0.10)	(0.08)	(0.10)	(0.09)	(0.09)	(0.09)
No. of Obs.	1511580	1915140	853140	237180	244260	421260	127440	70800	3104580	428340	601800	1270860	782340	2000100	831900	746940	477900	669060

Notes: Standard errors, clustered by source country, are in parentheses. Coefficients reported are multiplied by $-\theta$, as the effects of the independent variable of interest and the trade elasticity are not separately identified by the gravity estimation. The implied percentage effect of each coefficient on the ad valorem tariff equivalent trade cost is $100 \times (e^{-coeff/\theta} - 1)$, where *coeff* is the reported coefficient.

Table A4: Multi-Sector Trade Costs Coefficient Estimates (Gamma PML)

(a) Aggregate Model

	Food	Textiles	Apparel	Leather	Wood	Paper	Printing	Petr/Coal	Chemicals	Rub/Plstc	Minerals	Bas. Metal	Fab. Metal	Computing	Electrical	Medical	Transport	Furniture
mean($\ln(d_i)$)	−2.30	−0.25	−1.58	−1.34	−2.49	−1.12	−2.85	−2.03	−0.92	−0.93	−2.14	−1.16	−1.82	−0.71	−0.88	−1.43	−1.69	−1.76
Distance	−1.23	−1.21	−1.28	−1.04	−1.25	−1.49	−1.57	−1.88	−1.28	−1.25	−1.39	−1.37	−1.23	−1.20	−1.23	−1.06	−1.05	−1.08
	(0.08)	(0.14)	(0.14)	(0.10)	(0.11)	(0.08)	(0.10)	(0.16)	(0.08)	(0.09)	(0.13)	(0.13)	(0.09)	(0.09)	(0.09)	(0.08)	(0.08)	(0.08)
Shared Border	0.42	0.45	−0.45	0.28	0.62	0.21	0.19	−0.43	0.06	0.25	−0.19	−0.16	0.32	−0.04	0.12	0.28	0.42	0.13
	(0.18)	(0.29)	(0.28)	(0.21)	(0.23)	(0.20)	(0.23)	(0.40)	(0.17)	(0.19)	(0.23)	(0.35)	(0.15)	(0.15)	(0.11)	(0.17)	(0.15)	(0.17)
Common Language	0.31	0.31	0.38	0.46	1.02	0.99	0.88	−0.34	0.82	1.13	0.83	−0.36	0.62	0.64	0.45	0.48	0.63	0.59
	(0.19)	(0.38)	(0.50)	(0.29)	(0.37)	(0.26)	(0.24)	(0.32)	(0.15)	(0.24)	(0.31)	(0.19)	(0.20)	(0.20)	(0.17)	(0.15)	(0.20)	(0.19)
Colonial Ties	0.60	0.76	0.87	0.62	−0.02	−0.01	0.26	0.53	0.26	0.40	0.54	0.81	0.51	0.47	0.58	0.52	0.10	0.62
	(0.25)	(0.36)	(0.42)	(0.44)	(0.29)	(0.21)	(0.20)	(0.36)	(0.18)	(0.24)	(0.24)	(0.23)	(0.16)	(0.13)	(0.17)	(0.17)	(0.17)	(0.16)
RTA	0.40	0.67	0.45	0.35	−0.21	0.52	−0.01	−0.06	0.22	0.81	0.93	0.31	0.79	0.42	0.51	0.31	0.56	0.79
	(0.14)	(0.36)	(0.34)	(0.24)	(0.29)	(0.16)	(0.26)	(0.25)	(0.11)	(0.20)	(0.30)	(0.27)	(0.19)	(0.16)	(0.17)	(0.19)	(0.19)	(0.19)
No. of Obs.	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540

(b) Product-Level Model

	Food	Textiles	Apparel	Leather	Wood	Paper	Printing	Petr/Coal	Chemicals	Rub/Plstc	Minerals	Bas. Metal	Fab. Metal	Computing	Electrical	Medical	Transport	Furniture
mean($\ln(d_i)$)	−2.76	−0.72	−2.09	−1.18	−2.55	−1.16	−3.44	−3.57	−1.00	−1.43	−2.61	−2.33	−2.25	−0.59	−0.67	−0.39	−1.21	−0.59
Distance	−1.05	−1.36	−0.96	−1.25	−1.23	−1.74	−1.28	−0.41	−1.28	−1.34	−1.29	−0.60	−1.30	−1.20	−0.99	−1.00	−1.13	−1.07
	(0.15)	(0.09)	(0.10)	(0.12)	(0.15)	(0.28)	(0.08)	(0.17)	(0.13)	(0.15)	(0.11)	(0.18)	(0.08)	(0.10)	(0.07)	(0.18)	(0.13)	(0.07)
Shared Border	1.45	−0.18	0.02	0.41	0.45	−0.33	0.04	1.83	0.04	0.18	0.33	1.65	−0.20	0.20	0.32	1.33	0.44	0.17
	(0.29)	(0.22)	(0.15)	(0.16)	(0.19)	(0.23)	(0.15)	(0.52)	(0.41)	(0.24)	(0.24)	(0.38)	(0.22)	(0.17)	(0.09)	(0.38)	(0.18)	(0.22)
Common Language	0.54	0.42	0.41	0.21	0.93	1.03	0.78	−0.85	0.95	0.33	0.31	0.42	0.62	0.55	1.48	−0.55	0.49	0.59
	(0.33)	(0.20)	(0.20)	(0.20)	(0.18)	(0.59)	(0.15)	(0.33)	(0.37)	(0.25)	(0.23)	(0.17)	(0.15)	(0.13)	(0.33)	(0.37)	(0.27)	(0.26)
Colonial Ties	0.02	0.66	0.81	0.62	0.53	1.06	0.32	2.07	0.27	0.53	0.54	0.12	0.13	0.33	0.46	1.20	0.39	1.53
	(0.33)	(0.30)	(0.29)	(0.33)	(0.29)	(0.40)	(0.19)	(0.19)	(0.25)	(0.20)	(0.20)	(0.21)	(0.19)	(0.29)	(0.20)	(0.24)	(0.21)	(0.26)
RTA	0.09	0.17	0.21	−0.32	−0.24	−1.28	0.33	−0.63	−0.25	0.41	0.33	−0.72	0.83	0.01	0.05	0.45	−0.12	0.40
	(0.31)	(0.24)	(0.16)	(0.25)	(0.22)	(0.37)	(0.17)	(0.32)	(0.32)	(0.18)	(0.19)	(0.31)	(0.28)	(0.24)	(0.36)	(0.44)	(0.25)	(0.26)
No. of Obs.	1511580	1915140	853140	237180	244260	421260	127440	70800	3104580	428340	601800	1270860	782340	2000100	831900	746940	477900	669060

Notes: Standard errors, clustered by source country, are in parentheses. Coefficients reported are multiplied by $-\theta$, as the effects of the independent variable of interest and the trade elasticity are not separately identified by the gravity estimation. The implied percentage effect of each coefficient on the ad valorem tariff equivalent trade cost is $100 \times (e^{-coeff/\theta} - 1)$, where *coeff* is the reported coefficient.

Table A5: Multi-Sector Trade Costs Coefficient Estimates (Poisson PML)

(a) Aggregate Model

	Food	Textiles	Apparel	Leather	Wood	Paper	Printing	Petr/Coal	Chemicals	Rub/Plstc	Minerals	Bas. Metal	Fab. Metal	Computing	Electrical	Medical	Transport	Furniture
mean(ln(d_i))	−3.00	−1.18	−1.83	−2.21	−2.98	−2.49	−4.21	−1.65	−2.44	−2.14	−3.35	−1.35	−2.78	−2.37	−2.84	−3.28	−3.27	−2.93
Distance	−0.94	−1.02	−0.98	−0.67	−1.12	−1.03	−0.71	−1.80	−0.86	−0.95	−1.09	−0.95	−0.92	−0.59	−0.67	−0.68	−0.53	−0.55
	(0.09)	(0.10)	(0.14)	(0.20)	(0.18)	(0.09)	(0.13)	(0.14)	(0.07)	(0.09)	(0.09)	(0.07)	(0.09)	(0.07)	(0.10)	(0.08)	(0.10)	(0.15)
Shared Border	0.45	0.31	0.50	0.51	0.81	0.45	0.45	0.52	0.12	0.51	0.57	0.56	0.55	0.26	0.38	0.07	0.49	0.63
	(0.12)	(0.14)	(0.14)	(0.16)	(0.21)	(0.12)	(0.18)	(0.25)	(0.11)	(0.10)	(0.08)	(0.11)	(0.10)	(0.12)	(0.15)	(0.17)	(0.10)	(0.12)
Common Language	0.39	0.69	0.80	0.43	−0.03	0.12	0.74	−0.01	0.47	0.42	0.44	0.25	0.49	0.36	0.15	0.38	0.41	0.58
	(0.14)	(0.15)	(0.26)	(0.24)	(0.18)	(0.13)	(0.20)	(0.26)	(0.15)	(0.09)	(0.13)	(0.15)	(0.11)	(0.12)	(0.13)	(0.19)	(0.14)	(0.18)
Colonial Ties	0.34	−0.17	−0.04	0.26	0.21	0.03	0.20	0.40	−0.10	0.09	0.00	0.26	0.14	0.06	−0.18	−0.05	−0.35	0.04
	(0.14)	(0.24)	(0.27)	(0.22)	(0.12)	(0.14)	(0.13)	(0.29)	(0.17)	(0.13)	(0.11)	(0.15)	(0.09)	(0.11)	(0.15)	(0.15)	(0.18)	(0.16)
RTA	0.66	0.95	0.73	0.73	0.37	0.81	1.27	−0.67	0.41	1.01	0.30	0.62	0.82	0.92	0.89	0.24	1.14	1.23
	(0.19)	(0.18)	(0.25)	(0.39)	(0.22)	(0.16)	(0.30)	(0.26)	(0.13)	(0.18)	(0.20)	(0.14)	(0.14)	(0.14)	(0.24)	(0.17)	(0.15)	(0.27)
No. of Obs.	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540

(b) Product-Level Model

	Food	Textiles	Apparel	Leather	Wood	Paper	Printing	Petr/Coal	Chemicals	Rub/Plstc	Minerals	Bas. Metal	Fab. Metal	Computing	Electrical	Medical	Transport	Furniture
mean(ln(d_i))	−2.37	−0.65	−1.38	−1.33	−2.21	−1.47	−3.99	−1.44	−2.05	−1.84	−2.87	−0.74	−2.37	−1.67	−2.22	−2.66	−2.53	−2.08
Distance	−1.25	−1.23	−1.16	−0.91	−1.59	−1.41	−0.77	−2.01	−0.86	−1.05	−1.13	−1.25	−1.01	−0.72	−0.81	−0.60	−0.63	−0.79
	(0.08)	(0.08)	(0.12)	(0.15)	(0.11)	(0.07)	(0.13)	(0.14)	(0.06)	(0.08)	(0.08)	(0.07)	(0.10)	(0.06)	(0.07)	(0.05)	(0.10)	(0.10)
Shared Border	0.52	0.34	0.41	0.43	0.75	0.41	0.49	0.48	0.21	0.49	0.68	0.51	0.54	0.23	0.29	0.26	0.51	0.65
	(0.11)	(0.11)	(0.14)	(0.15)	(0.15)	(0.11)	(0.17)	(0.29)	(0.09)	(0.09)	(0.10)	(0.07)	(0.09)	(0.12)	(0.15)	(0.15)	(0.10)	(0.13)
Common Language	0.55	0.75	0.76	0.67	0.21	0.37	0.78	0.07	0.39	0.51	0.62	0.23	0.59	0.40	0.29	0.43	0.38	0.36
	(0.11)	(0.15)	(0.27)	(0.21)	(0.19)	(0.10)	(0.20)	(0.28)	(0.11)	(0.08)	(0.11)	(0.11)	(0.11)	(0.10)	(0.11)	(0.11)	(0.10)	(0.15)
Colonial Ties	0.29	−0.10	−0.01	0.26	0.20	0.02	0.19	0.35	0.03	0.04	0.00	0.21	0.14	−0.02	−0.21	−0.04	−0.39	0.06
	(0.14)	(0.17)	(0.25)	(0.20)	(0.12)	(0.10)	(0.14)	(0.27)	(0.10)	(0.11)	(0.11)	(0.08)	(0.08)	(0.09)	(0.11)	(0.11)	(0.15)	(0.10)
RTA	0.82	0.85	0.85	0.80	0.11	0.88	1.22	−0.75	0.70	1.01	0.45	0.83	0.76	0.98	0.85	0.45	1.15	0.97
	(0.16)	(0.13)	(0.22)	(0.38)	(0.17)	(0.13)	(0.28)	(0.25)	(0.11)	(0.15)	(0.17)	(0.12)	(0.13)	(0.12)	(0.19)	(0.11)	(0.15)	(0.20)
No. of Obs.	1511580	1915140	853140	237180	244260	421260	127440	70800	3104580	428340	601800	1270860	782340	2000100	831900	746940	477900	669060

Notes: Standard errors, clustered by source country, are in parentheses. Coefficients reported are multiplied by $-\theta$, as the effects of the independent variable of interest and the trade elasticity are not separately identified by the gravity estimation. The implied percentage effect of each coefficient on the ad valorem tariff equivalent trade cost is $100 \times (e^{-coeff/\theta} - 1)$, where *coeff* is the reported coefficient.

Table A6: Multi-Sector Trade Costs Coefficient Estimates (Gaussian PML)

(a) Aggregate Model

	Food	Textiles	Apparel	Leather	Wood	Paper	Printing	Petr/Coal	Chemicals	Rub/Plstc	Minerals	Bas. Metal	Fab. Metal	Computing	Electrical	Medical	Transport	Furniture
mean($\ln(d_i)$)	−3.32	−1.87	−2.27	−4.07	−3.82	−3.17	−5.29	−2.35	−2.81	−2.29	−3.63	−1.69	−3.06	−2.53	−2.92	−3.38	−3.83	−3.78
Distance	−0.81	−1.05	−1.27	−0.02	−0.63	−0.82	−0.24	−1.50	−0.88	−0.76	−1.04	−0.83	−0.58	−0.47	−0.67	−0.81	−0.49	0.09
	(0.06)	(0.09)	(0.06)	(0.08)	(0.14)	(0.09)	(0.09)	(0.09)	(0.06)	(0.09)	(0.11)	(0.05)	(0.11)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)
Shared Border	0.21	0.25	0.60	0.42	1.15	0.34	0.40	1.13	−0.07	0.37	0.51	0.50	0.52	0.09	0.41	−0.23	0.28	0.83
	(0.06)	(0.07)	(0.11)	(0.08)	(0.19)	(0.05)	(0.11)	(0.27)	(0.05)	(0.06)	(0.06)	(0.08)	(0.08)	(0.08)	(0.14)	(0.10)	(0.06)	(0.07)
Common Language	0.33	0.69	1.30	−0.20	−0.39	0.07	0.89	0.45	0.73	0.31	0.43	0.41	0.43	0.39	−0.05	0.48	0.55	0.38
	(0.17)	(0.18)	(0.27)	(0.25)	(0.17)	(0.10)	(0.17)	(0.29)	(0.16)	(0.14)	(0.10)	(0.09)	(0.12)	(0.09)	(0.11)	(0.13)	(0.15)	(0.11)
Colonial Ties	0.37	−0.99	−0.91	−0.14	0.23	−0.13	−0.05	−0.20	−0.48	0.24	−0.11	0.01	0.29	0.06	−0.21	−0.20	−0.73	0.18
	(0.07)	(0.30)	(0.32)	(0.36)	(0.10)	(0.09)	(0.11)	(0.28)	(0.17)	(0.11)	(0.10)	(0.12)	(0.06)	(0.07)	(0.12)	(0.09)	(0.14)	(0.13)
RTA	1.10	1.09	−0.50	2.49	1.06	1.03	1.96	−1.22	0.26	1.83	0.31	0.78	1.91	1.47	1.05	0.08	1.14	2.65
	(0.14)	(0.15)	(0.24)	(0.30)	(0.23)	(0.12)	(0.18)	(0.41)	(0.17)	(0.27)	(0.25)	(0.13)	(0.25)	(0.15)	(0.20)	(0.14)	(0.17)	(0.17)
No. of Obs.	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540	3540

(b) Product-Level Model

	Food	Textiles	Apparel	Leather	Wood	Paper	Printing	Petr/Coal	Chemicals	Rub/Plstc	Minerals	Bas. Metal	Fab. Metal	Computing	Electrical	Medical	Transport	Furniture
mean($\ln(d_i)$)	−2.68	−2.97	−0.87	−1.82	−2.81	−2.13	−5.82	−2.22	−3.28	−2.19	−2.67	−1.32	−2.08	−2.13	−2.39	−2.38	−3.43	−2.63
Distance	−0.91	−1.53	−1.68	−0.41	−2.22	−1.32	−0.00	−1.69	−0.48	−0.89	−1.60	−1.24	−0.86	−0.73	−0.83	−0.96	−0.45	−0.68
	(0.09)	(0.12)	(0.10)	(0.15)	(0.17)	(0.11)	(0.14)	(0.08)	(0.07)	(0.11)	(0.08)	(0.12)	(0.13)	(0.09)	(0.11)	(0.09)	(0.06)	(0.09)
Shared Border	0.01	0.27	0.43	0.26	1.18	0.27	0.54	1.25	−0.03	0.32	0.57	0.34	0.52	−0.17	0.10	−0.19	0.44	0.60
	(0.06)	(0.07)	(0.12)	(0.10)	(0.17)	(0.08)	(0.12)	(0.35)	(0.07)	(0.07)	(0.09)	(0.10)	(0.08)	(0.09)	(0.14)	(0.18)	(0.12)	(0.07)
Common Language	0.71	0.69	1.02	1.00	−1.55	0.00	0.53	0.37	0.70	0.35	1.15	−1.00	0.52	0.51	0.10	1.29	0.94	−0.32
	(0.12)	(0.26)	(0.28)	(0.12)	(0.33)	(0.09)	(0.26)	(0.35)	(0.09)	(0.13)	(0.09)	(0.23)	(0.10)	(0.13)	(0.10)	(0.08)	(0.16)	(0.21)
Colonial Ties	0.68	−3.18	−0.44	−0.34	−0.06	−0.02	0.01	−0.05	−0.28	−0.03	−0.72	0.75	0.35	−0.41	−0.34	−0.88	−1.21	−0.09
	(0.17)	(0.55)	(0.37)	(0.35)	(0.15)	(0.08)	(0.25)	(0.23)	(0.07)	(0.11)	(0.15)	(0.20)	(0.06)	(0.08)	(0.13)	(0.07)	(0.16)	(0.12)
RTA	1.73	0.92	0.02	3.62	−0.60	0.86	2.35	−1.60	0.73	1.76	−0.52	1.18	1.85	1.29	1.27	0.24	1.09	1.97
	(0.33)	(0.30)	(0.35)	(0.48)	(0.22)	(0.16)	(0.32)	(0.41)	(0.12)	(0.22)	(0.32)	(0.16)	(0.22)	(0.18)	(0.25)	(0.14)	(0.13)	(0.15)
No. of Obs.	1511580	1915140	853140	237180	244260	421260	127440	70800	3104580	428340	601800	1270860	782340	2000100	831900	746940	477900	669060

Notes: Standard errors, clustered by source country, are in parentheses. Coefficients reported are multiplied by $-\theta$, as the effects of the independent variable of interest and the trade elasticity are not separately identified by the gravity estimation. The implied percentage effect of each coefficient on the ad valorem tariff equivalent trade cost is $100 \times (e^{-coeff/\theta} - 1)$, where *coeff* is the reported coefficient.

B Data

B.1 Trade Data

Product-level, bilateral trade data is taken from the U.N. Comtrade database classified into six-digit Harmonized System (HS) product codes. For 2003, the database contains trade flow data for 155 reporting countries classified according to the HS1996 classification system.[14] These 155 reporting countries report trade with an additional 74 non-reporting countries and territories. However, to ensure a complete trade flow matrix, only reporting countries are included in the sample.

For pairs of reporting countries, bilateral trade flows are typically reported in both directions by both countries. Trade flows reported by the exporting country were used because these flows are more likely to be consistent with the manufacturing output data, which is reported by the producing country, and because exports are typically reported "free on board", as opposed to "cost, insurance, and freight", and the former is consistent with the measure of trade flows in the model. This results in a dataset of 155 countries, 5,122 product codes, and 4,481,143 non-zero bilateral, product-level trade flow observations.

To combine the trade flow data with manufacturing output data, trade in non-manufacturing HS codes was dropped from the dataset. These are identified using the mapping from HS1996 codes to ISIC (revision 3) codes available from the U.N. Statistics Division.[15] This concordance was developed by the U.N. Statistics Division based on the mapping between the HS1996 classification and the CPC 1.0 classification and the mapping between the CPC 1.0 and the ISIC rev. 3. All HS codes not mapped to ISIC 2-digit industries 15-37 are dropped. This reduces the number of HS codes in the sample to 4,608 and the number of observations to 4,255,517.

B.2 Gravity Variables

The bilateral relationship variables used to estimate trade costs are from the Gravity dataset available from CEPII (see **?**). The variables used in the estimation are population-weighted distance (*distw*), whether countries share a border (*contig*), whether they share a common

[14]The year was chosen to maximize the number of countries for which both product-level trade data from Comtrade and manufacturing gross output data form INDSTAT were available. Of these 155 reporting countries, 105 originally reported their trade data using the HS2002 system, and the data was subsequently converted to the HS1996 system by Comtrade. To evaluate whether this conversion is likely to have affected the results of this paper, I also conducted the analysis using data for 2001, when nearly all reporting countries reported in the HS1996 system, and that the results were very similar.

[15]This is available for free download from the following url:
http://unstats.un.org/unsd/cr/registry/regdntransfer.asp?f=183.

official language (*comlang_off*), whether they have ever had a colonial link (*colony*), and whether they are currently members of a common regional trade agreement (*rta*).

B.3 Manufacturing Output

Data on gross manufacturing output is is obtained from three sources. Where it is available, the data is taken from the OECD STAN database. For countries not in this database, data is obtained from the Industrial Statistics Database (INDSTAT4), 2011 Edition, CD-ROM available from the United Nations Industrial Development Organization. Where data for 2003 is not available but is available for other years both before and after 2003, the log of 2003 output is taken as the linear interpolation between the values of log output from the most recent year pre- and post-2003. Where no data is available from either of these sources, gross manufacturing output is imputed from total manufacturing value added obtained from the World Development Indicators database of the World Bank. Manufacturing value added is scaled up by a factor of 3.04 based on a cross-sectional regression of gross output on value added with no constant term, which has an R^2 of 0.99.

Industry-level data on gross manufacturing is also obtained from the STAN database, where available, and the INDSTAT4 database, otherwise. Both sources report data using the ISIC Revision 3 system. STAN reports data at the 2-digit industry level, and INDSTAT4 at the 4-digit level. However, in the INDSTAT database, many countries report data using combinations of categories, and many appear to report data for related industries using either one or the other industry code but not both. In addition different countries report data only in more aggregated categories. Because of such issues, the data was aggregated to the 2-digit level, and several 2-digit industries were combined. Table A2 lists the industries that are used, their definitions, the number of 6-digit HS-1996 codes within each industry, and the industry's share in total world manufacturing expenditure. As with the aggregate data, industry-level output data was interpolated for observations for which data was not available for 2003 but was available for years before and after 2003.

B.4 Constructing the Sample

To be included in the sample, data must be available for a country from the Comtrade database and at least one of the STAN, INDSTAT, or WDI databases. Beginning with the 155 countries that make up the sample of product-level trade data, lack of manufacturing output data reduces the sample size to 141 countries. To avoid problems related to entrepot trade, China, Hong Kong, and Macao are merged into a single country. There were also several other cases in which there were apparent problems of entrepot trade – i.e. reported exports exceeded reported gross output – which resulted in 7 countries being dropped from the sample.[16] These two steps together reduced the sample to 132 countries. Once the trade and manufacturing data were merged, domestic absorbtion of domestic manufacturing output, X_{ii}, was then calculated as total manufacturing output minus total manufacturing exports to all countries (including non-reporters), and total manufacturing absorbtion, X_i, was calculated as X_{ii} plus total imports from countries in the sample, yielding an internally consistent bilateral trade flow matrix.

In constructing the sample of industry-level output and trade flows, great care was taken to ensure the quality and consistency of the data, which included inspecting the data line-by-line for many countries in the sample. Countries with significant discrepancies, for instance between the sum of industry-level output and reported total output, were excluded from the sample. Even after excluding these countries, for about 12% of observations, reported exports exceeded reported gross output. For these observations, output was imputed based on the value of exports and the country's overall ratio of exports to output for the entire manufacturing sector. When this resulted in an imputed measure of industry-level output that exceeded the reported value by more than 30%, the country was removed from the sample. This resulted in a final sample of 60 countries, 18 manufacturing industries, and 2,360,978 observed product-level bilateral trade flows. The set of countries that make up the aggregate and industry-level samples, along with the source of output data, is reported in Table A1.

[16]The excluded counties are Armenia, Belgium, Guyana, Luxembourg, Mali, Mongolia, and Singapore.