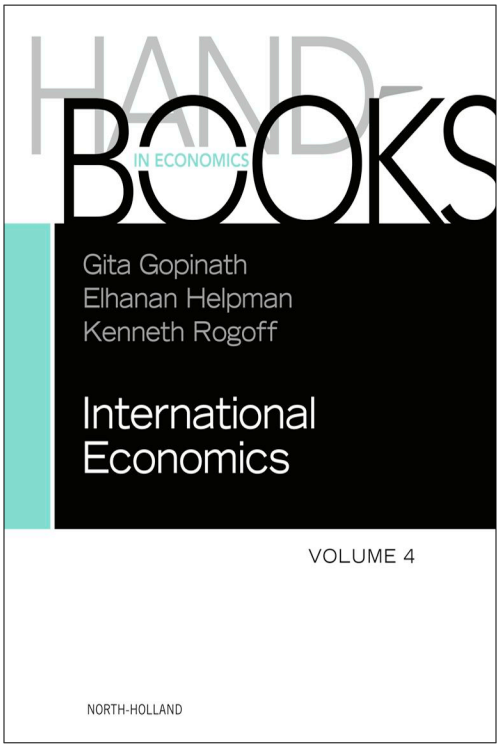

From Keith Head and Thierry Mayer, Gravity Equations: Workhorse,Toolkit, and Cookbook. In: Elhanan Helpman, Kenneth Rogoff and Gita Gopinath, editor, Handbook of International Economics. Vol 4, Oxford: BV, 2015, p. 131-195.
ISBN: 978-0-444-54314-1

CHAPTER 3

Gravity Equations: Workhorse,Toolkit, and Cookbook*

Keith Head[*,†] **and Thierry Mayer**[†,‡,§]

[*]Sauder School of Business, University of British Columbia, Canada
[†]Centre for Economic Policy Research, London, UK
[‡]Sciences-Po, Paris, France
[§]Centre d'études prospectives et d'informations internationales, France

Abstract

This chapter focuses on the estimation and interpretation of gravity equations for bilateral trade. This necessarily involves a careful consideration of the theoretical underpinnings since it has become clear that naive approaches to estimation lead to biased and frequently misinterpreted results. There are now several theory-consistent estimation methods and we argue against sole reliance on any one method and instead advocate a toolkit approach. One estimator may be preferred for certain types of data or research questions but more often the methods should be used in concert to establish robustness. In recent years, estimation has become just a first step before a deeper analysis of the implications of the results, notably in terms of welfare. We try to facilitate diffusion of best-practice methods by illustrating their application in a step-by-step cookbook mode of exposition.

Keywords

Bilateral trade, Heterogeneous firms, Distance, Borders, Trade cost elasticity, Poisson

JEL classification codes

F1

* The chapter has a companion website, https://sites.google.com/site/hiegravity/, with an appendix, Stata® code, and related links. We thank Leo Fankhänel and Camilo Umana for outstanding assistance with the programming and meta-analysis in this chapter, Soledad Zignago for great help with providing and understanding subtleties of some of the data used, and Julia Jauer for her update of the gravity data. Scott Baier, Sebastian Sotelo, and João Santos Silva generously provided computer code. Andres Rodríguez-Clare answered many questions we had about welfare calculations but is not responsible, of course, for any mistakes we may have made. Arnaud Costinot, Gilles Duranton, Thibault Fally, Mario Larch, Marc Melitz, Gianmarco Ottaviano, João Santos Silva, and Daniel Trefler made very useful comments on previous drafts. We are especially grateful to Jose de Sousa: his careful reading identified many necessary corrections in an early draft. Participants at presentations at the Hitotsubashi GCOE Conference on International Trade and FDI 2012, National Bank of Belgium, and Clemson University also contributed to improving the draft. Finally, we thank our discussants at the handbook conference, Rob Feenstra and Jim Anderson, for many helpful suggestions. We regret that because of limitations of time and space, we have not been able to fully respond to all of the valuable suggestions we received. This research has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) Grant Agreement no. 313522.

Handbook of International Economics, Volume 4
ISSN 1573-4404, http://dx.doi.org/10.1016/B978-0-444-54314-1.00003-3

1. INTRODUCTION

As the name suggests, gravity equations are a model of bilateral interactions in which size and distance effects enter multiplicatively. They have been used as a workhorse for analyzing the determinants of bilateral trade flows for 50 years since being introduced by Tinbergen (1962). Krugman (1997) referred to gravity equations as examples of "social physics," the relatively few law-like empirical regularities that characterize social interactions.[1] Over the last decade, concentrated efforts of trade theorists have established that gravity equations emerge from mainstream modeling frameworks in economics and should no longer be thought of as deriving from some murky analogy with Newtonian physics. Meanwhile empirical work—guided in varying degrees by the new theory—has proceeded to lay down a raft of stylized facts about the determinants of bilateral trade. As a result of recent modeling, we now know that gravity estimates can be combined with trade policy experiments to calculate implied welfare changes.

This chapter focuses on the estimation and interpretation of gravity equations for bilateral trade. This necessarily involves a careful consideration of the theoretical underpinnings since it has become clear that naive approaches to estimation lead to biased and frequently misinterpreted results. There are now several theory-consistent estimation methods and we argue against sole reliance on any one method and instead advocate a toolkit approach. One estimator may be preferred for certain types of data or research questions but more often the methods should be used in concert to establish robustness. In recent years, estimation has become just a first step before a deeper analysis of the implications of the results, notably in terms of welfare. We try to facilitate diffusion of best-practice methods by illustrating their application in a step-by-step cookbook mode of exposition.

1.1. Gravity Features of Trade Data

Before considering theory, we use graphical displays to lay out the factual basis for taking gravity equations seriously. The first key feature of trade data that mirrors the physical gravity equation is that exports rise proportionately with the economic size of the destination and imports rise in proportion to the size of the origin economy. Using GDP as the economy size measure, we illustrate this proportionality using trade flows between Japan and the European Union. The idea is that the European Union's area is small enough and sufficiently far from Japan that differences in distance to Japan can be ignored. Similarly because the EU is a customs union, each member applies the same trade policies on Japanese imports. Japan does not share a language, religion, currency, or colonial history with any EU members either.

[1] Other examples of social physics include power function distributions thought to characterize incomes, firm and city sizes, and network linkages.

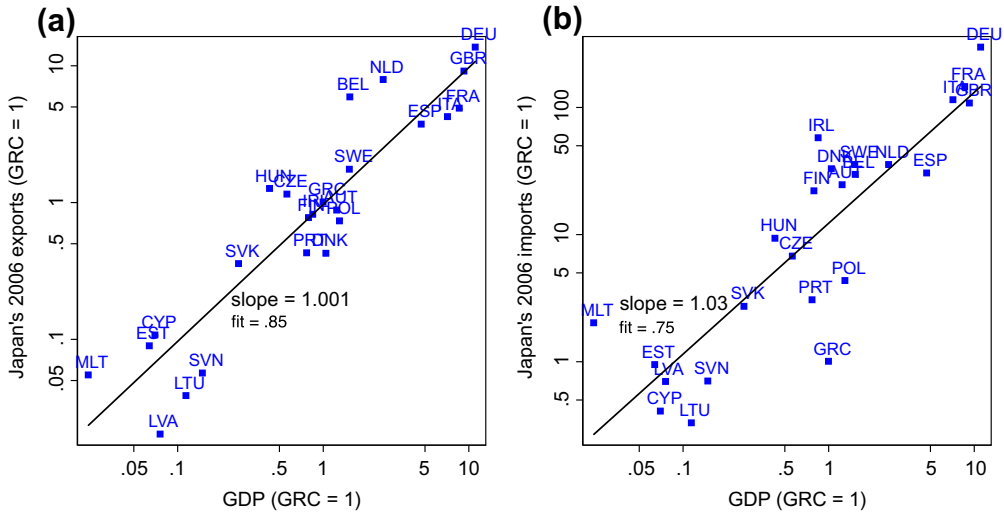

Figure 3.1 Trade is Proportional to Size; (a) Japan's Exports to EU, 2006; (b) Japan's Imports from EU, 2006. GRC: Greece

Figure 3.1(a) shows Japan's bilateral exports on the vertical axis and (b) shows its imports. The horizontal axes of both figures show the GDP (using market exchange rates) of the EU trade partner. The trade flows and GDPs are normalized by dividing by the corresponding value for Greece (a mid-size economy).[2] The lines show the predicted values from a simple regression of log trade flow on log GDP. For Japan's exports, the GDP elasticity is 1.00 and it is 1.03 for Japan's imports. The near unit elasticity is not unique to the 2006 data. Over the decade 2000–2009, the export elasticity averaged 0.98 and its confidence intervals always included 1.0. Import elasticities averaged a somewhat higher 1.11 but the confidence intervals included 1.0 in every year except 2000 (when 10 of the EU25 had yet to join). The gravity equation is sometimes disparaged on the grounds that any model of trade should exhibit size effects for the exporter and importer. What these figures and regression results show is that the size relationship takes a relatively precise form—one that is predicted by most, but not all, models.

Figure 3.2 illustrates the second key empirical relationship embodied in gravity equations—the strong negative relationship between physical distance and trade. Since we have just seen that GDPs enter gravity with a coefficient very close to one, one can pass GDP to the left-handside, and show how bilateral imports or exports as a fraction of GDP varies with distance. Panels (a) and (b) of Figure 3.2 graph recent export and import data from France. These panels show deviations from the distance effect associated

[2] The trade data come from International Monetary Fund (IMF) Direction of Trade Statistics (DOTS) and the GDPs come from World Development Indicators (WDI). The web appendix provides more information on sources of gravity data.

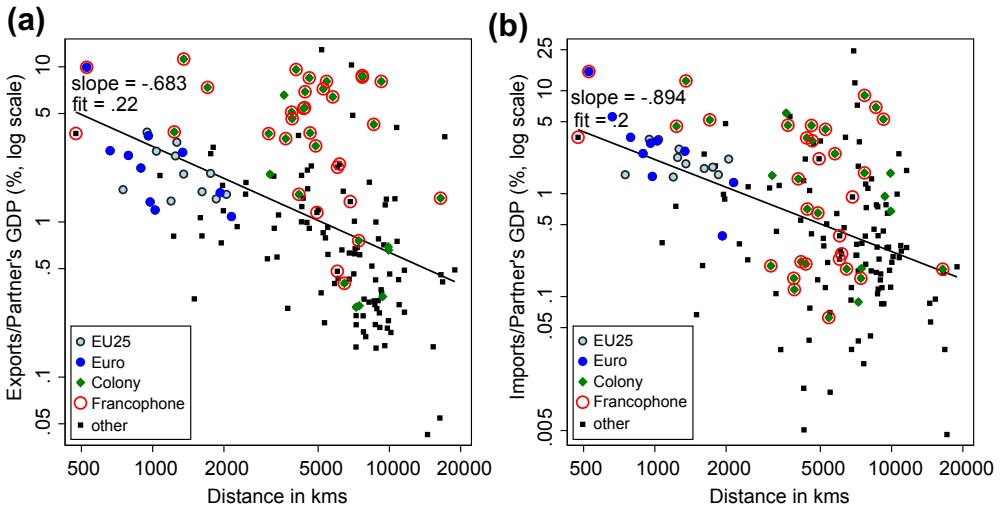

Figure 3.2 Trade is Inversely Proportional to Distance; (a) France's Exports (2006); (b) France's Imports (2006)

with Francophone countries, former colonies, and other members of the EU or of the Eurozone. The graph expresses the "spirit" of gravity: it identifies deviations from a *benchmark* taking into account GDP proportionality and systematic negative distance effects. Those deviations have become the subject of many separate investigations.

This chapter is mainly organized around topics with little attention paid to the chronology of when ideas appeared in the literature. But we do not think the history of idea development should be overlooked entirely. Therefore in the next section we give our account of how gravity equations went from being nearly ignored by trade economists to becoming a focus of research published in the top general interest journals.

1.2. A Brief History of Gravity in Trade

While economists have been estimating gravity equations on bilateral trade data since Tinbergen (1962), this work lay outside of the mainstream of trade research until 1995. One of the barriers to mainstream acceptance was the lingering perception that gravity equations were more physics analogy than economic analysis. In the first volume of this handbook series, Deardorff (1984, p. 503) characterized the "theoretical heritage" of gravity equations as being "dubious." Given the traditional importance of theory in the field of international trade, this was damning criticism. It was not entirely fair to the economists who had begun the work of grounding the gravity equation in theory long before. Savage and Deutsch (1960) contains a multiplicative model of bilateral trade published two years before the empirical work of Tinbergen (1962). Although that model was purely probabilistic, Anderson (1979) set forth a conventional economic model of

gravity. The model did not penetrate the consciousness of trade economists. Leamer and Levinsohn (1995, fn. 13), write "An attempt to give a theoretical foundation by Anderson (1979) is formally fruitful but seems too complex to be part of our everyday toolkit."

By contrast with 1995, gravity is now an integral and important part of international trade. We view its recent inclusion as a core element of the field as being articulated in three distinct steps. Firstly, the "admission" wherein researchers realized there was a surprisingly large amount of missing trade, and admitted that gravity was one way to measure and explain it. Then came the "multilateral resistance/fixed effects revolution," a burst of papers that established the relationship between fixed effects in gravity and underlying theories with origins as varied as Ricardo, monopolistic competition, and Armington. The final step was one of "convergence" of the gravity and heterogeneous firms literatures.

Admission (1995): 1995 was a very important year for gravity research. In that year Trefler (1995) introduced the idea of "missing trade." A key empirical problem for the Heckscher-Ohlin-Vanek (HOV) model is that it predicts much higher trade in factor services than is actually observed. Trefler invoked "home bias" rather than distance to explain missing trade but his work pointed to the importance of understanding the impediments to trade. In a *Handbook of International Economics* chapter, Leamer and Levinsohn (1995) pointed out that gravity models "have produced some of the clearest and most robust findings in economics. But paradoxically they have had no effect on the subject of international economics." They asked provocatively, "Why don't trade economists 'admit' the effect of distance into their thinking?" Their explanation was that "human beings are not disposed toward processing numbers, and empirical results will remain unpersuasive if not accompanied by a graph." Their solution was to produce a version of Figure 3.2(a) for Germany.[3] Krugman's (1995) chapter in the same handbook also considers the role of remoteness and intuitively states why bilateral distance cannot be the only thing that matters as in the standard gravity equation (see the end of its Section 3.1.2). Krugman's thought experiment of moving two small countries from the middle of Europe to Mars provides the intuition for why we need the multilateral resistance (MR) terms that Anderson (1979) originated and Anderson and van Wincoop (2003) popularized.

One irony of the history of the gravity equation is that trade economists "discovered" the empirical importance of geographic distance and national border just as some prominent journalists and consultants had dismissed these factors as anachronisms. Thus the business press was proclaiming the "borderless world," "the death of distance," and the "world is flat" while empirical research was categorically demonstrating the opposite. McCallum (1995) used the gravity equation and previously unexploited data on interprovincial trade to decisively refute the notion that national borders had lost their

[3] Forty years earlier Isard and Peck (1954) had offered the same graphical device to complain about the lack of consideration for distance (space in general) in international trade theory.

economic relevance. McCallum's article not only showed the usefulness of the gravity equation as a framework for estimating the effects of trade integration policies, but also launched a literature attempting to understand "border effects." While we now think of Anderson and van Wincoop (2003) as being first and foremost a paper about the gravity methodology, it was framed as a resolution to the puzzle McCallum had exposed.

The MR/fixed effects revolution (2002–2004): With the publication of Eaton and Kortum (2002), and Anderson and van Wincoop (2003), the conventional wisdom that gravity equations lacked micro-foundations was finally dismissed. Since neither model relied on imperfect competition or increasing returns, there was no longer a reason to believe that gravity equations should only apply to a subset of countries or industries. Perhaps most importantly, these papers pointed the way toward estimation methods that took into account the structure of the models. In 2004, it became clear, with the chapter by Feenstra (2004) and the article by Redding and Venables (2004), that importer and exporter fixed effects could be used to capture the multilateral resistance terms that emerged in different theoretical models. The combination of being consistent with theory and quite easy to implement (in most cases) leads to rapid adoption in empirical work.

Convergence with the heterogeneous firms literature (2008): 2008 was the third pivotal year for research on gravity as it saw the publication of three papers—Chaney (2008), Helpman et al. (2008), and Melitz and Ottaviano (2008)—that united recent work on heterogeneous firms with the determination of bilateral trade flows. In this final step, the toolkit nature of gravity again appeared as it became a useful tool to measure the new distinction between intensive and extensive margins of adjustment to trade shocks (Bernard et al., 2007; Mayer and Ottaviano, 2007; Chaney, 2008). The "merger" of the two literatures implied changes to the way gravity equations should be estimated and to how the estimated coefficients should be interpreted. It was also a sign of the rising intellectual stature of the gravity equation that the three 2008 papers make a point of showing that their heterogeneous firms models are compatible with gravity.

Clearly, the useful tool of the early 1990s had by then become an object respected by theorists, who even tried to add to the sophistication of it. In a field that has historically been so dominated by pure theory, this sounds like the definitive recognition, which has recently been expanded further, by incorporating gravity as a central component of the theory and measurement of welfare gains from trade (Chapter 4 by Costinot and Rodriguez-Clare in this handbook probably being the best illustration).

Because none of this would probably have happened if the theoretical underpinnings of gravity had not been made clearer, we start with those in Section 2. We then turn in Section 3 to the estimation issues, to cover the many existing practices and give our views on best practice. Section 4 focuses on what has been and probably will remain the main use of gravity: a tool for quantifying the impacts of trade policies. This section focuses particularly on what recent advances mean for the implementation of those evaluations.

We finish with Section 5, covering areas of current, mostly unsettled research and progress: the frontiers of gravity equations, before concluding.

2. MICRO-FOUNDATIONS FOR GRAVITY EQUATIONS

> *"The equation has...gone from an embarrassing poverty of theoretical foundations to an embarrassment of riches!"*
>
> *Frankel et al. (1997, p. 53)*

As the quote above suggests, the conventional wisdom that gravity equations had no sound theoretical underpinnings has been forcefully dismissed. Indeed, in the 15 years following the Frankel's comment, the "embarrassment of riches" has become substantially more acute. It seems reasonable to credit the empirical success of gravity equations with attracting the attention of theorists. This section of the chapter will proceed by first defining what we mean when we use the term gravity equation and then setting out the theories that conform with the definitions. We close the theory section by summarizing successful efforts to transfer the gravity modeling techniques to interactions beyond trade in goods.

2.1. Three Definitions of the Gravity Equation

While the term gravity equation has been used to refer to a variety of different specifications of the determinants of bilateral trade, we consider three definitions to be particularly useful.

Definition 1. General gravity comprises the set of models that yield bilateral trade equations that can be expressed as

$$X_{ni} = GS_i M_n \phi_{ni}. \tag{1}$$

The S_i factor represents "capabilities" of exporter i as a supplier to all destinations. M_n captures all characteristics of destination market n that promote imports from all sources. Bilateral accessibility of n to exporter i is captured in $0 \leq \phi_{ni} \leq 1$: it combines trade costs with their respective elasticity to measure the overall impact on trade flows. Lastly, G can be termed the "gravitational constant," although it is only held constant in the cross-section.

Definition 1 has two important features. The most obvious one is the insistence that each term enters multiplicatively. A second important feature is that this definition requires that third-country effects, if there are any, must be mediated via the i and n multilateral terms.[4] The multiplicative form derives from the original analogy with the gravity equation in physics. It is convenient because, after taking logs, equation (1) can be

[4] For example, ϕ_{nj} can influence X_{ni} but only by changing S_i or M_n. Thus it would be impossible for a trade agreement between j and n to reduce n's imports from i but leave all its other imports unchanged.

estimated by regressing log exports on exporter and importer fixed effects and a vector of bilateral trade costs variables. However, the multiplicative form is not necessary for estimation. Both the linear demand system used by Ottaviano et al. (2002) or the translog form used by Feenstra (2003) and Novy (2013) are relatively straightforward to estimate despite not being multiplicatively separable in the S_i, M_n, and ϕ_{ni} terms, and therefore not obeying Definition 1.[5] Thus the main reason to insist on the multiplicative form in the definition of gravity is historical usage. It is therefore possible that future work would abandon the multiplicative form and redefine gravity to allow other functional forms.

By imposing a small set of additional conditions, one can express the exporter and importer terms in equation (1)—S and M—as functions of observables, leading to a second way to define the gravity equation.

Definition 2. Structural gravity comprises the subset of general gravity models in which bilateral trade is given by

$$X_{ni} = \underbrace{\frac{Y_i}{\Omega_i}}_{S_i} \underbrace{\frac{X_n}{\Phi_n}}_{M_n} \phi_{ni}, \tag{2}$$

where $Y_i = \sum_n X_{ni}$ is the value of production, $X_n = \sum_i X_{ni}$ is the value of the importer's expenditure on all source countries, and Ω_i and Φ_n are "multilateral resistance" terms defined as

$$\Phi_n = \sum_\ell \frac{\phi_{n\ell} Y_\ell}{\Omega_\ell} \quad \text{and} \quad \Omega_i = \sum_\ell \frac{\phi_{\ell i} X_\ell}{\Phi_\ell}. \tag{3}$$

Definition 2 corresponds, as discussed below, to a surprisingly large set of models. It can be validated against alternatives, by comparing estimated fixed effects to the theoretical counterparts. Because the Φ and Ω terms can be solved for a given set of trade costs, Definition 2 allows for a more complete calculation of the impacts of trade costs changes, something we come back to in Section 4.3.

Structural gravity can be estimated at the aggregate or industry level.[6] At the aggregate level one should measure Y_i as gross production (not value-added) of traded goods (assuming X_{ni} is merchandise trade) and X_n should be apparent consumption of goods (production plus imports minus exports). However, in practice GDP is often used as a proxy for both Y_i and X_n.[7]

Definition 3. Naive gravity equations express bilateral trade as

$$X_{ni} = G Y_i^a Y_n^b \phi_{ni}. \tag{4}$$

[5] As we will see later, *heterogeneous firms* versions of the linear and translog models *do* fit equation (1) under Pareto-distributed heterogeneity.

[6] In a series of papers Anderson and Yotov (2010a,b, 2012) estimate structural gravity at the industry level, arguing that this practice reduces aggregation bias.

[7] The web appendix provides details on data sources for aggregate and industry level Y_i and X_n.

Definition 3 is pedagogically useful, was long viewed as empirically successful, and contains the important insight that bilateral trade should be roughly proportional to the product of country sizes. The naive gravity is at once more general and more restrictive than definitions derived from theory. The presence of $a \neq b \neq 1$ is a generalization that has been included in estimation starting with Tinbergen (1962). However, as we shall see, most theories predict unit GDP elasticities and Figures 3.1(a) and (b) suggest the data appear happy to comply (to a reasonable approximation). On the other hand, as pointed out by Krugman (1995), theoretical justifications for Definition 3 impose the implausible restriction that ϕ_{ni} is a constant. This cancels the need for multilateral terms, but cannot be reconciled with the overwhelming evidence that trade costs do vary across bilateral pairs. Baldwin and Taglioni (2007) refer to the omission of $1/(\Omega_i \Phi_n)$ in Definition 3 as the "gold medal mistake" of gravity equations, almost universally characterizing papers appearing before Anderson and van Wincoop (2003).

In the next subsections, we will consider the assumptions underlying structural gravity, before turning to detailed micro-foundations of this relationship. Then we will consider a small number of recent models that fit Definition 1, but violate Definition 2.

2.2. Assumptions Underlying Structural Gravity

Structural gravity relies on two important conditions. The first governs spatial allocation of expenditure for the importer. The second imposes market-clearing for the exporter.

Let i be the origin (exporter) and n be the destination. Importer n's total expenditures, X_n, can be thought of as the "pie" to be allocated. The share of the pie allocated to country i is denoted π_{ni}. As an accounting identity we have

$$X_{ni} = \pi_{ni} X_n, \tag{5}$$

where $\pi_{ni} \geq 0$ and $\sum_i \pi_{ni} = 1$.

The critical requirement is that π_{ni} can be expressed in the following multiplicatively separable form:

$$\pi_{ni} = \frac{S_i \phi_{ni}}{\Phi_n}, \quad \text{where} \quad \Phi_n = \sum_\ell S_\ell \phi_{n\ell}. \tag{6}$$

The definition of Φ_n as the accessibility-weighted sum of the exporter capabilities is required to ensure that the budget allocation shares sum to one. Φ_n therefore measures the set of opportunities of consumers in n or, equivalently, the degree of competition in that market. We will see below that a wide range of different micro-foundations yield equation (6). While (6) might seem an innocuous assumption, it requires that budget shares should be independent of income. This rules out several demand systems, such as quasi-linear models with outside goods. Those models might still fit the conditions of general gravity, as is the case for Melitz and Ottaviano (2008).

A second accounting identity holds that the sum of i's exports to all destinations—including i—equals the total value of i's production, which in aggregate is just Y_i.

$$Y_i = \sum_n X_{ni} = S_i \sum_n \frac{\phi_{ni} X_n}{\Phi_n}. \tag{7}$$

Solving for S_i, one obtains

$$S_i = \frac{Y_i}{\Omega_i}, \quad \text{where} \quad \Omega_i = \sum_\ell \frac{\phi_{\ell i} X_\ell}{\Phi_\ell}. \tag{8}$$

The Ω term is familiar in economic geography as an index of market potential or access (see Redding and Venables, 2004; Head and Mayer, 2004b or Hanson, 2005). Relative access to individual markets is measured as $\phi_{\ell i}/\Phi_\ell$. Hence, Ω_i is an expenditure-weighted average of relative access. Substituting (8) into equation (6) gives

$$\Phi_n = \sum_\ell \frac{\phi_{n\ell} Y_\ell}{\Omega_\ell}, \tag{9}$$

which, once plugged back into (5), provides (2):

$$X_{ni} = \frac{Y_i}{\Omega_i} \frac{X_n}{\Phi_n} \phi_{ni}.$$

Anderson and van Wincoop (2003) assume $X_i = Y_i$ (balanced trade) and $\phi_{ni} = \phi_{in}$ (symmetric trade costs), which implies that $\Phi_i = \Omega_i$. This in turn would imply $S_i = M_i$ in the general gravity equation, leading to a symmetric gravity equation.

2.3. Main Variants of Gravity for Trade

The next step is to show the range of established theories that comply with the structural gravity assumptions. All the specifications we consider specify trade costs (transport for goods, travel for many services, search, and other transaction costs for both goods and services) using the iceberg form. Under this assumption, $\tau_{ni} - 1$ is the *ad valorem* tariff equivalent of all trade costs. Most models work with a single factor of production, denoted L. Factor income is w, and hence GDP is given by $X_n = Y_n = w_n L_n$. Below we specify the different set of assumptions characterizing each of the models, and summarize the theoretical content of S_i, ϕ_{ni}, and M_n in Table 3.1 (see page 149).

We group the models under the category "demand-side" and "supply-side." In the demand-side models the exogenous wage combined with constant returns to scale or constant markups neutralizes the supply side of the model. The models we call supply-side derivations also have demand sides but distributional assumptions used in these models (Fréchet or Pareto) cause the demand-side terms to be eliminated from the final formulation.

2.3.1. Demand-Side Derivations

CES National Product Differentiation (Anderson-Armington)

The earliest "modern" derivation of the gravity equation for trade is Anderson (1979). As in Armington (1969), each country is the unique source of each product (there is National Product Differentiation, NPD). Consumers in country n consume q_{ni} units of the product from country i. Utility exhibits a constant elasticity of substitution (CES), $\sigma > 1$, over all the national products:

$$U_n = \left(\sum_i (A_i q_{ni})^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}. \tag{10}$$

A_i is a utility shifter that can be thought of as an index of the quality of country i's product.[8] Simple maximization of (10) under budgetary constraint provides optimal demand for each variety. The two terms of equation (6) are then given by $S_i = A_i^{\sigma-1} w_i^{1-\sigma}$, and $\phi_{ni} = \tau_{ni}^{1-\sigma}$.

Following Okawa and Van Wincoop (2010, Section 3.1), we can modify the Armington utility function, adding consumption of a homogeneous "outside" good, here denoted q_n^0, to equation (10). For each differentiated good i has sales of $(w_i/A_i)^{1-\sigma}$. Note that demand for the differentiated goods does not depend on income of country n; all residual income is spent on the homogeneous good. The resulting gravity equation still has $S_i = A_i^{\sigma-1} w_i^{1-\sigma}$, and $\phi_{ni} = \tau_{ni}^{1-\sigma}$ but $M_i = 1$, because $X_n/\Phi_n = 1$ (assuming X_n corresponds to expenditures on the differentiated goods only). Adding an outside good that enters utility linearly therefore leads to a specification that fits the general definition for gravity but not the one we call "structural gravity."[9]

CES Monopolistic Competition (Dixit-Stiglitz-Krugman)

The gravity equation based on standard symmetric Dixit-Stiglitz-Krugman (DSK) monopolistic competition assumptions was derived by multiple authors.[10] It assumes that each country has N_i firms supplying one variety each to the world from a home-country production site. Utility features a constant elasticity of substitution, denoted σ, between all varieties available in the world. Dyadic accessability is given by $\phi_{ni} = \tau_{ni}^{1-\sigma}$. The exporter attribute is given by $S_i = N_i w_i^{1-\sigma}$, where the difference compared to the NPD model is that the N_i term replaces $A_i^{\sigma-1}$. Thus the exporter attribute reflects the monopolistic competition among the symmetric varieties in the DSK model and competitively supplied national varieties in the NPD model. Note that prices are also different

[8] Anderson and van Wincoop (2003) use $\beta_i = 1/A_i$ in their formulation. We prefer the one specified above because it allows us to think of A_i as the attractiveness of country i's product, whereas Anderson and van Wincoop's (2003) β_i is an *inverse* measure of quality.

[9] This is an unfortunate aspect of the terminology but we could not find a suitable alternative.

[10] One early derivation based on Krugman (1979) is contained in the unpublished paper of Wei (1996).

since they are a constant positive markup over marginal costs in DSK, and just equal to marginal cost in NPD.

While the Dixit-Stiglitz model is usually interpreted as firms supplying differentiated goods to consumers, the fact that the majority[11] of trade involves intermediates suggests the benefits of generalizing to that case. If we follow Ethier (1982) in assuming that each firm produces a differentiated variety of intermediate input, the S_i, M_n, and ϕ_{ni} terms remain the same.

CES Demand with CET Production

The earliest derivation of a gravity equation using monopolistic competition (MC) of the Dixit-Stiglitz form is Bergstrand (1985). Bergstrand used a more general set of functional forms that were not retained in later work. In particular, he allowed for a nested structure in which domestic varieties are closer substitutes for each other than are foreign varieties. Bergstrand also generalized the production side to allow for the possibility that output might not be transferable to the export sector on a one-for-one basis. Instead he allows for a "constant elasticity of transformation" (CET). The idea is that output to one destination cannot be costlessly transformed into output for a different destination. The elasticity of transformation is denoted γ and ranges from 0, where it is impossible to reallocate output, to infinity, in which case transformation is costless.

Here we follow Baier and Bergstrand (2001) in assuming a finite CET, while retaining the single-layer CES. This specification still yields structural gravity with

$$S_i = L_i w_i^{\frac{\gamma(1-\sigma)}{\sigma+\gamma}} \quad \text{and} \quad \phi_{ni} = \tau_{ni}^{\frac{(1+\gamma)(1-\sigma)}{\sigma+\gamma}}.$$

The model has $M_n = X_n/\Phi_n$ and therefore has a unit income elasticity if X_n is proportional to income.[12] The wage and trade costs elasticities now include the supply-side CET, γ, and the wage elasticity is $\gamma/(1+\gamma)$ times the trade elasticity.

Baier and Bergstrand (2001) motivate the finite CET by arguing that it could reflect distribution costs of entering foreign markets. We believe it is better to think of it as a way of generating upward sloping marginal costs of serving each market. This has the effect of lowering both the wage and trade elasticities. That is, trade is less responsive to wages and trade costs than it would be if only the demand parameter σ mattered.

[11] Chen et al. (2005) construct the share of intermediates in total trade for 10 OECD countries using input-output tables in various years between 1968 and 1998. The US share averages 50% while the other countries have higher averages, with Japan above 80% until the 1990s.

[12] Equation (12) of Bergstrand (1985) gives the appearance that the model predicts a less than unit elasticity but this is because it retains the price index. After solving for the price index the elasticity is predicted to be one which implies that the estimated income elasticity cannot be used to back out γ.

Heterogeneous Consumers

The taste for variety present in the CES utility functions may be plausible in some contexts but it does not fit products like laundry detergents or (except for the very rich) passenger cars. In those and many other cases, the natural way to think about consumer choice is that the large variety of products purchased results from consumers making different decisions. If they face the same prices, then the different selections result from a variety of tastes.[13] Anderson et al. (1992) show that two strong functional form assumptions are enough to yield a demand equation that is observationally equivalent to the CES. This equivalence breaks down if there are only a finite number of buyers. In that case the heterogeneous consumer model can account for zeros. This makes it worth laying out rather than just invoking the equivalence result.

Consumers from country n, indexed with $n\ell$, have utility functions defined over the products made by each supplier s in each country i, $u_{n\ell is} = \ln[\psi_{n\ell is} q_{n\ell is}]$, where $q_{n\ell is}$ represents the quantity of products consumed and $\psi_{n\ell is}$ is the idiosyncratic preference shock. The heterogeneity is assumed to be distributed Fréchet with a cumulative distribution function (CDF) of $\exp\{-(\psi/(A_i a_{ni}))^{-\theta}\}$, where θ is an inverse measure of consumer heterogeneity and A_i is a location parameter that is specific to the origin country. In an analogous way to equation (10), an increase in A_i shifts up the utility derived from varieties produced in i, which can be interpreted as an increase in perceived quality. a_{ni} also shifts utility upwards, and is a bilateral preference parameter.

Each of the L_n consumers chooses the product giving highest utility and then spends w_n on it. Hence, individual demand is $q_{n\ell is} = w_n/p_{ni}$ for the selected variety and zero on all other varieties. $p_{ni} = p_i \tau_{ni}$ is the price consumers in country n face for product varieties from country i. On the supply side, we assume constant markups (allowing for competitive pricing $p_i = w_i$). The conditional indirect utility function is given by

$$v_{n\ell is} = \ln w_n - \ln(w_i \tau_{ni}) + \ln \psi_{n\ell is}. \tag{11}$$

The Fréchet form for ψ implies a Gumbel form for $\ln \psi$ and thereby implies multinomial logit forms for the probabilities of choosing one of the N_i varieties produced in country i for consumers in n:

$$\mathbb{P}_{ni} = \frac{w_i^{-\theta} A_i^{\theta} \tau_{ni}^{-\theta} a_{ni}^{\theta}}{\sum_{\ell} w_{\ell}^{-\theta} A_{\ell}^{\theta} \tau_{n\ell}^{-\theta} a_{n\ell}^{\theta}}. \tag{12}$$

This equation has a second interpretation that applies to settings in which products are allocated to consumers via auctions. $\mathbb{P}_{ni}$ becomes the probability that i has the highest valuation and therefore makes the winning bid for a good from n.[14]

[13] Income differences would also produce different choices if utility were not homothetic. Fajgelbaum et al. (2011) is a recent combination of the two effects, introducing non-homothetic preferences over quality in a discrete choice logit-type demand system.

[14] Hortaçsu et al. (2009) apply such a model to eBay transactions.

Summing over the set of N_i varieties, $\mathbb{E}[\pi_{ni}] = N_i\mathbb{P}_{ni}$. With a continuum of consumers, the expectation is no longer needed, and $\pi_{ni} = N_i\mathbb{P}_{ni}$. This formulation meets the separability requirement of Definition 2. The exporter attribute and the accessability terms are given by $S_i = N_i w_i^{-\theta} A_i^{\theta}$, and $\phi_{ni} = \tau_{ni}^{-\theta} a_{ni}^{\theta}$. The key difference in this model compared to the first two models lies in the parameter $-\theta$ substituting for $1-\sigma$ when the demand system is CES. There is a very strong parallel though, since an increase in σ means that products are becoming more homogeneous, and an increase in θ means that consumers are becoming less heterogeneous. Whether consumers are becoming more alike in their tastes, or whether products are becoming more substitutable yields similar aggregate predictions for trade flows, which is quite intuitive.

Note that allowing for the bilateral shock a_{ni} to enter preferences of consumers makes it possible for variables like distance to affect trade not only though freight costs, but also through preferences. Another advantage of this model is that, for finite numbers of consumers in the importing country n, it is possible for imports from i to have realized values of zero, an issue we return to in Section 5.2.

2.3.2. *Supply-Side Derivations*

Heterogeneous Industries (Ricardian Comparative Advantage)

Eaton and Kortum (2002) derive a gravity equation that departs from the CES-based approaches in almost every respect and yet the results they obtain bear a striking resemblance. In contrast to the CES-NPD approach, each country produces a very large number of goods (modeled as a continuum) that are homogeneous across countries. In contrast to the CES-MC approach, every industry is perfectly competitive.[15] Productivity z is assumed to be distributed Fréchet with a cumulative distribution function (CDF) of $\exp\{-T_i z^{-\theta}\}$, where T_i is a technology parameter that increases the share of goods for which i is the low-cost supplier and θ determines the amount of heterogeneity in the productivity distribution. Note that the θ parameter now corresponds inversely to dispersion in productivity rather than tastes. However, since this parameter plays the same key role in both models, we maintain the notation in order to emphasize the similarity in resulting terms.

Delivered costs of good g from origin i to destination n are $(c_i/z_{ig})\tau_{ni}$, where c_i is an input price index. Consider one of the goods, the probability of buying it from i is

$$\Pr\left[\ln z_\ell < \ln z_i + \ln\left(\frac{c_\ell \tau_{n\ell}}{c_i \tau_{ni}}\right), \forall h\right]. \tag{13}$$

The Fréchet for z implies Gumbel for $\ln z$, which gives a multinomial logit probability. With a continuum of goods, the share of goods for which consumers in n choose i as

[15] Bernard et al. (2003) reformulate the Eaton and Kortum (2002) model to allow for Bertrand competition in each sector but this reformulation does not change the form of the gravity equation.

their supplier is given by

$$\pi_{ni} = \frac{T_i(c_i\tau_{ni})^{-\theta}}{\sum_\ell T_\ell(c_\ell\tau_{n\ell})^{-\theta}}. \tag{14}$$

Total bilateral flow aggregates over each g good and multiplies expenditure on each good by the above probability. With a CES demand structure over goods, countries spread their overall expenditure X_n according to $X_{ng} = X_n \times p_{ng}^{1-\sigma} / \sum_g p_{ng}^{1-\sigma}$, where p_{ng} is the best price available for good g to country n. Total flow is therefore $X_{ni} = \sum_g X_{ng} \times \pi_{ni} = \pi_{ni} X_n$.

Using the Eaton and Kortum (2002) input cost assumption that $c_i = w_i^\beta P_i^{1-\beta}$ where the price index P_i is proportional to $\Phi_i^{-\theta}$ implies that the two structural gravity terms are given by $S_i = T_i w_i^{-\beta\theta} \Phi_i^{(1-\beta)}$, and $\phi_{ni} = \tau_{ni}^{-\theta}$. The trade cost elasticity, $-\theta$, is equal to the input cost elasticity but the wage elasticity will be smaller since $\beta < 1$.

Heterogeneous Firms

Models covered up to this point have allowed consumers to be heterogeneous in their preferences and industries to differ in terms of production costs. The next step is to let each realization of unit input requirement α be unique so that they can be used to identify individual firms. The CDF of unit input requirements is denoted $G(\alpha)$. Suppose there is a mass of active firms in country i given by N_i. A key variable in heterogeneous firms models is the threshold α_{ni}^*, above which firms do not enter a market. It is a dyadic variable since the threshold must depend on trade costs between i and n. We can now use this notation to obtain an expression for the aggregate share of the market. Chaney (2008) and Helpman et al. (2008) embed heterogeneous firms in a Dixit-Stiglitz framework generalizing the Melitz (2003) paper to multiple countries. The pricing equation is now specific to each firm indexed with their α:

$$p_{ni}(\alpha) = \frac{\sigma}{\sigma - 1} w_i \tau_{ni} \alpha. \tag{15}$$

The resulting market share of i firms in n is therefore:

$$\pi_{ni} = \frac{N_i \int_{\underline{\alpha}}^{\alpha_{ni}^*} p_{ni}(\alpha)^{1-\sigma} dG(\alpha)}{\sum_\ell N_\ell \int_{\underline{\alpha}}^{\alpha_{n\ell}^*} p_{n\ell}(\alpha)^{1-\sigma} dG(\alpha)} = \frac{N_i w_i^{1-\sigma} V_{ni} \tau_{ni}^{1-\sigma}}{\sum_\ell N_\ell w_\ell^{1-\sigma} V_{n\ell} \tau_{n\ell}^{1-\sigma}}, \tag{16}$$

where V_{ni} is defined as in Helpman et al. (2008):

$$V_{ni} \equiv \int_{\underline{\alpha}}^{\alpha_{ni}^*} \alpha^{1-\sigma} dG_i(\alpha).$$

When the threshold entry costs, α_{ni}^*, are less than the lower support, $\underline{\alpha}$, then $V_{ni} = 0$ and there will be no exports from i to n. To specify π_{ni}, we need to solve for V_{ni}, and therefore to specify α_{ni}^* and $G_i(\alpha)$.

In this model, the equilibrium threshold α_{ni}^* such that the corresponding firm is the last one to serve market n (zero profit condition with f_{ni} the fixed cost of serving n from i) is

$$\alpha_{ni}^* = \sigma^{\frac{\sigma}{\sigma-1}}(\sigma-1)\left(\frac{X_n}{f_{ni}\Phi_n}\right)^{\frac{1}{\sigma-1}}\frac{1}{w_i\tau_{ni}}. \tag{17}$$

Since α_{ni}^* depends on destination country characteristics X_n and Φ_n, and on i-specific distribution parameters in $G_i(\alpha)$, we generally cannot separate V_{ni} multiplicatively as would be required to obtain the structural form of gravity. The only functional form known to generate a multiplicable closed form for V_{ni} is the Pareto distribution. Hence we follow Helpman et al. (2008) in setting $G_i(\alpha) = (\alpha^\theta - \underline{\alpha}^\theta)/(\bar{\alpha}_i^\theta - \underline{\alpha}^\theta)$, where θ is the shape parameter and the support of input requirements is $\underline{\alpha}, \bar{\alpha}_i$. The lower bound of $\underline{\alpha} > 0$ is the mechanism through which Helpman et al. (2008) generate aggregate bilateral trade flows of zero. However, to obtain the structural gravity form we need to follow Chaney (2008) and Arkolakis et al. (2012b) in making zero the lower bound for α.[16]

Imposing Pareto (with $\underline{\alpha} = 0$ and country-specific $\bar{\alpha}_i$) and solving for V_{ni}, the aggregate market share of i firms in n is

$$\pi_{ni} = \frac{N_i(w_i\bar{\alpha}_i)^{-\theta}\tau_{ni}^{-\theta}f_{ni}^{-\left[\frac{\theta}{\sigma-1}-1\right]}}{\sum_\ell N_\ell(w_\ell\bar{\alpha}_\ell)^{-\theta}\tau_{n\ell}^{-\theta}f_{n\ell}^{-\left[\frac{\theta}{\sigma-1}-1\right]}}. \tag{18}$$

The first point to note, made originally by Chaney (2008), is that the elasticity of trade with respect to trade costs is now $-\theta$ a supply-side parameter, rather than $1-\sigma$, the preference parameter that determines the elasticity of trade for individual firms (and aggregate trade flows in symmetric firms models). Both parameters can be interpreted as inverse measures of heterogeneity. However, while dispersion in the consumer tastes are increasing in $1/(\sigma-1)$, differences in productive efficiency of firms are what rises with $1/\theta$. The disappearance of the demand parameter is purely a consequence of the Pareto assumption, under which the elasticity of V_{ni} with respect to trade costs is given by $-\theta+\sigma-1$. When adding this elasticity to the intensive margin elasticity, the $1-\sigma$ term drops out.

Equation (18) shows that, in models with an extensive margin of firms' entry, bilateral trade is affected by both variable and fixed trade costs. Eaton et al. (2011a) use $\tilde{\theta}$ to denote $\theta/(\sigma-1)$. Since θ needs to be bigger than $\sigma-1$ for the integral defined by V_{ni} to be finite, $\tilde{\theta} > 1$. Thus, the elasticity of trade with respect to bilateral fixed costs, $-(\tilde{\theta}-1)$ is negative. The fixed costs of entering markets may involve some costs incurred in the domestic economy, w_i, as well as costs incurred in the destination market, w_n. Following

[16] Since α is the inverse of productivity this means that productivity has no upper bound. In that case the continuum assumption implies positive mass of exporters for all country pairs ni.

Arkolakis et al. (2012b), we specify $f_{ni} = \xi_{ni} w_i^{\mu} w_n^{1-\mu}$. Substituting this expression for f_{ni} into (18), we obtain

$$S_i = N_i \bar{\alpha}_i^{-\theta} w_i^{-\theta-\mu\left[\frac{\theta}{\sigma-1}-1\right]} \quad \text{and} \quad \phi_{ni} = \tau_{ni}^{-\theta} \xi_{ni}^{-\left[\frac{\theta}{\sigma-1}-1\right]}.$$

Many of the underlying determinants of variable trade costs, τ_{ni}, such as distance, common language, and colonial history, can reasonably be expected to also contribute to the determination of ξ_{ni}. Two implications follow from this observation: (i) the elasticity of trade with respect to distance now includes both θ and σ, and (ii) even if one could find a variable determining the fixed costs of entry only, equation (18) reveals that its impact is not confined to the binary observation of whether i and n trade at all. It also enters the equation for the value of aggregate trade, and therefore cannot be validly used as an exclusion restriction in a Heckman-type estimation. Note that the procedure used by Helpman et al. (2008) goes beyond simple Heckman-type estimation, and essentially controls for V_{ni} (which is the only channel through which ξ_{ni} enters bilateral flows) in equation (16).

An important limit of CES monopolistic competition models is their constant markup property. This motivated Melitz and Ottaviano (2008) to propose a model with heterogeneous firms that could allow for pro-competitive effects on markups. While, when combined with Pareto, their approach maintains tractability for bilateral trade flows, it does require the assumption of an outside good, which as we see below, leads to a departure from our definition of structural gravity.

In Melitz and Ottaviano (2008), the bilateral exporter's cost threshold c_{ni}^* is simply a function of the domestic production threshold c_n^*, such that $c_n^* = c_{ni}^* \tau_{ni}$. With the linear demand structure used

$$p_{ni}(c) = \frac{1}{2}(c_n^* + \tau_{ni} c) \quad \text{and} \quad q_{ni}(c) = \frac{L_n}{2\gamma}(c_n^* - \tau_{ni} c). \tag{19}$$

Integrating over all firms' individual exports $p_{ni}(c) q_{ni}(c)$ and dividing by X_n, one obtains the collective share of the market

$$\pi_{ni} = \frac{N_i \bar{\alpha}_i^{-\theta} w_i^{-\theta} c_n^{*\theta+2} \tau_{ni}^{-\theta} L_n}{2\gamma(\theta+2) X_n}. \tag{20}$$

The exporter and bilateral terms of general gravity are given by $S_i = N_i \bar{\alpha}_i^{-\theta} w_i^{-\theta}$ and $\phi_{ni} = \tau_{ni}^{-\theta}$. The importer term is $M_n = L_n c_n^{*\theta+2}$. Appendix A.2 of Melitz and Ottaviano (2008) shows that the cutoff in country n is a function of its population and of a market access index that sums trade costs over all source countries: $c_n^{*\theta+2} = \lambda_3 C_n / L_n$, where C_n is a geographical remoteness index (resembling Φ_n of other models) and λ_3 is a constant. After substitution, the importer term in the gravity equation becomes $M_n = \lambda_3 C_n$. Thus, holding the intensity of competition constant in n, M_n is increasing in the *population* of the

importing country but not in the per-capita income. This is due to the non-homotheticity of preferences. In the linear-quadratic utility structure, a higher income individual lowers the share of income spent on the traded varieties and spends a higher share on the outside good. However, the competition-increasing effect of L_n in this model exactly offsets the positive demand effect of country size. Note also that in contrast to the version with Dixit-Stiglitz preferences, ϕ_{ni} does not depend on a bilateral fixed export cost. This is because the linear demand system generates zero trade flows through a choke price.

Arkolakis et al. (2012a) investigate a broader class of variable markup demand systems also featuring choke prices. The general demand system they define is

$$\ln q(p_{ni}(j), p_n^*, x_n) = -\beta \ln p_{ni}(j) + \gamma \ln x_n + d(\ln p_{ni}(j) - \ln p_n^*), \tag{21}$$

for each consumer, where $\gamma \leq 1$ is the income elasticity of demand and $\beta \leq 1$ is a parameter that enters the price elasticity of demand. The $d()$ function shows what happens to demand as $p(j)$ approaches the choke price, p^*. Arkolakis et al. (2012a) results depend on the assumption that $d''() < 0$. They also assume that if the choke price is exceeded, $d()$ goes to negative infinity. Note that p_n^* is also an aggregator of the prices of all other varieties available in market n. Arkolakis et al. (2012a) show that this demand system encompasses a large set of different preferences that have been used in the literature to generate variable markups (Behrens et al. (2009), Feenstra (2003), and a version of Melitz and Ottaviano (2008) where the outside good is omitted). On the supply side of their economy they maintain the Pareto distribution of the productivity of firms competing under monopolistic competition. The two structural gravity terms are given by $S_i = N_i \bar{\alpha}_i^{-\theta} w_i^{-\theta}$ and $\phi_{ni} = \tau_{ni}^{-\theta}$.

Table 3.1 summarizes the results from nine models that fit Definition 1, seven of which fit the stronger requirements of Definition 2. The final column shows trade elasticities with respect to variable trade costs, ϵ. Note that in most structural gravity models, the elasticity of trade with respect to wages is also given by ϵ. For CES-CET, this occurs in the limit as $\gamma \to \infty$ (reallocation of output across destination is costless), for heterogeneous industries it occurs as $\beta \to 1$ (labor is the only input), and for heterogeneous firms as $\mu \to 0$ (fixed costs are paid in units of foreign labor). In principle, if one had reliable estimates of both wage and trade elasticities, one could infer something about these parameters. An important difficulty is to find good instruments for cross-country variation in wages of the origin country that can be excluded from the trade equation.

2.4. Gravity Models Beyond Trade in Goods

The same modeling tools that yield gravity equations for trade in goods can also be applied to other types of flows and interactions. Head et al. (2009) adapt the Eaton and Kortum (2002) (EK) model to the case of service offshoring. Anderson (2011) presents a migration gravity model drawing on discrete choice techniques. Ahlfeldt et al. (2012) draw on

Table 3.1 Theoretical Content of Monadic, Dyadic Terms, and Elasticities of Gravity

Model:	Term:	S_i Exporter	M_n Importer	ϕ_{ni} Bilateral	ϵ Tr. elas.
			Naive Gravity		
N/A		Y_i^a	Y_n^b	ad hoc	N/A
			Structural Gravity		
CES NPD		$A_i^{-\epsilon} w_i^{\epsilon}$	X_n/Φ_n	τ_{ni}^{ϵ}	$1-\sigma$
CES MC (DSK)		$N_i w_i^{\epsilon}$	X_n/Φ_n	τ_{ni}^{ϵ}	$1-\sigma$
CES MC CET		$L_i w_i^{\frac{\epsilon\gamma}{1+\gamma}}$	X_n/Φ_n	τ_{ni}^{ϵ}	$\frac{(1+\gamma)(1-\sigma)}{\sigma+\gamma}$
Heterogeneous consumers		$A_i^{-\epsilon} N_i w_i^{\epsilon}$	X_n/Φ_n	$\tau_{ni}^{\epsilon} a_{ni}^{-\epsilon}$	$-\theta$
Het. industries (EK)		$T_i w_i^{\beta\epsilon} \Phi_i^{1-\beta}$	X_n/Φ_n	τ_{ni}^{ϵ}	$-\theta$
Het. firms (CES)		$N_i \bar{\alpha}_i^{\epsilon} w_i^{\epsilon-\mu\left[\frac{\theta}{\sigma-1}-1\right]}$	X_n/Φ_n	$\tau_{ni}^{\epsilon} \xi_{ni}^{\frac{\theta}{\sigma-1}-1}$	$-\theta$
Het. firms (log-concave)		$N_i \bar{\alpha}_i^{\epsilon} w_i^{\epsilon}$	X_n/Φ_n	τ_{ni}^{ϵ}	$-\theta$
			General Gravity		
CES NPD (outside good)		$A_i^{-\epsilon} w_i^{\epsilon}$	1	τ_{ni}^{ϵ}	$1-\sigma$
Het. firms (linear pref. + outside good)		$N_i \bar{\alpha}_i^{\epsilon} w_i^{\epsilon}$	$L_n c_n^{*\theta+2}$	τ_{ni}^{ϵ}	$-\theta$

Eaton and Kortum (2002) to specify a commuting gravity model. With a few minor changes, the discrete choice framework can easily produce a gravity equation for tourism.

Portes et al. (2001) and Portes and Rey (2005) establish that gravity equations ("naive" definition) can explain cross border portfolio investment patterns as well as they explain trade flows. Martin and Rey (2004) propose a two-country model that they use to justify a gravity equation for bilateral portfolio investment. Coeurdacier and Martin (2009) generalize the framework to multiple countries and apply it using different types of assets and a fixed effects estimation technology very close to the one used by trade economists. Okawa and van Wincoop (2012) suggest an alternative foundation for gravity in international finance.

Gravity equations have also been shown to do a good job fitting stocks of foreign direct investment (FDI). Head and Ries (2008) consider a model in which FDI takes the form of acquisitions. Using the discrete choice framework in a way that resembles Eaton and Kortum (2002), they develop a gravity equation for FDI which fits the data well. de Sousa and Lochard (2011) extend the model to greenfield investment by imagining that instead of bidding for assets, each corporation selects the best "investment project" across all host countries.

In summary, one of the contributions of the development of micro-foundations for the gravity equation for trade is that they can be applied to a range of other bilateral flows and interactions. The key ingredients tend to be "mass" effects that come from adding up constraints and bilateral and multilateral "resistance" terms. Once these gravity equations are specified, they can usually be estimated using the same techniques that are appropriate for trade flows.

3. THEORY-CONSISTENT ESTIMATION

After having described the different theoretical setups that give rise to the gravity prediction, we turn to estimation methods that are consistent with the theory predictions, in particular because they do account for the multilateral resistance terms that are a key feature of general and structural gravity. Historically, the very first approach was to proxy multilateral resistance with remoteness terms. This approach progressively appeared as too weak once the theoretical modeling of gravity became clearer. Researchers then switched to more structural approaches. Because of the influence of Anderson and van Wincoop (2003) in the literature, we start with a version of their approach (their original approach using non-linear least squares has actually been hardly followed), that applies the full structure of the structural gravity framework. We then describe fixed effects estimation that imposes much less structure, but still complies with general gravity. This method can however encounter computational difficulties when using very large datasets, which is not uncommon in the literature. We therefore turn to alternatives when fixed effects are not feasible, and end with Monte Carlo comparisons of all those methods.

3.1. Proxies for Multilateral Resistance Terms

A few early studies have included variables proxies for $1/\Omega_i$ and $1/\Phi_n$ and referred to them as "remoteness." Wei (1996) used a monopolistic competition model to show the theoretical counterparts of these variables but settled for using "log(GDP)-weighted average distances" in his regressions.[17] This bears little resemblance to its theoretical counterpart. Some other remoteness measures differ from their theoretical counterparts in ways that are even more problematic. For instance, Helliwell (1998) measures remoteness as $\text{REM1}_n = \sum_i \text{Dist}_{ni}/Y_i$. This measure has the feature of giving extraordinary weight to tiny countries: as $Y_i \to 0$, REM1 explodes. A better measure of remoteness is $\text{REM2}_n = (\sum_i Y_i/\text{Dist}_{ni})^{-1}$, that is the inverse of the Harris market potential.[18] Tiny countries have negligible effects on REM2 and the size of very distant countries becomes irrelevant.

[17] It is interesting to note that the literature has kept "circling" around those GDP-weighted averages of trade costs as proxies for the MR terms. Baier and Bergstrand (2009), discussed below, can be viewed as the latest approach in that tradition, but one that maintains a clear connection (via approximation) back to the model.

[18] Baldwin and Harrigan (2011) use REM2 to explain the bilateral zero trade flows and Martin et al. (2008) use something close to REM2 as an instrument for trade.

Supposing $\phi_{ni} \sim \text{Dist}_{ni}^{-1}$ and $X_n = Y_n$, the correct Φ_n and Ω_i are $\sum_\ell (Y_\ell/\text{Dist}_{n\ell})\Omega_\ell^{-1}$ and $\sum_\ell (Y_\ell/\text{Dist}_{\ell i})\Phi_\ell^{-1}$. Thus we see that REM2 is on the right track by summing up GDP to distance ratios but it ends up wide off the mark because it implicitly assumes that Φ_ℓ and Ω_ℓ equal one. This makes no sense when the whole point is to obtain a proxy for those variables. Furthermore, while Dist_{ni}^{-1} is an important factor in determining ϕ_{ni} many other trade costs besides distance ought to be considered. In sum, proxy variables do not take the theory seriously enough, a concern that underlines the need for *gravitas*.

3.2. Iterative Structural Estimation

Our implementation of the Anderson and van Wincoop (2003) method involves assuming initial values of $\Omega_i = 1$ and $\Phi_n = 1$, then estimating the vector of parameters determining ϕ_{ni}, then using a contraction mapping algorithm to find fixed points for Ω_i and Φ_n given those parameters. We then run OLS using $\ln X_{ni} - \ln Y_i - \ln X_n + \ln\hat{\Omega}_i + \ln\hat{\Phi}_{ni}$ as the dependent variable. This gives a new set of ϕ_{ni} parameter estimates. We iterate until the parameter estimates stop changing. This method exploits the structural relationship between Ω_i, Φ_n, and ϕ_{ni}. We therefore call the estimator SILS (structurally iterated least squares). Although it is not identical to the Anderson and van Wincoop (2003) method—which is estimated using a non-linear least squares routine in Gauss—SILS does have the advantage of being available as a Stata® ado file (available on our companion website). On the other hand, while SILS uses OLS only, the iteration is time-consuming. Also, the structural methods require data on trade with self and distance to self, both of which may be problematic.

3.3. Fixed Effects Estimation

Standard estimating procedure involves taking logs of equation (1), obtaining

$$\ln X_{ni} = \ln G + \ln S_i + \ln M_n + \ln \phi_{ni}. \tag{22}$$

The naive form of gravity equations involved using log GDPs (and possibly other variables) as proxies for the $\ln S_i$ and $\ln M_n$ but modern practice has been moving toward using fixed effects for these terms instead (Harrigan (1996) seems to be the first paper to have done so). Note that estimating gravity equations with fixed effects for the importer and exporter, as is now common practice and recommended by major empirical trade economists, does not involve strong structural assumptions on the underlying model. As long as the precise modeling structure yields an equation in multiplicative form such as (1), using fixed effects will yield consistent estimates of the components of ϕ_{ni}, which are usually the items of primary interest.[19]

[19] Although the particular model underlying the fixed effects does not matter for the ϕ_{ni} coefficients, it does affect the mapping from the S_i and M_n estimates back to primitives such as technology or demand parameters.

We focus the exposition and our Monte Carlo investigation on cross-sections. However, most current gravity estimations employ datasets that span many years. In such cases the importer and exporter fixed effects should be time-varying as well. The same is true if the data pools over several industries. The S_i and M_n have no reason to be identical across industries since supply capacity of i and total expenditure of n will vary across industries, because of differences in comparative advantages or in consumer's preferences for instance. For panels of trade flows with a large number of years and/or industries, the estimation might run into computational feasibility issues due to the very large number of resulting dummies to be estimated, a challenge that now appears to be solved, as we shall discuss below.

Using country fixed effects has an additional advantage that has nothing to do with being consistent with theory. There can be systematic tendencies of a country to export large amounts relative to its GDP and other observed trade determinants. As an example consider the Netherlands and Belgium. Much of Europe's trade flows through Rotterdam and Antwerp. In principle the production location should be used as the exporting country and the consumption location as the importing country. In practice use of warehouses and other reporting issues makes this difficult so there is reason to expect that trade flows to and from these countries are overstated. Fixed effects can control for this, since they will account for any unobservable that contributes to shift the overall level of exports or imports of a country.

3.4. Ratio-Type Estimation

As mentioned above, the use of fixed effects can sometimes hit a computational constraint imposed upon the number of separate parameters that can be estimated by a statistical package. A solution that has been explored involves using the multiplicative structure of the gravity model to eliminate the monadic terms, S_i and M_n. Head and Mayer (2000) and Eaton and Kortum (2002) normalize bilateral flows X_{ni} by trade with self[20] (X_{nn}) for a given industry/year, delivering a ratio we call the *odds* specification:

$$\frac{X_{ni}}{X_{nn}} = \left(\frac{S_i}{S_n}\right)\left(\frac{\phi_{ni}}{\phi_{nn}}\right). \tag{23}$$

While this specification simplifies greatly the issue by removing any characteristic of the importer, the origin country term S remains to be measured, presumably with substantial error. A related issue is that constructing S_i requires knowledge of the trade cost elasticity, which is also contained in the ϕ_{ni} to be estimated through (23).

Head and Ries (2001) propose a simple solution to cancel those exporter terms, multiplying (23) by $\frac{X_{in}}{X_{ii}}$. If one is ready to assume symmetry in bilateral trade costs

[20] Those manipulations can be done with a reference country other than self. Martin et al. (2008) and Anderson and Marcouiller (2002) use the United States as the reference country.

($\phi_{ni} = \phi_{in}$), and frictionless trade inside countries ($\phi_{nn} = \phi_{ii} = 1$), we end up with a very simple index that Eaton et al. (2011b) call the *Head-Ries Index* (HRI),

$$\hat{\phi}_{ni} = \sqrt{\frac{X_{ni}X_{in}}{X_{nn}X_{ii}}}, \tag{24}$$

and which can be used to assess the *overall* level of trade integration between any two countries.[21]

The problem with the HRI is that it cannot be calculated without a measure of trade *inside* a country (X_{nn}). In principle, it can be proxied using production minus total exports of a country/industry/year combination. Disturbingly, this procedure generates some negative observations, notably for countries like Belgium and the Netherlands, pointing to potential measurement issues related, in particular, to transit shipments, as stated above. Alternative, but related, solutions exist that omit the need for internal trade. Romalis (2007) and Hallak (2006) have used ratios of ratios methods, involving four different international trade flows and thus named the *Tetrads* method by Head et al. (2010). Choosing a reference importer k and a reference exporter ℓ, provides a tetradic term such that

$$\frac{X_{ni}/X_{ki}}{X_{n\ell}/X_{k\ell}} = \frac{\phi_{ni}/\phi_{ki}}{\phi_{n\ell}/\phi_{k\ell}}. \tag{25}$$

The tetradic term can then be used as the LHS to estimate the impact of the usual set of dyadic covariates, with the caveat that all of those covariates need to be "tetrad-ed" as well.[22]

A recent paper that has utilized an alternative trade ratio method is Caliendo and Parro (2012). Their aim is to estimate the trade cost elasticity from tariff data, using asymmetries in protectionism as an identification strategy. Suppose trade costs can be described as $\phi_{ni} = \left[(1 + t_{ni})d_{ni}^{\delta}\right]^{\epsilon}$, where $d_{ni} = d_{in}$ captures all symmetric trade costs (such as distance) in $X_{ni} = GS_iM_n\phi_{ni}$. Introducing a third country h, and multiplying the three ratios X_{ni}/X_{nh}, X_{ih}/X_{hi}, and X_{hn}/X_{in} gives the following estimable equation:

$$\frac{X_{ni}X_{ih}X_{hn}}{X_{nh}X_{hi}X_{in}} = \left(\frac{(1 + t_{ni})(1 + t_{ih})(1 + t_{hn})}{(1 + t_{nh})(1 + t_{hi})(1 + t_{in})}\right)^{\epsilon}. \tag{26}$$

[21] Head and Ries (2001) apply it to the US/Canada FTA (free trade agreement), Head and Mayer (2004a) to a comparison of North American and European integration, Jacks et al. (2008) use it to measure trade integration over the very long run using trade data of France, Germany, and the UK from 1870 to 2000, and Eaton et al. (2011b) use it to quantify the effects of the 2008–2009 crisis on trade integration. $\hat{\phi}_{ni}$ can also be used as the left-hand side (LHS) of a regression trying to explain the bilateral determinants of trade integration (Combes et al. (2005) and Chen and Novy (2011) are examples following that path).

[22] A difficulty in implementing Tetrads in practice is the choice of the reference countries, since doing all potential combinations of k and ℓ would drive the number of observations into the billions in most applications. Romalis (2007) focuses on the impact of the North American Free Trade Agreement (NAFTA) where he considers EU12 as a reference importer, and each of the NAFTA countries in turn as a reference exporter. Head et al. (2010) study the erosion of colonial preferences and therefore face a higher dimensional issue. Their preferred specification takes the average of results when reference countries are chosen in turn in the set of the five biggest traders in the world. As shown in the Monte Carlo exercise below, Tetrads yields a very small bias when the share of missing values in the data is minimal.

3.5. Other Methods

The ratios approaches are one way to deal with an exceedingly large number of dummies required by theory. An intuitive alternative is to "double-demean" the gravity dataset, one demeaning for the exporter dimension, one for the importer. However, this solution only yields unbiased estimates if the dataset is completely full, with no missing flows. Another approach is to demean in one dimension only, and use dummies in the other dimension. This hybrid strategy does not require the matrix of trade flows to be full, and divides the computational problem by two, which however might prove insufficient in some cases (with 150 countries and 60 years for instance, 9000 dummies remain to be estimated). Following on the analysis of employer–employee datasets carried out by Abowd et al. (1999), iterative methods have been developed to solve the two-way Fixed Effect (FE) problem with unbalanced data and very large numbers of effects. The command we have employed is **reg2hdfe** by Guimaraes and Portugal (2010) which allows for clustered standard errors.

Another alternative, dubbed Bonus Vetus OLS, has been proposed by Baier and Bergstrand (2009). Define $\text{MRS}(v_{ni}) = \bar{v}_i + \bar{v}_n - \bar{\bar{v}}$. Similarly let $\text{MRD}(v_{ni})$ be the GDP-weighted version of these averages. Bonus Vetus adds $\text{MRD}(v_{ni})$ (or $\text{MRS}(v_{ni})$ in the unweighted version) for each trade cost variable to the regression and constrains it to have the opposite sign as v_{ni}. The unweighted version resembles double-demeaning in which one subtracts $\text{MRS}(v_{ni})$ from the dependent variable as well as all right-hand side (RHS) variables.

3.6. Monte Carlo Study of Alternative Estimators

In order to compare the major set of methods described above, we run a Monte Carlo exercise using structural gravity as a data generating process (DGP). For the determinants of trade, we use actual data for the 170 countries for which we have data on GDP, distance, and the existence of a Regional Trade Agreement (RTA) in 2006. The DGP specifies accessibility as a function of distance and RTA:

$$\phi_{ni} = \exp\left(-\ln \text{Dist}_{ni} + 0.5\ \text{RTA}_{ni}\right)\eta_{ni},$$

where η_{ni} is a log-normal random term. The η_{ni} is the only stochastic term in the simulation since the GDPs, distances, and RTA relationships are all set by actual data. We calibrate the variance of $\ln \eta_{ni}$ to replicate the root mean squared error (RMSE) of the least squares dummy variables (LSDV) regression on real data. As we will show later, the distance elasticity of -1 and the 0.5 coefficient on the RTA dummy are representative of the literature. Combining this with incomes of exporters and importers, we calculate the multilateral resistance terms, Φ_n and Ω_i using equation (3), which are used in (2) to generate bilateral trade flows.[23]

[23] Baier and Bergstrand (2009) adopt the same method to run the Monte Carlo comparison of their Bonus Vetus estimation method with other methods, with one important difference. Rather than including the random term in ϕ_{ni} before calculating the MR index, they introduce the log-normal perturbation just prior to estimation. They therefore adopt a statistical approach, rather than a structural approach to the error term, according to which MR terms should be calculated using the whole of ϕ_{ni} and not only its deterministic part.

Table 3.2 The Estimators Used in this Study

Abbrev.	Description	Introduced by
OLS	Linear-in-logs with GDPs	Tinbergen (1962)
SILS	Structurally Iterated Least Squares	Anderson and van Wincoop (2003)*
LSDV	Least squares w/country dummies	Harrigan (1996)
DDM	Double-Demeaning of LHS & RHS	None
BVU	Bonus Vetus OLS, simple averages	Baier and Bergstrand (2010)
BVW	Bonus Vetus OLS, GDP-weighted	Baier and Bergstrand (2009)
Tetrads	Ratios of reference exporter & importer	Head et al. (2010)

*Section 3.2 explains how SILS differs from the original method.

Since this DGP does not yield missing flows, and such missing flows are a substantial part of the computational issues (due to the problems raised by double-demeaning in unbalanced panels), we propose two ways to generate missing values (which due to the log specification can also be interpreted as zero flows). The first one suppresses X% of observations randomly, while the second method removes the smallest X% of the initial set of export flows. The first method can be thought of as representing haphazard data collection and reporting, whereas the second method can be thought of as eliminating exports that are too small to be profitable in the presence of fixed market entry costs. To consider minor, moderate, and major amounts of missing data we set X at 5%, 25%, and 50%.

Table 3.3 presents the results of a simulation of the seven different methods shown in Table 3.2. Each "cell" of the table is a method-sample-regressor combination. The top value in a cell shows the mean estimate over 1000 repetitions, that is the expected value of the estimator. The second value in parentheses is the average standard error and the third, in square brackets, is the standard deviation of the estimate. If the first number is equal to the true values of -1.0 and 0.5 the estimator is unbiased. If the last two values are equal, the estimator also gives unbiased standard errors.

The first point emerging from the simulations reported in Table 3.3 is that *OLS is a poor estimator under the structural gravity DGP*. Its estimates are biased toward zero for both explanatory variables. The method is not robust to deleting the smallest observations. These results validate the decision of Baldwin and Taglioni (2007) to bestow their "gold medal" mistake to gravity regressions that fail to include exporter and importer dummies.

SILS, the structural method we programmed based on Anderson and van Wincoop (2003), gives distance estimates that are close to the assumed true values when there is no missing data. A comparison of the standard deviations of the estimates between SILS and LSDV reveals that LSDV deliver substantially more precise estimates. SILS coefficients are stable in the presence of randomly missing data. With selectively missing data, both

Table 3.3 Monte Carlo Estimates of Distance and RTA Effects

Censoring Estimates	None		Observations Deleted: 5%				Observations Deleted: 25%				Observations Deleted: 50%			
			Random		Smallest Flows		Random		Smallest Flows		Random		Smallest Flows	
	Dist.	RTA	Dist.	RTA	Dist.	RTA	Dist.	RTA	Dist.	RTA	Dist.	RTA	Dist.	RTA
OLS	−0.836	0.276	−0.836	0.277	−0.726	0.444	−0.836	0.276	−0.578	0.485	−0.836	0.276	−0.478	0.324
	(0.021)	(0.063)	(0.022)	(0.064)	(0.021)	(0.062)	(0.025)	(0.072)	(0.022)	(0.062)	(0.030)	(0.089)	(0.023)	(0.063)
	[0.051]	[0.114]	[0.051]	[0.114]	[0.045]	[0.106]	[0.052]	[0.118]	[0.036]	[0.097]	[0.055]	[0.129]	[0.031]	[0.089]
SILS	−0.937	0.749	−0.937	0.750	−0.833	0.666	−0.937	0.748	−0.819	0.141	−0.937	0.752	−0.904	−0.471
	(0.021)	(0.060)	(0.021)	(0.062)	(0.021)	(0.059)	(0.024)	(0.069)	(0.022)	(0.060)	(0.030)	(0.085)	(0.024)	(0.063)
	[0.058]	[0.176]	[0.058]	[0.176]	[0.051]	[0.171]	[0.059]	[0.183]	[0.046]	[0.161]	[0.062]	[0.202]	[0.044]	[0.146]
LSDV	−1.000	0.501	−1.000	0.501	−0.934	0.596	−1.001	0.501	−0.799	0.651	−0.999	0.503	−0.691	0.545
	(0.021)	(0.058)	(0.022)	(0.060)	(0.021)	(0.058)	(0.024)	(0.067)	(0.022)	(0.059)	(0.030)	(0.083)	(0.024)	(0.062)
	[0.021]	[0.059]	[0.022]	[0.061]	[0.022]	[0.059]	[0.026]	[0.069]	[0.022]	[0.059]	[0.031]	[0.084]	[0.024]	[0.062]
DDM	−1.000	0.501	−0.999	0.501	−0.920	0.624	−0.997	0.499	−0.712	0.789	−0.988	0.497	−0.532	0.715
	(0.021)	(0.058)	(0.022)	(0.059)	(0.021)	(0.058)	(0.024)	(0.067)	(0.023)	(0.061)	(0.030)	(0.082)	(0.026)	(0.065)
	[0.021]	[0.059]	[0.022]	[0.061]	[0.022]	[0.059]	[0.025]	[0.068]	[0.022]	[0.061]	[0.030]	[0.084]	[0.023]	[0.063]
BVU	−1.000	0.501	−1.000	0.502	−0.933	0.583	−1.000	0.501	−0.859	0.431	−1.001	0.501	−0.839	0.060
	(0.025)	(0.067)	(0.025)	(0.069)	(0.024)	(0.067)	(0.028)	(0.078)	(0.026)	(0.069)	(0.035)	(0.095)	(0.029)	(0.074)
	[0.021]	[0.059]	[0.022]	[0.061]	[0.022]	[0.060]	[0.027]	[0.071]	[0.024]	[0.066]	[0.032]	[0.088]	[0.028]	[0.071]
BVW	−0.995	0.524	−0.491	1.230	−0.912	0.769	−0.140	1.626	−0.616	1.233	−0.055	1.516	−0.142	1.697
	(0.022)	(0.055)	(0.016)	(0.053)	(0.021)	(0.054)	(0.009)	(0.055)	(0.022)	(0.056)	(0.006)	(0.063)	(0.020)	(0.060)
	[0.049]	[0.157]	[0.093]	[0.187]	[0.046]	[0.142]	[0.048]	[0.154]	[0.052]	[0.132]	[0.029]	[0.191]	[0.052]	[0.122]
Tetrads	−0.998	0.509	−0.878	0.714	−0.936	0.570	−0.530	1.258	−0.925	0.474	−0.404	1.582	−0.962	0.294
	(0.131)	(0.355)	(0.160)	(0.413)	(0.129)	(0.347)	(0.213)	(0.569)	(0.131)	(0.338)	(0.234)	(0.668)	(0.139)	(0.339)
	[0.137]	[0.366]	[0.172]	[0.418]	[0.134]	[0.358]	[0.220]	[0.540]	[0.133]	[0.345]	[0.252]	[0.645]	[0.134]	[0.348]

Notes: Top value in each cell is the mean estimate (based on 1000 repetitions). The true parameters are −1 for distance and .5 for RTA. Average standard error in "( )" and standard deviation of estimate in "[ ]". Table 3.2 defines the estimators.

LSDV and SILS estimates deviate notably from the true parameters. We conclude that, even though SILS can be estimated with Stata®, it is not worth the computational effort.

Double-demeaning both log exports and the RHS variables (DDM) and the Bonus Vetus *unweighted* (BVU) approach of double-demeaning just the RHS variables deliver identical results (out to machine precision) when there is no missing data. Unfortunately real gravity data does tend to have missing data. DDM is one of the worst estimators when there are large numbers of non-random missing observations. BVU appears to have better robustness properties. In the worst case scenario with the smallest half of the original data eliminated, BVU gives somewhat better distance elasticities than LSDV but much worse RTA estimates. The GDP-weighted double-demeaning of the RHS variables (BVW) has several disadvantages. Its estimates are not robust to missing data and it is very imprecise as we see in the high standard deviation of the coefficients. Its standard errors appear to be biased downwards.

Tetrads seems to be unbiased except when there are substantial numbers of randomly missing observations. It does quite well with DGPs that eliminate the smallest trade flows. But even there it is imprecise. Fortunately the **cluster2** standard errors we use correctly measure this imprecision. Given the imprecision, the lack of robustness to randomly missing data and sensitivity of results to the choice of reference countries (see Head et al., 2010), the argument for Tetrads hinges on LSDV being computationally infeasible. This is because software such as Stata® cannot handle the large number of dummies needed for panel estimation of time-varying country fixed effects. Fortunately, two-way fixed effects based on the iterative method of Guimaraes and Portugal (2010) yield identical estimates to LSDV (which is why we do not report it separately) and are not subject to arbitrary limits. These 2WFE methods mean that "fixes" like DDM, BVU, and Tetrads are no longer advisable.[24]

These simulations have considered a DGP that follows closely from the major theories that deliver the form we call structural gravity. In this DGP there is a built in relationship between bilateral resistance terms, distance and RTAs, and the multilateral resistance terms. This covariance is sufficient in its own right to cause notably high bias of OLS. Fortunately LSDV solves this problem perfectly so long as there are no other econometric issues. In Section 5 we consider two particularly important additional problems that can undermine the argument for LSDV: heteroskedastic errors and structural zeros.

3.7. Identification and Estimation of Country-Specific Effects

In the presence of importer and exporter fixed effects a variety of potentially interesting trade determinants can no longer be identified in a gravity equation. Notably, (1) anything

[24] There is one case where we see Tetrads outperforming LSDV and that is when the smallest 25% of trade flows are selectively removed. In Section 5.2 we point to other methods better suited to such selective censoring of the data.

that affects exporters' propensity to export to all destinations (such has having hosted the Olympics or being an island), (2) variables that affect imports without regard to origin, such as country-level average applied tariff, and (3) sums, averages, and differences of country-specific variables. If any variables of these three forms is added to a trade equation estimated with importer and exporter fixed effects, programs such as Stata® will report estimates with standard errors. However the estimates are meaningless. They are identified by dropping one or more of the country dummies. This is the case for size variables Y_i and Y_n naturally, and country-level institutional variables (e.g. rule of law). Also problematic is the use of exchange rates in this respect. Since (like any relative price) the bilateral value of a currency is defined as a ratio, the fixed effects will swallow each of the price terms after the usual logarithmic transformation of the gravity regression.

To retain monadic variables, authors sometimes resort to creating new dyadic variables using functional form assumptions other than linear relationships. For example, one can create a bilateral institutions variable by multiplying quality of institutions in i times quality of institutions in n. This is identifiable even when having i and n FEs, but this is a sort of constructed identification, with no straightforward interpretation in many cases.[25] A second example is the case of using country-specific average tariff data to try to create a bilateral tariff variable. If one simply averages country i and country n tariffs, the effect is not identified. To get around this, one might take the log of the average tariff. In this case the bilateral tariff effect is identified but only by the choice of functional form: the log of the *product* of i and n tariffs would not work.

While most of the applications are in panel gravity equations, the time dimension clutters notation so we consider the case of a cross-section gravity equation. The underlying estimating equation is

$$\ln X_{ni} = \alpha_i + \beta V_i + \gamma_n + \delta D_{ni} + \varepsilon_{ni}. \tag{27}$$

V_i is a monadic variable of interest. It could be a direct measure of the cost or quality of exports from country i or some geographic or institutional characteristic that underlies cost and quality differences. The D_{ni} are the dyadic controls (e.g. distance, RTAs). The α_i term represents all the other i-level determinants of exports.

There are several possible ways to estimate β and we follow here the treatment of a similar problem in labor economics by Baker and Fortin (2001). The case they consider is the effect of the percent of female workers in an occupation (corresponding to our V_i) on the wages of individuals in that occupation (analogously, our $\ln X_{ni}$). The γ_n destination fixed effect would correspond to an individual worker effect (which Baker and Fortin

[25] One good example where the multiplication does seem appropriate is the case of Rauch and Trindade (2002). The idea is that trade is more likely when conducted by an exporting firm who is managed by someone of the same ethnicity as the corresponding importer. The probability that two randomly selected members of each population would encounter each other is given by the product of the ethnicity share in the two counties. Note that the paper itself does not use exporter and importer fixed effects.

do not consider presumably because workers do not move across occupations enough to identify such a term).

Probably the most common approach taken in labor or gravity equations is a one-step estimation. The simplest version combines α_i and ε_{ni} as the error term of equation (27). Even if α_i is uncorrelated with V_i, the error terms for the same exporter will be correlated. This will result in downward biased standard errors of β unless standard errors are clustered by exporter.

A two-step estimator is another way to solve the standard error problem and it has other potential advantages. In the two-step approach, one first estimates the two-way fixed effects version of equation (27) in which exporters fixed effect $\ln S_i$ replaces $\alpha_i + \beta V_i$. The second step is to regress $\widehat{\ln S_i}$ on V_i. Eaton and Kortum (2002) is an early example of the two-step approach in the gravity literature with cross-sectional trade data. Head and Ries (2008) is an example using FDI data.

As pointed out by Baker and Fortin (2001), both methods can be improved by modeling α_i as the sum of the effects of some i-specific controls, C_i, the average characteristics of each exporter, $\bar{D}_i = (\sum_n D_{ni})/N$, and an error term.

$$\alpha_i = \alpha_0 + \alpha_1 C_i + \alpha_2 \bar{D}_i + \psi_i. \tag{28}$$

Substituting this equation into (27) yields a superior version of the one-step equation:

$$\ln X_{ni} = \alpha_0 + \alpha_1 C_i + \beta V_i + \gamma_n + \delta D_{ni} + \alpha_2 \bar{D}_i + (\psi_i + \varepsilon_{ni}). \tag{29}$$

Standard errors should be clustered at the i-level since the presence of ψ_i causes the error to be correlated across n for a given i. This approach looks attractive because it recovers the *within* estimates for the dyadic variables and still allows one-step estimation of the monadic effects. That is, the presence of the $\bar{D}_i$ causes $\hat{\delta}$ to be estimated as if there were i-specific fixed effects. The estimate of β remains vulnerable to correlation between the ψ_i and V_i.

Lastly, we can also consider a two-step version of (29). It first estimates $\widehat{\ln S_i}$ in a fixed effects regression. Recognizing that the fixed effects are estimated with error, denoted ω_i, the estimated i fixed effects are then regressed on all the i-specific variables:

$$\widehat{\ln S_i} = \alpha_0 + \alpha_1 C_i + \alpha_2 \bar{D}_i + \beta V_i + (\psi_i + \omega_i). \tag{30}$$

Since different fixed effects are estimated with differing amounts of precision, the error $\psi_i + \omega_i$ is heteroskedastic. Estimating (30) by generalized least squares gives greater weight to observations with lower standard errors on $\widehat{\ln S_i}$. However, Baker and Fortin (2001) point out that there is no particular reason to expect ψ_i to be heteroskedastic in the same way as ω_i. If ψ_i is homoskedastic and has high variance then more efficient estimation will come from giving equal weight to all observations in the second step. It therefore seems sensible to estimate all three specifications—the one-step equation (29) and the GLS and OLS versions of the two-step equation (30).

Table 3.4 Estimates of Typical Gravity Variables

	All Gravity				Structural Gravity			
Estimates:	**Median**	**Mean**	**s.d.**	**#**	**Median**	**Mean**	**s.d.**	**#**
Origin GDP	.97	.98	.42	700	.86	.74	.45	31
Destination GDP	.85	.84	.28	671	.67	.58	.41	29
Distance	−.89	−.93	.4	1835	−1.14	−1.1	.41	328
Contiguity	.49	.53	.57	1066	.52	.66	.65	266
Common language	.49	.54	.44	680	.33	.39	.29	205
Colonial link	.91	.92	.61	147	.84	.75	.49	60
RTA/FTA	.47	.59	.5	257	.28	.36	.42	108
EU	.23	.14	.56	329	.19	.16	.5	26
NAFTA	.39	.43	.67	94	.53	.76	.64	17
Common currency	.87	.79	.48	104	.98	.86	.39	37
Home	1.93	1.96	1.28	279	1.55	1.9	1.68	71

Notes: The number of estimates is 2508, obtained from 159 papers. Structural gravity refers here to some use of country fixed effects or ratio-type method.

4. GRAVITY ESTIMATES OF POLICY IMPACTS

From the first time gravity equations were estimated, one of the main purposes has been to investigate the efficacy of various policies in promoting trade.[26] From this standpoint, production, expenditure, and geography are just controls with the real target being a policy impact coefficient. This section considers the evidence that has been gathered on the policy coefficients and then turns to the harder question of how to move from coefficients to economically meaningful impact measures.

4.1. Meta-Analysis of Policy Dummies

Using Disdier and Head (2008) as a starting point, we have collected a large set of estimates of important trade effects other than distance and extended the sample forward after 2005. The set of new papers augments the Disdier and Head (2008) sample by looking at all papers published in top-5 journals, the *Journal of International Economics* and the *Review of Economics and Statistics* from 2006 to available articles of 2012 issues. A second set of papers were added, specifically interested in estimating the trade costs elasticity. Since those are much less numerous, we tried to include as many as possible based on our knowledge of the literature. A list of included papers is available in the web appendix. The final dataset includes a total of 159 papers, and more than 2500 usable estimates. We provide in Table 3.4 meta-analysis type results for the most frequently used variables in gravity equations, including policy-relevant ones.

[26] Tinbergen (1962) found small increases in bilateral trade attributable to Commonwealth preferences (≈5%) and the Benelux customs union (≈ 4%).

The table is separated in two groups of four columns: one giving summary statistics of estimates across all papers, and one focusing on structural gravity papers. Here we must have a somewhat looser definition of what structural gravity is, since the use of theory-consistent methods has been quite diverse, and evolving over time. We choose to adopt a rather inclusive definition. For instance many papers include origin and destination country fixed effects, although their data is a panel, and should therefore include country-*year* dummies. We classify as structural the papers that include some form of country dummies or ratio type estimation. We also drop outliers for each of the gravity variables investigated, using a 5% threshold.

The first results are that GDP elasticities are close to unitary as predicted by theory and shown in Figure 3.1 for Japan–EU trade. This is particularly true for origin GDPs (mean of 0.98). The destination GDP elasticity is lower (0.84), a finding that Feenstra et al. (2001) pointed to as evidence of home market effects.

The average distance elasticity of −0.93 is close to the −0.91 reported by Disdier and Head (2008). Thus, the 368 additional estimates we obtained by updating the sample are not out of line with the earlier sample. Consistent with our Monte Carlo results above, we also find that the distance coefficient is biased toward zero empirically when committing the gold medal mistake of not controlling for MR terms. The magnitude of the bias even seems to be quite in line with our Monte Carlo.

Contiguity and common language effects seem to have very comparable effects, with coefficients around 0.5, about half the effects of colonial links. Common language and colonial linkage are frequent proxies for cultural/historical proximity. Those "non-traditional" determinants of economic exchange turn out to be important factors in trade patterns.

The two direct policy relevant variables, RTAs and common currency, have large estimated effects—albeit with large standard deviations. Interestingly, the meta-analysis by Cipollina and Salvatici (2010) on the trade effects of RTAs report a mean effect of 0.59 and median effect of 0.38 for their 1867 estimates. This is quite close to the characteristics of our smaller sample of 257 estimates (mean of 0.59 and median of 0.47). Interestingly, they find that structural gravity yields *stronger* estimates of RTA effects, whereas we find weaker effects (mean of 0.36). Many papers include dummies for RTAs of specific interest, notably the EU and NAFTA which involve some of the largest bilateral trade flows worldwide. Whether looking at the median or mean coefficients, estimated using naive or structural gravity, the North-American agreement seems to be associated with larger amounts of trade creation. Cipollina and Salvatici (2010) also find this pattern, with a mean coefficient for NAFTA (0.90) almost twice as big as the one for EU (0.52).

The trade effects of common currencies have been the subject of controversy. Our mean over 104 estimates is 0.79, which corresponds to a doubling of trade. This is substantially smaller than initial estimates by Rose (2000) who put the currency union coefficient at 1.21, implying more than tripling trade. However, the meta-analysis average is

substantially larger than the preferred estimates of some recent work. Baldwin (2006), synthesizing a stream of papers focusing mainly on the Euro, puts the currency effect at about 30%. Santos Silva and Tenreyro (2010) find virtually no effects on trade for the Euro, after taking account of the high level of trade integration of Eurozone members even *before* they formed a common currency. Berthou and Fontagné (2013) use firm-level exports by French firms and find a weakly significant 5% effect, coming mostly from average exports by the most productive firms. Frankel (2010) finds a more optimistic 15% increase of trade that takes about five years to take place, and then stabilizes.[27]

The numbers reported in Table 3.4 establish the *typical* findings but they should not be interpreted as *preferred* estimates of the causal effects of the policy variables. This is because by and large they fail to address the endogeneity related to many of the policy variables and especially to currency unions and RTAs. There are many examples where the countries that sign a trade enhancing agreement already trade a great deal together (NAFTA, EU). Since currency unions economize on transaction costs of converting exchange, they will be greater when there are more transactions, that is when countries trade a lot with each other. Cross-section or pooled panel estimates are therefore not reliable—even if they have country or country-year fixed effects. The textbook solution would be to find instrumental variables but we are not aware of any compelling instruments. Most variables that plausibly cause currency unions or RTAs also "belong" in the trade equation on their own (e.g. distance, colonial history). Lacking plausible instrumental variables (IVs), the most promising approach is to include country-pair fixed effects. This forces identification to come from the *within* dimension of the data. Studies that introduce dyadic fixed effects often obtain dramatically different coefficient estimates from the pooled OLS estimates.[28] Another strategy is to use a natural experiment. In the final part of the paper, Frankel (2010) uses the conversion of the French Franc to the euro in 1999 as an exogenous shock hitting Western African countries that had the CFA Franc (linked to the French Franc) as a currency. The switch to a common currency with members of the Eurozone other than France can reasonably be considered as exogenous for this group of African countries. The trade creating effect seems stronger with this method (around 50%) than with the more classical approach used in the first part of this paper. It also coincides with

[27] Differences might come from different sets of fixed effects, and from different estimators. Baldwin and Taglioni (2007, Table 4) turn the Eurozone coefficient from a significant positive 0.17 with OLS, to a significant −0.09 with the appropriate set of country-year and country-pair fixed effects that account for MR terms, and identify in the within dimension. Santos Silva and Tenreyro (2010) have similar identification strategy and results, with Poisson pseudo-maximum likelihood (PPML) rather than a linear-in-logs estimator. Frankel (2010) regressions have a country pair fixed effect, but not the country-year dummies that would control for MR terms.

[28] For instance, Baier and Bergstrand (2007) find that the RTA estimate is multiplied by more than two, while Glick and Rose (2002) find that the common currency effect is divided by around the same factor. Head et al. (2010) also conduct a dyadic fixed effect specification. Compared to a naive specification, they find a rise in the effect of the General Agreement on Tariffs and Trade (GATT)/World Trade Organization (WTO) (which is also the case in Rose (2004)), and confirm the fall in common currency effects. On RTA they find that the coefficient is halved, in contrast to the results of Baier and Bergstrand (2007).

the switch to the euro, although coefficients puzzlingly lose significance in the two last years of the sample (2005 and 2006).

4.2. The Elasticity of Trade with Respect to Trade Costs

Arkolakis et al. (2012b) show that a gravity equation is all that is needed to calculate welfare gains from trade. Indeed, of the two sufficient statistics required when their macro restrictions hold, one is directly observable (the import ratio), and the other, the trade cost elasticity of trade, can be estimated using a gravity equation for bilateral trade. While relatively few gravity papers estimate trade cost elasticities, we have identified 32 papers that do so, and we summarize their results in Table 3.5. We will refer to those as "gravity-based" estimates. They involve regressing bilateral trade on measures of bilateral trade costs or on exporter "competitiveness" such as wages or productivity.[29] About three quarters of our estimates of ϵ are of the first type, and come from regressions along the lines of

$$\ln X_{ni} = \ln S_i + \ln M_n + \epsilon \ln \tau_{ni}. \tag{31}$$

In many cases, equation (31) is estimated at the industry level. This explains in part the very large variance observed across estimates in the literature, reported in Table 3.5.[30] Most specifications measure $\ln \tau_{ni}$ as the log of one plus the *ad valorem* bilateral tariff rate. In some cases the *ad valorem* freight rates are used instead of or in addition to tariff rates (Hummels (1999) in particular).

We define the gravity-based method broadly enough to encompass estimates derived from regressing bilateral trade on proxies for exporter competitiveness such as wages, exchange rates, and prices. The precise implementation of the competitiveness-based estimate can take two forms: (1) estimate the exporter in a first stage and regress it on wages in a second stage, and (2) directly estimate the bilateral equation using the determinants of S_i, including wages. Eaton and Kortum (2002) is an example of the first approach. They regress exporter fixed effects (derived from a transformed bilateral trade variable) on proxies for technology (R&D expenditures, average years of education) and wages. Instrumenting for wages, they obtain an elasticity of -3.6. As with the trade cost methods, this approach is actually more general than the precise model used by Eaton and Kortum (2002). Indeed, when looking at Table 3.1, we see that the wage in the origin country exhibits an elasticity that is the same or closely related to the elasticity with respect to bilateral trade costs in most foundations of gravity. In Eaton and

[29] In addition to the gravity-based estimates included in our meta-analysis, there are two other influential approaches. One method, devised by Feenstra (1994) and applied more broadly by Broda and Weinstein (2006), is to estimate the "Armington" elasticity, σ, using Generalized Method of Moments (GMM) identification via heteroskedasticity. Then $1 - \sigma$ could be used as the estimate of ϵ. A second method originated by Eaton and Kortum (2002) and refined by Simonovska and Waugh (2011), estimates ϵ by relating trade variation to price gaps.

[30] Taking into account this heterogeneity has been shown recently to be particularly important for the estimation of welfare gains from trade, which are larger when ϵ varies across sectors (see Ossa (2012) and Costinot and Rodriguez-Clare (2013) for expositions and estimations of the aggregation bias in welfare gains calculations).

Table 3.5 Descriptive Statistics of Price Elasticities in Gravity Equations

Estimates:	Median	Mean	s.d.	#
Full sample	−3.19	−4.51	8.93	744
Naive gravity	−1.31	−1.35	5.17	122
Structural gravity	−3.78	−5.13	9.37	622
Split structural estimates by:				
Estimation method:				
Country FEs	−3.5	−4.12	8.2	447
Ratios	−4.82	−7.7	11.49	175
Identifying variable:				
Tariffs/Freight rates	−5.03	−6.74	9.3	435
Price/Wage/Exchange rate	−1.12	−1.38	8.46	187

Notes: The number of statistically significant estimates is 744, obtained from 32 papers.

Kortum (2002), the wage elasticity is $-\beta\theta$ whereas the trade elasticity is $-\theta$. Thus, if we know β (the share of wages in the cost function in their model), the effect of wage variation on estimated $\widehat{\ln S_i}$ is an alternative source of identification for the same key parameter.[31] The second approach is chosen by Costinot et al. (2012), who estimate a trade elasticity of −6.5. Their method regresses log corrected exports on log productivity, which they capture based on producer prices data, as their theory is one of perfect Ricardian competition.

Table 3.5 reports the average value and standard deviation of 744 coefficients obtained for the full sample of 32 papers. We then split the sample according to several important characteristics: (1) estimates dealing with the multilateral resistance terms through country fixed effects, ratios or not treating the MR problem, and (2) the variable identifying the price elasticity in the regression (tariffs/freight rates vs exchange rates, relative producer prices, or productivity). This last decomposition is done on the set of estimates that treat the MR problem (structural gravity estimates).

Results in Table 3.5 show that estimates of price elasticities vary immensely with a standard deviation twice as large as the mean. On average the elasticity of trade is −4.51, but when using a median to reduce the influence of outliers, this falls to −3.19. Much of the variance in the estimates can be related to estimation methods: structural gravity (defined the same loose way as in Table 3.4) yields much larger responses of trade flows to price shifters than naive gravity. Also important in the debate between international trade and international macro-economists is the difference between coefficients estimated

[31] Since wages are likely to be simultaneously determined with trade patterns, it seems important to instrument for wages, and indeed, the estimated parameter seems to be systematically larger (in absolute value) when instrumenting. It is the case for Eaton and Kortum (2002), Costinot et al. (2012), and Erkel-Rousse and Mirza (2002).

using bilateral tariffs vs exchange rate changes. The latter tend to be much smaller than the former, related to the different usage in the accepted values (for calibrating models in particular) of the two academic populations.[32] Note that the difference between elasticities identified through relative prices or bilateral tariffs holds *within* papers. Studies such as De Sousa et al. (2012) and Fitzgerald and Haller (2012) which estimate the effects of exchange rates and tariffs in the same regressions find comparable differences to ones seen in Table 3.5. Overall, our preferred estimate for ϵ is -5.03, the median coefficient obtained using tariff variation, while controlling for multilateral resistance terms.

Armed with this estimate of the trade elasticity, we can do a simple calculation to determine if estimated RTA effects in Table 3.4 are reasonable. Let ρ be the estimated coefficient on the RTA dummy variable. Since it measures the reduction in trade costs achieved by the RTA, $\rho = \epsilon(\ln \tau_{ni}^{\text{MFN}} - \ln \tau_{ni}^{\text{RTA}})$, where τ_{ni}^{MFN} is the "most-favored-nation" trade cost factor that n would apply to imports from i were they not in a free trade agreement. Denote as t the Most Favoured Nation (MFN) tariffs that must be removed in the RTA (as per GATT article 24) and let κ capture the *ad valorem* tariff equivalent of all trade barriers that remain in force after the implementation of the RTA. Then $\tau_{ni}^{\text{MFN}} = 1 + \kappa + t$ and $\tau_{ni}^{\text{RTA}} = 1 + \kappa$. After some algebra, we obtain $t = (1 + \kappa)[\exp(\rho/\epsilon) - 1]$. Martin et al. (2012) estimate $\rho = 0.26$, implicitly assume $\kappa = 0$, and set $\epsilon = 4$ to calculate $t = \exp(0.26/4) = 6.7\%$. With our median structural gravity estimate of $\rho = 0.28$ and tariff-based $\epsilon = -5.03$, assuming $\kappa = 0$ implies $t = 5.7\%$. The problem with this assumption is that "home" (trade with self) coefficients are estimated at 1.55. This implies $\kappa = \exp(1.55/5.03) - 1 = 36\%$. Substituting this value back in yields $t = 7.8\%$. This is considerably higher than the current 3.83% weighted world MFN tariff but lower than the 2000 world simple average MFN tariff of 12.8% (both reported by World Bank WDI database). Thus, our results on border effects, RTA impacts, and trade elasticities are mutually consistent with the proposition that the main channel through which RTAs liberalize trade is the elimination of MFN tariffs. The more general point is that to be in line with actual tariffs, the trade elasticity should be somewhat over 10 times the RTA coefficient; our meta-analysis suggests this is indeed the case.

4.3. Partial vs General Equilibrium Impacts on Trade

The consideration of price index and multilateral resistance terms is not only important from the point of view of estimating the correct β for each of the variables that comprise the trade cost determinants. A second point that Anderson and van Wincoop (2003) emphasize is that the indexes change when trade costs change. Thus, merely exponentiating the coefficients on dummy variables (which we will call the Partial Trade Impact,

[32] See Imbs and Méjean (2009) and Feenstra et al. (2010) for recent presentations of the different estimates used by trade and macro economists.

PTI) may not give a reliable estimate of the full impact on trade. Indeed, one of the points emphasized by Anderson and van Wincoop (2003) is that taking into account price index changes leads to substantially smaller trade impacts of borders. The trade impact holding production and expenditure constant but adjusting Φ_n and Ω_i via the contraction mapping does not have an obvious name. It should not really be thought of as a general equilibrium impact because it holds GDP constant and the GDPs depend on factor prices. Anderson (2011) emphasizes the *modular* nature of the structural gravity model: the determination of output and expenditures occurs in a different module from the allocation of bilateral flows. Hence, we will label the trade impact that observes this feature of the model the Modular Trade Impact (MTI). We reserve the title of General Equilibrium Trade Impact (GETI) for the case where wages (and therefore GDPs) also adjust to trade cost changes.

Suppose that B_{ni} is one of the bilateral variables determining τ_{ni}. Further suppose that $\ln \phi_{ni}$ is linear in B_{ni} with coefficient β. We want to see the impact on trade of changing B_{ni} to B'_{ni}. Holding the multilateral terms constant, the ratio of new to original trade is just the ratio of new to original trade freeness. Thus the partial trade impact is given by

$$\text{PTI}_{ni} = \hat{\phi}_{ni} = \phi'_{ni}/\phi_{ni} = \exp[\beta(B'_{ni} - B_{ni})]. \tag{32}$$

Note that $\text{PTI}_{ni} = 0$ for any country pair that does not change bilateral linkages, i.e. $B'_{ni} = B_{ni}$. Thus, the PTI omits third-country effects, which are to be expected because the multilateral resistance terms change whenever other countries change their trade costs.

For any trade equation fitting into structural gravity, the ratio of new bilateral trade, X'_{ni}, to original trade *taking MR changes into account* (but leaving incomes unchanged) is obtained from equation (2) as

$$\text{MTI}_{ni} = \frac{X'_{ni}}{X_{ni}} = \underbrace{\exp[\beta(B'_{ni} - B_{ni})]}_{\text{PTI}} \times \underbrace{\frac{\Omega_i}{\Omega'_i}\frac{\Phi_n}{\Phi'_n}}_{\text{MR adjustment}}. \tag{33}$$

The procedure to implement is therefore to retrieve $\ln \phi_{ni}$, including coefficient β for B_{ni} either using estimates from the literature or estimating ϕ_{ni} through an implementation of equation (22). Then, using ϕ_{ni}, Y_i and X_n in equation (3), a contraction mapping gives us Φ_n and Ω_i. The third step is to do a counterfactual change to B_{ni} (for instance, turn off all RTAs), which results in a new freeness of trade index ϕ'_{ni}. Re-running the contraction mapping provides us with Φ'_n and Ω'_i. We have all the needed elements to calculate X'_{ni}/X_{ni}. Contrary to the PTI approach, a change in a variable specific to a pair of countries using this approach will provide counterfactual changes in trade flows for *all country pairs*.

A growing number of papers, following the initial motivation for structural estimation, do counterfactuals using MTI (although the terminologies vary). Glick and Taylor (2010) is an example where MTI is used to estimate the costs of military conflicts. Anderson and Yotov (2010a) apply this method to assess the impact of an agreement on trade between Canadian provinces that took place in 1995.

An important issue with the MTI is that while (33) does account for changes in MR terms (Φ_n and Ω_i), it assumes constant expenditure (X_n) and output (Y_i) for all countries, which raises a question of interpretation. Recall that $S_i = Y_i/\Omega_i$. Holding Y_i constant, it must be that $S_i' = Y_i/\Omega_i'$. The conceptual problem is that S_i in many models summarized in Table 3.1 depends on wages and exogenous parameters such as quality, A, or technology, T, of all products manufactured in i. Changes in trade costs are not permitted to change wages since that would affect Y_i, but it is peculiar to allow trade costs to change deep parameters.

A second issue is that MTI may omit potentially important effects. For instance, if the thought experiment is the removal of trade costs with a major partner, it is very unlikely that such a drastic change in the trade cost matrix, and therefore in predicted trade flows, would leave incomes unchanged. The MTI remains an interesting entity but we think it also worth calculating the GETI allowing for wage/income changes.

Anderson and van Wincoop (2003) was probably the first paper to calculate the GETI counterfactuals of a removal of national borders, taking into account income changes. Their approach is very related to the "exact hat algebra" methods developed by Dekle et al. (2007) and followers for calculating counterfactual welfare changes. The exact hat algebra approach has a big advantage as a pedagogical tool: it makes it very clear what is the equation driving the wage/income adjustment.[33]

Dekle et al. (2007, 2008) develop a methodology to investigate the consequences in terms of changed wages and welfare of closing trade deficits of all countries. Costinot and Rodriguez-Clare (2013) in this volume show how to adapt the method to determine the welfare impact of trade costs shocks.[34] While the goal of this approach is to provide a quantitative evaluation for welfare, it also yields the GETI as an intermediate step. Here we express the method in terms of our notation and allow for trade deficits (a feature of the data which applications cannot ignore).

The GETI calculation adjusts the income terms Y_i and X_n following the change in trade costs. Denoting $\hat{x} = x'/x$ as the change between new and initial situation of all variables x, the resulting change in bilateral trade is now expressed as

$$\text{GETI}_{ni} = \frac{X_{ni}'}{X_{ni}} = \underbrace{\exp[\beta(B_{ni}' - B_{ni})]}_{\text{PTI}} \times \underbrace{\frac{\Omega_i \Phi_n}{\Omega_i' \Phi_n'}}_{\text{MR adj.}} \times \underbrace{\frac{Y_i' X_n'}{Y_i X_n}}_{\text{GDP adj.}} = \frac{\hat{Y}_i \hat{X}_n}{\hat{\Omega}_i \hat{\Phi}_n} \hat{\phi}_{ni}. \tag{34}$$

To calculate changes in Y, recall that the value of production in the origin country is given by $Y_i = w_i L_i$. Considering the labor endowment as fixed, the change in Y_i will therefore be completely determined by the change in w_i: we have $\hat{w}_i = \hat{Y}_i$. Bilateral trade

[33] Egger and Larch (2011) is an example of a set of papers inspired by Anderson and van Wincoop (2003) that calculate trade effects including a GDP updating step. However, as in the inspiring paper, it does not provide a wage updating equation, making it less transparent what are the assumptions that underlie their approach.

[34] Ossa (2011) and Caliendo and Parro (2012) have related implementations.

is a function of the output of the origin country Y_i, but the expenditure at destination X_n also enters. In general, $X_n \neq Y_n$, because of trade deficits, denoted as D_n. There are different ways to handle the presence of trade deficits, which are all ad hoc in the absence of a fully specified intertemporal model. The most straightforward way to incorporate those deficits is to assume that deficit is exogenously given on a per capita basis, that is $D_n = L_n d_n$. With this assumption (which implies that trade deficits are specified in units of labor of country n), $X_n = w_n L_n (1 + d_n)$, so that $\hat{X}_n = \hat{w}_n = \hat{Y}_n$.

At this stage we therefore need to derive the equilibrium change in income, $\hat{Y}$. Note first that market clearing implies that $\hat{Y}_i = Y_i'/Y_i = \frac{1}{Y_i}\sum_n \pi_{ni}' X_n'$. Recall that $\pi_{ni} = X_{ni}/X_n$ is the share of n's expenditure spent on goods from i. In all the models we call structural gravity, changes in π resulting from trade cost shocks take the following form (first demonstrated in Dekle et al. (2007)):

$$\hat{\pi}_{ni} = \frac{(\hat{Y}_i \hat{\tau}_{ni})^{\epsilon}}{\sum_{\ell} \pi_{n\ell} (\hat{Y}_{\ell} \hat{\tau}_{n\ell})^{\epsilon}}. \tag{35}$$

Plugging this back into the market clearing condition, one can solve for the changes in production of each origin country.

$$\hat{Y}_i = \frac{1}{Y_i} \sum_n \hat{\pi}_{ni} \pi_{ni} \hat{Y}_n X_n = \frac{1}{Y_i} \sum_n \frac{\pi_{ni} \hat{Y}_i^{\epsilon} \hat{\phi}_{ni}}{\sum_{\ell} \pi_{n\ell} \hat{Y}_{\ell}^{\epsilon} \hat{\phi}_{n\ell}} \hat{Y}_n X_n. \tag{36}$$

The method for calculating the GETI involves four steps.

1. Retrieve β as the coefficient on B_{ni} from a gravity equation in which B_{ni} is a dummy for a trade-cost changing event such as a free trade agreement or a currency union formation (or dissolution). An alternative is to take values of the β vector from the literature. If an *ad valorem* trade cost is included in the study, recover the trade elasticity, ϵ. We use $\epsilon = -5.03$, the median value from our meta-analysis (structural gravity results from tariff rates), which is also the source of each β.
2. The exponential of the coefficient is our estimator of the impact of the trade cost change. That is let $\hat{\phi}_{ni} = \exp(\beta)$ for the ni for whom $B_{ni} = 1$ and $\hat{\phi}_{ni} = 1$ for all other ni pairs.
3. Along with the value of production of each country (Y_i), the original trade share matrix (π_{ni}), plug the estimated $\hat{\phi}_{ni}$ into equation (36), which defines a system of equations determining $\hat{Y}_i$ for each country. Using the estimated value of ϵ, substitute the $\hat{\phi}_{ni}$ and $\hat{Y}_i^{\epsilon}$ into equation (35) to derive the matrix of trade changes, $\hat{\pi}_{ni}$. Iterate using a dampening factor until $\hat{\pi}_{ni}$ stops changing.[35]
4. The GETI for each country pair is $\hat{\pi}_{ni}\hat{Y}_n$. The welfare change is $\hat{\pi}_{nn}^{1/\epsilon}$.

[35] Stata® code is provided online.

We implement the methodologies for PTI, MTI, GETI, and welfare calculations just outlined on a dataset of bilateral trade for 84 countries, and the year 2000. The choice of datasets and sample is dictated by the need to include trade with self X_{ii} in order to calculate meaningful MR terms, needed from MTI to welfare computations. With a few exceptions where "true" internal flows are available (such as trade between and within Canadian provinces), trade with self must be inferred from production and export data as $X_{ii} = Y_i - \sum_{n \neq i} X_{ni}$. Calculating this for total trade is difficult, since the GDP of i includes many service sectors that are hardly traded at all. Furthermore GDP, as a value-added measure, excludes purchases of intermediates, which should be included in trade with self. Data for manufacturing industries is more useful, since comparing the value of production with total exports for the same industry raises less issues. We therefore rely on the CEPII trade and production database, developed for De Sousa et al. (2012), and used in Anderson and Yotov (2010b) and Anderson and Yotov (2012) recently for similar purposes. We take the year 2000 because the production data has a very long lag in release dates, and this makes available a larger set of countries with complete data.[36] We aggregate all 23 industries available in the database to obtain an overall manufacturing sector (with the exception of two sectors—misc. petroleum and other manuf.—which seem to exhibit a large share of negative internal flows, probably due to classification errors).

The results of the trade impacts are displayed in Table 3.6. The two first columns simply gives estimated coefficients and PTIs for the set of variables we want to evaluate: RTAs, Currency Unions, Common Language, Colonial Linkage, and the Border Effect. MTI, GETI, and welfare calculations allow for a separate calculation for members and non-members for each variable. For instance, when evaluating RTAs, the GETI for pairs like the United States and Canada that have an RTA is 1.205 whereas it is 0.96 for pairs like the United States and France which do not have an RTA. Egger et al. (2011) apply a similar methodology to a different dataset and obtain GETIs of 1.39 for members and 0.95 for non-members.

The experiment is to turn off all those dummy variables, in order to calculate the counterfactual trade flows for all pairs, and therefore reveal the amount of trade created by those variables under each methodology. Note first that the MTI is systematically smaller than the PTI. The intuition can be illustrated with RTAs. When signing those, PTI only takes the downward impact on τ_{ni}, when MTI also adjusts the MR terms, in particular Φ_n. Because RTAs make access to n easier, competition is fiercer there, raising Φ_n, and counteracting the direct τ_{ni} effect.

Also note that the difference between MTI and GETI is usually quite small, except for the removal of the effect of national borders, which is a much larger shock. This similarity in the two types of estimates was noted in the original work by Anderson

[36] The constraint that internal trade should be available is only binding for the MTI to welfare stages where counterfactual MR and income terms have to be calculated.

Table 3.6 PTI, MTI, GETI, and Welfare Effects of Typical Gravity Variables

	Coeff.	PTI	MTI		GETI		Welfare	
Members:	**Yes**	**Yes**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
RTA/FTA (all)	.28	1.323	1.129	.946	1.205	.96	1.011	.998
EU	.19	1.209	1.085	1.007	1.136	1.001	1.013	.999
NAFTA	.53	1.699	1.367	1.005	1.443	1	1.048	1
Common currency	.98	2.664	1.749	1.028	2.203	1.003	1.025	.998
Common language	.33	1.391	1.282	.974	1.303	.99	1.005	.999
Colonial link	.84	2.316	2.162	.961	2.251	.984	1.004	.999
Border effect	1.55	4.711	4.647	.938	3.102	.681	.795	n/a

Notes: The MTI, GETI, and welfare are the median values of the real/counterfactual trade ratio for countries relevant in the experiment.

and van Wincoop (2003). Although they only report PTI and GETI, their footnote 26 states that the changes in incomes only affect marginally the outcome (even though their experiment removes the Canada–US border). It is also interesting that the results by Anderson and van Wincoop (2003) from the counterfactual removal of the US–Canada border reveals a steep decline when comparing GETI to PTI (2.43 vs 5.26), a finding we also observe in the last row of Table 3.6 (3.1 vs 4.7), using a quite different dataset.

Looking at welfare effects, it is striking that strong trade impacts may have small welfare consequences. The welfare effects in this class of model are linked to the change in the share of trade that takes place inside a country. Therefore a given variable, colonial link for instance, can turn out to have very large factor effects on the considered flows but very small welfare effects overall, because the initial π_{ni} is very small. Intuitively, because the initial flows are so small, even doubling trade with ex-colonies will result in very tiny changes in the share of expenditure that is spent locally. In contrast, adding even a few percentage points of trade with a major partner will be much more important for welfare.

Finally, it should be kept in mind that the GETI and welfare results shown in Table 3.6 are intended for exposition of the methods, rather than as definitive calculations. There are very important omissions in the analytical framework we used: it lacks sector-level heterogeneity in ϵ, input-output linkages, and other complexities that could alter results in a substantial way. Costinot and Rodriguez-Clare (2013) provide a more complete treatment of the question in their chapter dedicated to welfare effects (see Chapter 4).

4.4. Testing Structural Gravity

The GETI approach to quantifying trade impacts of various policy changes builds a counterfactual world based on a general equilibrium modeling of the economy. Structural gravity is the common core of this modeling. Anderson and van Wincoop (2003) rely

on the CES-NPD version of it, Dekle et al. (2007, 2008) or Caliendo and Parro (2012) use the heterogeneous industries version, Bergstrand et al. (2013) and Egger et al. (2011) use the CES-MC view, but all those GETI-related exercises rely on structural gravity, and hence need it to hold empirically. It is also true of Arkolakis et al. (2012b) welfare gains formula, since the assumptions underlying structural gravity overlap to a large degree with the assumptions of that formula. However none of those papers actually test for the empirical relevance of it: the usual approach is to assume it holds, estimate or calibrate a value of ϵ, and then run the counterfactual.

Anderson and Yotov (2010b, 2012) propose that estimated fixed effects can be used to validate the structural gravity model and hence to justify its use for comparative statics. They regress the *estimated* FEs on their counterparts constructed using structural gravity theory and bilateral trade cost estimates. They find very high R^2 and interpret this as confirmation of the theory. One way to think about the issue is that if fixed effects mainly arise due to data issues or unobserved multilateral trade costs, then the estimated fixed effects might be expected to show little relationship to their theoretical determinants. We see some important caveats. The most important is a point raised by Fally (2012). Anderson and Yotov (2012) use Poisson pseudo-maximum likelihood (PML) to estimate the fixed effects and gravity coefficients. Fally (2012) shows that the use of Poisson PML has an unintended consequence: it leads to a perfect fit between the fixed effects and their structural gravity counterparts (the MR terms). To be more precise, if $Y_i = \sum_n X_{ni}$ and $X_n = \sum_n X_{ni}$ as implied by the market clearing and budget allocation assumptions, then $\hat{S}_i = Y_i/\hat{\Omega}_i$ and $\hat{M}_n = X_n/\hat{\Phi}_n$, when using PPML as an estimator for $\hat{S}_i$ and $\hat{M}_n$. The test is therefore bound to succeed perfectly if using this estimation procedure. Even putting that problem on the side, there are important issues to be mentioned with that approach.

First, fit that comes merely from size effects cannot be interpreted as support for the theory. Trade has to go somewhere so larger countries must export and import more as a matter of *accounting* identities, not theory. The real challenge should focus on whether theory-constructed Φ_n and Ω_i are good predictors of the importer and exporter fixed effects after they have been purged by the size effects of X_n and Y_i. In Anderson and Yotov (2010b, 2012), the resistance indexes appear to have much smaller coefficients than the size effects when theory states that they should have the same unit elasticities.

Lai and Trefler (2002) propose a related and potentially devastating critique of structural gravity. Although they specifically address only the CES monopolistic competition model, their results apply to all the models that yield observationally equivalent multilateral resistance terms. The crux of their argument is to show that changes in their constructed price term (a combination of our ϕ_{ni}, Φ_n, and Ω_i) "literally contributes nothing to the analysis of changing trade patterns." They illustrate this finding with a scatter plot showing no relationship between changes in trade and changes in a CES-based computed price index. The price term aggregates tariffs, which Lai and Trefler (2002) established

earlier in the paper to have strong effects on bilateral trade.[37] Thus, it is surprising that a tariff-based index term cannot predict trade changes.

Table 3.1 helps to clarify the underlying issue. It shows two versions of the Armington CES model, with and without an outside good. In both cases the fixed effect gravity equation would estimate the same trade elasticity based on bilateral tariffs. However, the "content" of those fixed effects would be very different. With standard CES preferences, the importer term is an index of tariffs. Hence, under the assumptions of that model, changes in that index should lead to changes in trade. On the other hand, with an outside good, the importer fixed effect is just 1 and is obviously not a function of tariffs. We speculate that the absence of tariff-index effects found by Lai and Trefler (2002) does not invalidate CES but rather the homothetic version without an outside good. Nevertheless, the standard CES model is too entrenched—partly because it is so useful!—that it will not be abandoned based on one finding. It seems clear that more research that follows up on Lai and Trefler (2002) is needed to verify just how much violence the structural gravity model does to the data.

5. FRONTIERS OF GRAVITY RESEARCH

This section investigates three areas of ongoing research. The first issue is how to appropriately model the error term in the gravity equation, in particular considering the problem of heteroskedasticity in multiplicative models. The second topic is the appropriate estimation response to large numbers of zero trade flows, a phenomena at odds with a model in which predicted trade is a multiple of strictly positive numbers. The last item covers the rising use of firm-level trade data with its associated set of new issues regarding estimation and interpretation.

5.1. Gravity's Errors

Part of the original attraction of the gravity equation—and of other multiplicative models such as the Cobb-Douglas production function—is that after taking logs they can be estimated with OLS. Santos Silva and Tenreyro (2006) (hereafter SST) brought to the attention of the field that this seemingly innocuous approach involves taking a much stronger stance on the functional form of the error than we do when estimating truly linear models with OLS.

SST frame the problem in terms of heteroskedasticity but this begs the question of which error is not homoskedastic. There are two ways of expressing the error in a gravity equation. Suppose that the exporter and importer fixed effects as well as all determinants of ϕ_{ni} have been combined into a k-length vector $\mathbf{z}_{ni}$ and that the coefficients on these variables are vector $\boldsymbol{\zeta}$. The conventional way to express the error is as the *difference* between

[37] Their tightly estimated elasticity of −5 is almost the same mean as found in the Section 4.1 meta-analysis comprising hundreds of estimates.

data and prediction: $\varepsilon_{ni} \equiv X_{ni} - \exp(\mathbf{z}'_{ni}\boldsymbol{\zeta})$. The second way to express the error is as a ratio of data to prediction: $\eta_{ni} \equiv X_{ni}/\exp(\mathbf{z}'_{ni}\boldsymbol{\zeta})$.

After taking logs, the linear regression error term is given by $\ln X_{ni} - \mathbf{z}'_{ni}\boldsymbol{\zeta} = \ln \eta_{ni}$. In standard OLS regressions, heteroskedasticity of $\ln \eta_{ni}$ is a minor concern. The grave concern is whether $\ln \eta_{ni}$ is independent from the $\mathbf{z}_{ni}$. SST point out that if the variance of η depends on $\mathbf{z}_{ni}$ then the log transformation will prevent $\ln \eta_{ni}$ from having a zero conditional expectation and will therefore lead to inconsistent coefficient estimates in linear (in logs) regression.

Should we then try to minimize the sum of the squared residuals, $\hat{\varepsilon}^2_{ni}$, using non-linear least squares? A homoskedastic additive error is an unappealing assumption. It defies common sense to think that deviations of true trade from predicted trade would be of the same order of magnitude for the US and Aruba. Moreover, SST find that non-linear least squares (NLLS) performs very badly in Monte Carlo simulations.

SST argue that Poisson PML is an attractive alternative to linear-in-logs OLS for multiplicative models like the gravity equation. Poisson is not the only PML that could be applied to gravity equations and SST also consider Gamma PML. To understand what each estimator is doing it is useful to compare their first-order conditions side by side.

Actual exports are given by $X_{ni} = \exp(\mathbf{z}'_{ni}\boldsymbol{\zeta})\eta_{ni}$, where η is a multiplicative error term. Using $\sum$ to represent summation over all ni dyads, the moment conditions for the Poisson PML, OLS, and Gamma PML are

$$\underbrace{\sum \mathbf{z}_{ni} \cdot (X_{ni} - \hat{X}_{ni}) = \mathbf{0}}_{\text{Poisson}}, \quad \underbrace{\sum \mathbf{z}_{ni} \cdot (\ln X_{ni} - \ln \hat{X}_{ni}) = \mathbf{0}}_{\text{OLS}}, \quad \underbrace{\sum \mathbf{z}_{ni} \cdot (X_{ni}/\hat{X}_{ni} - 1) = \mathbf{0}}_{\text{Gamma}}, \tag{37}$$

where $\hat{X}_{ni} \equiv \exp(\mathbf{z}'_{ni}\hat{\boldsymbol{\zeta}})$ denotes the prediction for X_{ni} conditional on the observables. The first set of first-order conditions are the ones used for Poisson "true" maximum likelihood estimator (MLE) on count data. Comparing with the OLS first-order conditions we see that the Poisson involves *level* deviations of X_{ni} from its expected value whereas the OLS involves *log* deviations.[38] The term in parentheses in the Gamma PML first order condition (FOC) is just the percent deviation of actual trade from predicted trade. Since percent deviations are approximately equal to log deviations, the Gamma PML pursues an objective that is very similar to that of OLS shown in equation (37). A useful feature of the two PMLs is that FOC permit the inclusion of zeros, unlike the linear-in-logs form. However, we delay treatment of the zero issue to the following section so as to focus on the role of assumptions about the error term.

Both the Poisson and Gamma PML deliver consistent $\boldsymbol{\zeta}$ estimates regardless of the distribution of η_{ni} so long as $\mathbb{E}[X_{ni} \mid \mathbf{z}_{ni}] = \exp(\mathbf{z}'_{ni}\boldsymbol{\zeta})$. The question of which one is

[38] Wooldridge (2010, p. 741) provides further detail on the robustness and efficiency properties of Poisson PML. SST have provided responses to a variety of potential concerns about the Poisson PML estimator on their "log of gravity" page.

more efficient depends on how the variance of X_{ni} relates to its expected value. Consider the following (fairly) general case:

$$\text{Var}[X_{ni} \mid \mathbf{z}_{ni}] = h\mathbb{E}[X_{ni} \mid \mathbf{z}_{ni}]^{\lambda}. \tag{38}$$

If $\lambda = 1$, a case we will call the Constant Variance to Mean Ratio (CVMR) assumption, then Poisson PML is efficient. The CVMR assumption is a generalization of the Poisson variance assumption in which $h = \lambda = 1$. The Gamma PML is the efficient PML if $\lambda = 2$, that is if the standard deviation is proportional to the mean. We will therefore refer to a DGP that satisfies $\lambda = 2$ as one that adheres to the Constant Coefficient of Variation (CCV) assumption. As the log-normal has a CCV, this provides the intuition for why the Gamma PML estimates tend to be similar to the OLS (on logs), since the latter is the MLE under the assumption of homoskedastic log-normality (which we abbreviate as log-normality). Given that both Poisson and Gamma PML are consistent under the same conditional expectation assumption, their estimates, $\hat{\boldsymbol{\zeta}}$, should be approximately the same if the sample is large enough. Their estimates will only converge on the OLS estimates under log-normality of η_{ni}.

Poisson and Gamma PML remain consistent (and efficient for the corresponding cases for λ) even if $h > 1$, i.e. what is called "over-dispersion." Thus, the finding that variance exceeds the mean does not justify use of estimators such as the negative binomial, suggested by De Benedictis and Taglioni (2011). This estimator is alluring because it has Poisson as a special case but estimates what appears to be more general variance function with a parameter estimating the amount of over-dispersion. We urge researchers to resist the siren song of the Negative Binomial. The most important reason, pointed out by Boulhol and Bosquet (2013), is that Negative Binomial PML estimates depend on the units of measurement for the dependent variable. The web appendix uses actual data to show that measuring trade in thousands of dollars instead of billions not only leads to large changes in the magnitudes of estimated elasticities, it even reverses the signs on some of the indicator variables.[39]

Here we conduct Monte Carlo simulations that re-express some key insights derived from SST. We illustrate attractive robustness features of PML estimators. For each repetition of the simulation we also estimate a test statistic proposed by Manning and Mullahy (2001) to diagnose the error term. This MaMu test—referred to by SST as a "Park-type" test—takes the log of equation (38) and replaces the variance and expected value terms with their sample counterparts to obtain

$$\ln \hat{\varepsilon}_{ni}^2 = \text{constant} + \lambda \ln \hat{X}_{ni}, \tag{39}$$

where $\hat{\varepsilon} = X_{ni} - \exp(\mathbf{z}'_{ni}\hat{\boldsymbol{\zeta}})$ and $\ln \hat{X}_{ni} = \mathbf{z}'_{ni}\hat{\boldsymbol{\zeta}}$. Equation (39) is estimated using OLS.

[39] Other drawbacks of the negative binomial include: (a) as pointed out in Wooldridge (2010, p. 738), the one-step form of Stata®'s **negbin** command lacks the robustness properties of the other PML, and (b) even the two-step method does not nest the CVMR assumption.

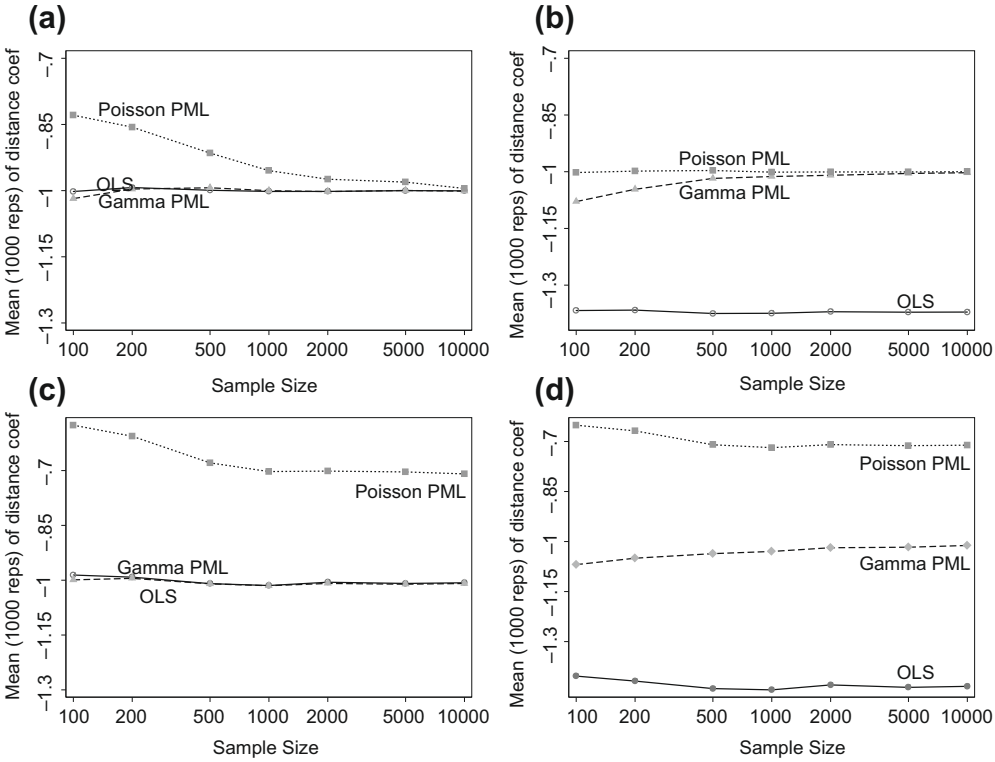

Figure 3.3 Monte Carlo Investigation of PMLs: (a) Log-normal Homoskedastic (CCV); (b) Homoskedastic (CVMR); (c) Model Mis-specification (CCV); (d) Model Mis-specification (CVMR)

In order to focus on issues related to the distribution of the error term, the DGP does not contain i, n, and mi-level components. Rather, as with the Monte Carlos of SST, it is a single-dimensional cross-section. Also following SST, we include a continuous trade determinant, denoted Dist for distance and assumed to be log-normal, as well as a binary variable, called RTA. The results for the binary variable did not offer additional insights but they are available to interested readers by running the program, which is available on the chapter companion website.

Figure 3.3 displays results for four versions of the following data generating process (where u_i denotes a standard normal pseudo-random term):

$$X_i = \exp[\theta_i \ln \text{Dist}_i + 0.5\text{RTA}_i + \sigma_i \times u_i].$$

Versions (a) and (b) have the usual distance elasticity ($\theta_i = -1$) and differ regarding the assumed error term. Case (a) considers log-normal errors with a constant variance parameter $\sigma = 2$. Case (b) departs from log-normal η by assuming a constant mean-variance ratio (CVMR), i.e. heteroskedasticity *á la* Poisson. The σ_i is set to satisfy $\text{Var}[X_i \mid \text{RTA}_i, \text{Dist}_i] = h \times \exp[-1 \ln \text{Dist}_i + 0.5\text{RTA}_i]$. Case (c) reverts to log-normal errors with

$\sigma = 2$ but introduces a mis-specification. True $\theta_i = -0.5$ for $\text{Dist}_i < \overline{\text{Dist}}$ (where we set $\overline{\text{Dist}}$ equal to the median distance) and $\theta_i = -1.5$ for $\text{Dist}_i \geq \overline{\text{Dist}}$. Case (d) uses the same break in the distance elasticity but follows case (b) in using a CVMR error. The regressions in (c) and (d) are mis-specified because they estimate a constant distance elasticity.

The two left panels, (a) and (c), of Figure 3.3 consider the error term structure that is most favorable to OLS and Gamma, that of log-normality with a σ parameter that is constant across all observations. The key result in panel (a) is that Poisson PML underestimates the (absolute) magnitude of the distance effect (and, while not shown, the RTA effect), but the estimates converge on the true value as the sample size rises. The bias exhibited by Poisson PML in panel (a) is increasing in the assumed σ parameter—we have assumed a realistic value of $\sigma = 2$. With $\sigma = 1$, Poisson shows very little bias even in relatively small samples.

Panel (b) of Figure 3.3 replicates the key finding of SST that OLS on log exports becomes an inconsistent estimator in the presence of heteroskedasticity. When the variance of exports is proportional to the mean, OLS overestimates distance and RTA effects. Fortunately, both of the PMLs estimate the effects of distance and RTAs consistently. But now it is the Gamma PML that shows small-sample bias.

Figure 3.3(c) and (d) use a DGP that does not appear in SST, one that features a major error in the specification of the conditional expectation. This DGP features small distance elasticities for flows that travel less than the median distance but much larger (in absolute value) distance elasticities at longer distances. Under such mis-specification, the Poisson and Gamma would be expected to have different probability limits. In the version of this specification with log-normal errors, we find that both OLS and Gamma PML estimate distance elasticities about -1, the average of the short- and long-distance elasticities. In contrast, because Poisson's FOC emphasizes absolute deviations, it puts more emphasis on the high-expected-trade observations, delivering an elasticity, -0.7, that is much closer to the short-distance elasticity. The final simulation shown in Figure 3.3(d) combines the CVMR error with the distance elasticity mis-specification. The two PMLs are hardly changed compared to frame (c), but OLS now estimates the absolute distance elasticity to be over 1.3, as was the case in frame (b).

The Monte Carlo simulations also attest to the usefulness of estimating the MaMu regression. We find that in the log-normal DGP that $\hat{\lambda} \approx 2$. On the other hand, under the CVMR DGP, $\hat{\lambda} \approx 1.6$ with a range of 1.55 to 1.66 if there are 10,000 observations. Even though the MaMu regression does not robustly estimate true λ, it appears to be a reliable method for distinguishing between the two DGPs. Estimates of $\hat{\lambda}$ significantly below two were a near perfect predictor of a CVMR DGP.[40]

Our Monte Carlo results suggest that rather than selecting the Poisson PML as the single "workhorse" estimator of gravity equations, it should be used as part of a robustness-

[40] Specifically under log-normality, in only 6 in 1000 cases with a sample size of 10,000 did the MaMu test find $\lambda < 2$ at the 5% significance. With CVMR errors, the MaMu test rejects $\lambda = 2$ in all repetitions.

exploring ensemble that also includes OLS and Gamma PML. Upon comparing the results of each method, we suggest the following conclusions be drawn.

1. If all three estimates are similar, then we can relax because the model appears to be well specified and η is approximately log-normal with a constant σ parameter.
2. If the Poisson and Gamma PML coefficients are similar to each other and both are distinct from the OLS, then it is reasonable to conclude that heteroskedasticity is a problem and the OLS estimates are unreliable.
3. If the Gamma and OLS coefficients on trade cost proxies are similar and the Poisson coefficients are smaller in absolute magnitude—as occurs in Figure 3.3 (a) and (c)—our simulations suggest two possible interpretations.
   - **(a)** If the root mean squared error is large and sample size is not very large, this pattern might be arising from small-sample bias of the Poisson PML.
   - **(b)** If the sample size is large enough to dismiss small-sample bias, then trade cost elasticities may be falling in absolute value as trade itself rises.
4. More generally, major divergence in large samples between Poisson and Gamma PML—as exhibited in Figure 3.3 (c) and (d)—can signal model mis-specification.

5.2. Causes and Consequences of Zeros

The structural gravity models we have considered in this paper express trade as the multiple of strictly positive variables. Hence, they do not naturally generate zero flows. Most actual trade datasets exhibit substantial fractions of zeros, which become more frequent with disaggregation at the firm or product level. Haveman and Hummels (2004) is an early paper tabulating the frequency of zeros. Even at the country level, Helpman et al. (2008) report that country pairs that do not trade with each other or trade in only one direction account for about half the observations. The high frequency of zeros calls for two things. First, we need to adjust our trade models in order to accommodate zeros since they are an important feature of the data. Second, we need to revise our methods of estimation to allow for consistent estimates in the presence of a dependent variable that takes on zeros frequently.

There a number of possible modifications to the structural gravity model to incorporate zeros. The simplest approach is to assume that zeros are simply a data recording issue, i.e. that there are no "structural zeros" but only "statistical zeros." This would occur due to rounding or declaration thresholds. Structural models of zeros mainly work by adding a fixed cost of exporting a positive amount from i to n. In the Chaney (2008) model, fixed costs are not enough to cause zeros because of the assumption of a continuum of firms with unbounded productivity. Helpman et al. (2008) truncate the productivity distribution and this leads to zeros for some dyads. In contrast, Eaton et al. (2012) generate zeros by abandoning the assumption of a continuum of firms. With a finite number of draws there will be (in realization) a maximum productivity firm even if the productivity distribution has infinite upper support. If the most productive firm from i cannot export profitably to n, then there will be no trade between these countries. Also, as we noted

in Section 2.3, with a finite number of consumers, each selecting a single supplier, there will be realizations of the random utility model in which two countries do not trade.

These models all share the feature that zeros are more likely when bilateral trade is expected to be low, i.e. between distant and/or small countries.[41] Unobserved trade costs will endogenously create zeros. When taking logs of the zeros we remove those observations. That leads to the systematic selection bias illustrated in Table 3.3. For this reason, it is important to determine which estimators can deliver good results even when zeros are an endogenous component of the data generating process.

We now proceed to consider several candidate estimating methods prior to judging them using a Monte Carlo simulation. One commonly used method that does not deserve Monte Carlo treatment is the practice of adding one to observed exports and then taking logs. This gives a lower limit of 0 so Tobit is sometimes applied. The method should be avoided because results depend on the units of measurement. Thus, the interpretation of coefficients as elasticities is lost. In the web appendix we show that distance elasticities range from -1.93 to -0.09 as we change the exports units from dollars to billions of dollars. The estimated impact of common currencies switches from negative and significant to positive and significant simply by changing units from millions to billions.

Eaton and Tamura (1994) developed an early solution to incorporate zeros that can be thought of as a model of $\ln(a + X_{ni})$ where instead of arbitrarily setting $a = 1$, it is instead treated as a parameter to be estimated. One could think of a as a fixed amount of trade that "melts" away before the trade flow is measured by government. More formally, the method, which we refer to as ET Tobit, defines a strictly positive latent variable X^*_{ni} and a threshold a such that when $X^*_{ni} > a$ we observe $X_{ni} = X^*_{ni} - a$ and when $X^*_{ni} \leq a$ we observe $X_{ni} = 0$. Unfortunately, $\hat{a}$ lacks a compelling structural interpretation. Another drawback of ET Tobit is that it is not a "canned" program.

Eaton and Kortum (2001) propose another method that has the advantage of being both easier to implement and interpret. Suppose that there is minimum level of trade, a, such that if "ideal" trade, X^*_{ni}, falls below a we observe $X_{ni} = 0$ but otherwise we observe $X_{ni} = X^*_{ni}$. Each a_n is estimated as the minimum X_{ni} for a given n, which we denote as $\underline{X}_{ni}$. To estimate the model, all the observed zeros in X_{ni} are replaced with $\underline{X}_{ni}$ and the new bottom-coded $\ln X_{ni}$ is the dependent variable in a Tobit command that allows for a user-specified lower limit of $\ln \underline{X}_{ni}$. The EK Tobit, as we will refer to this method, has the advantages of (a) not requiring exclusion restrictions and (b) being easily estimable using Stata®'s **intreg** command.

Helpman et al. (2008) take a Heckman-based approach to zeros. This involves first using probit to estimate the probability that n imports a positive amount from i. The second step estimates the gravity equation on the positive-flow observations including a selection correction. A challenge, common to Heckman-based methods, is that it is

[41] Baldwin and Harrigan (2011) find this pattern of zeros in the US product-level trade data.

difficult to find an exclusion restriction. Thus, one ideally would like to use a variable in the export status probit that theory tells you can be excluded from the gravity equation. Since both equations have country fixed effects, this variable needs to be dyadic in nature. Helpman et al. (2008) consider overlap in religion and the product of dummies for low entry barriers in countries i and n. While their model deals with zeros, the main focus of their method is to remove the effect of the extensive firm margin so as to estimate intensive margin effects. Thus, they are designed to uncover a different set of parameters than the other approaches which estimate coefficients that combine extensive and intensive margins. Consequently, we omit this method from the Monte Carlo simulations.

In any model that abandons the continuum assumption, the market shares π_{ni} that appear in all structural gravity formulations should be reinterpreted as *expectations*. For a wide class of models featuring finite numbers of buyers and sellers, we conjecture that it is reasonable to stipulate $\mathbb{E}(X_{ni}/X_n) = \pi_{ni}$. In that case, the appropriate estimator is the Multinomial PML, a solution advanced by Eaton et al. (2012) for the case of a finite number of firms. Fortunately, as proven in unpublished notes by Sebastian Sotelo, the Multinomial PML can be estimated by applying the **poisson** command to the market share variable X_{ni}/X_n, along with country-specific fixed effects.[42] By using a dependent variable that divides raw trade by the importing country's total expenditure, the multinomial pseudo-maximum likelihood (MNPML) accords less importance to large levels of trade. This is because the biggest dyadic flows tend to be imported by countries with large aggregate expenditures. Shares prevent this dependent variable from obtaining values over one.[43]

Since one of the original draws of the Poisson PML method was that it allows for easy incorporation of zeros, we will consider the performance of both Poisson and Gamma PML in the Monte Carlo simulation. Previous simulation evidence had produced mixed results. While Poisson PML performs well in Santos Silva and Tenreyro (2006), their simulation uses statistical zeros, obtained via rounding. Santos Silva and Tenreyro (2011) propose a mixture model to generate zeros. Total bilateral exports are given as the product of a random number of exporters and a random level of exports per firm. Santos Silva and Tenreyro (2011) set the share of zeros between 62% and 83% by choosing high variance parameters for the assumed negative binomial count distribution determining the number of exporters. Even with such high zero frequencies, they find both Poisson and Gamma PML outperform alternatives such as linear-in-logs OLS (on the positives), log of one plus exports, and the ET Tobit. These simulations make it clear that the mere presence of large shares of zeros does not undermine the performance of PML estimators such as Poisson and Gamma.

[42] The Eaton et al. (2012) specification also includes country dummies interacted with a dummy for trade with self ($n = i$).

[43] The potential drawback is the maintained assumption of an expenditure elasticity of one. In gravity models featuring quasi-linear utility for instance, that elasticity is zero.

The issue left unresolved by Santos Silva and Tenreyro (2011) is whether a DGP that followed modern theory by explicitly featuring fixed costs might be problematic for the PML estimators. In particular, the number of exporting firms should not be purely random but should instead depend on trade costs and market sizes, just as the volume of exports does. Martin and Pham (2011) consider DGPs involving threshold values and find that Tobit and Heckman methods outperform the Poisson PML. However, as noted by Santos Silva and Tenreyro (2011), their DGP is not multiplicative so it does not embed the fundamental problem of linear estimation in the presence of heteroskedasticity.

We consider a DGP that takes as its starting point the structural gravity model. We make a straightforward modification of the heterogeneous firms version of structural gravity seen in Section 2.3.2 so that it can generate zeros. The simple idea is that profits for firm α from i exporting to market n in the CES monopolistic competition model are given by $x_{ni}(\alpha)/\sigma - f_n$. The threshold level of sales at which zero profits would be earned is $x^*_{ni} = \sigma f_n$. Therefore if the initial prediction for aggregate trade X_{ni} falls below σf_n then it would be impossible for any firm to enter and break even. The result would be an observation of $X_{ni} = 0$. Thus this data generating process corresponds to the assumptions of the EK Tobit so long as the disturbances are log-normally distributed.

We do not observe the market-specific entry costs but instead assume that f_n is log-normal, with mean and variance parameters chosen so as to replicate the 25% of zeros in the DOTS data for 2006 that we also used in the first Monte Carlo exercise of Section 3. The procedure is also very much in line with the first Monte Carlo, modified to generate and account for the zeros. The assumed parameters on log distance and RTAs are maintained at -1 and 0.5 respectively, such that $\tau_{ni}^{-\theta} = \exp(-\ln \text{Dist}_{ni} + 0.5\text{RTA}_{ni})\eta_{ni}$. As before the model has an error term η_{ni} that is assumed to come from unobserved variable trade costs. We first specify η_{ni} as a homoskedastic log-normal term and then consider a second specification in which η_{ni} is heteroskedastic, such that bilateral trade has a constant variance to mean ratio (CVMR).

The contraction mapping algorithm generating simulated trade flows requires both i and n incomes, combined with bilateral trade freeness, $\phi_{ni} = \tau_{ni}^{-\theta} f_n^{-\left[\frac{\theta}{\sigma-1}-1\right]}$ in this model. Based on our meta-analysis in Section 4.2, we assume $\theta = 5$. Based on Eaton et al. (2011a) we set $\theta/(\sigma - 1) = 2.5$. Combining these assumptions implies $\sigma = 3$. This completes the set of data and parameters needed to generate the predicted aggregate trade X_{ni}, which is cut to zero when falling under σf_n.

Table 3.7 shows the performance of six candidate estimation methods in the presence of zero trade flows. It begins with the most commonly used approach of taking logs and running least squares on the logs of the positive values of bilateral trade. This omits 25% of the sample and does so in a highly selective manner. Under both error DGPs, the coefficients are biased, by about 45%. This result was already anticipated in Table 3.3 in the column where we removed the smallest 25% of the observations and found a 20% bias for distance and a 30% bias for RTA.

Table 3.7 Monte Carlo Results with 25% Structural[a] Zeros

Estimates:	Distance (−1)		RTA (0.5)		Bias[c] (%)	
Error:	Log-normal	CVMR[b]	Log-normal	CVMR	Best	Worst
LSDV on $\ln(X)$ positives	−0.81	−1.07	0.63	0.69	45	46
	[0.02]	[0.01]	[0.06]	[0.03]		
ET Tobit: $\ln[a + X_{ni}]$	−0.94	−1.06	0.53	0.68	12	43
	[0.02]	[0.01]	[0.06]	[0.03]		
EK Tobit: $\ln[X_n^{\min}]$ for 0s	−0.99	−1.23	0.50	0.57	1	36
	[0.02]	[0.01]	[0.06]	[0.03]		
Poisson PML	−0.73	−1.00	0.29	0.50	0	70
	[0.14]	[0.00]	[0.43]	[0.01]		
Gamma PML	−1.05	−1.10	0.41	0.38	23	34
	[0.04]	[0.03]	[0.11]	[0.07]		
Multinomial PML (EKS)	−0.79	−1.00	0.36	0.50	0	49
	[0.06]	[0.02]	[0.15]	[0.03]		

Notes: Mean estimates based on 1000 repetitions. The true parameters are −1 for distance and .5 for RTA. Standard deviation of estimate in "[]". All estimators include exporter and importer fixed effects.
[a]DGP sets trade flows to 0 when $X_{ni} < \sigma f_n$.
[b]CVMR is a Poisson-like error with a constant variance to mean ratio.
[c]"Bias" is the absolute bias in percentage points, with "Best" being the error process that minimizes bias for a given estimator.

The Eaton and Tamura (1994) Tobit-like method estimating $\ln(a+X)$ via MLE works better than LSDV under the maintained assumption of log-normal errors but remains biased. The second Tobit we consider generates better results. Since the EK Tobit is also easier to estimate and has a sound structural interpretation, it dominates the ET Tobit. However, it remains inconsistent under the CVMR assumption. In that case Poisson or Multinomial PMLs are unbiased.

The first and third columns of Table 3.7 reveal that the Poisson PML are biased toward zero under the log-normal DGP. Cameron and Trivedi (1998, Chapter 9, pp. 281–282) prove that Poisson can obtain consistent estimates in panel data even with the number of years fixed and the number of individuals going to infinity. Charbonneau (2012) provides an analytic proof that *with two-way fixed effects*, Poisson PML suffers from an incidental parameters problem. However, we do not think this is the precise problem here. The reason is that the simulation results shown in the second and fourth columns show that Poisson with country effects is unbiased with CVMR errors. Furthermore, the underestimates under log-normality appeared in the previous section even in the absence of fixed effects. We conjecture that asymptotic properties of PPML are not achieved due to the high coefficient of variation in the simulation (calibrated on real data) and the need to estimate $(170 - 1) \times 2 = 338$ importer and exporter effects.

Gamma PML does badly in the presence of the CVMR DGP and is even biased under the log-normal assumption, where it had performed well in the previous simulation.

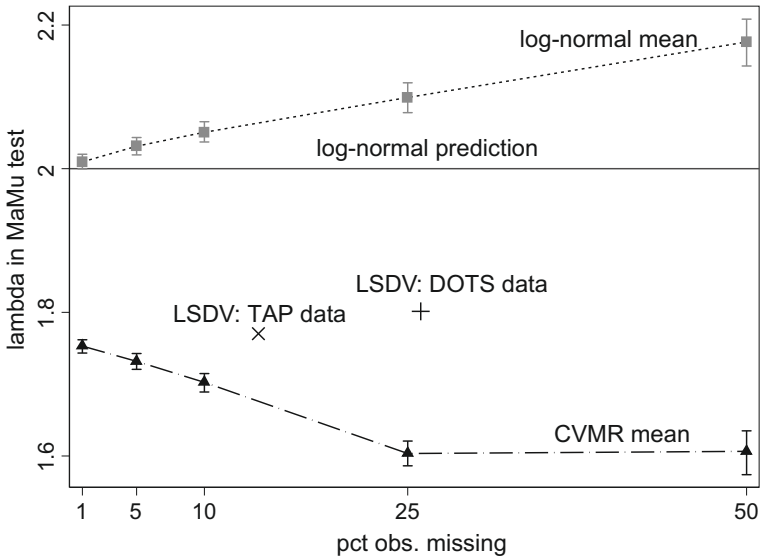

Figure 3.4 Discriminating Between Different DGPs with Structural Zeros

Evidently, the presence of zeros undermines its performance here. The most positive comment on Gamma PML in the presence of structural zeros is that it exhibits the lowest worst-case bias (34%).

The selection of the appropriate estimator therefore appears to be contingent on the process generating the error term. Under the CVMR we would want to use Poisson or Multinomial PML but under log-normality EK Tobit is preferred. This points to the potential of the MaMu test for log-normal errors introduced in the previous subsection. We ran 1000 repetitions of the MaMu test for DGPs featuring varying shares of "structural zeros." We estimated λ by applying OLS to the logged squared residuals from the LSDV model. The results are reported in Figure 3.4.

Since LSDV excludes the zeros, there is reason to be doubtful that a MaMu test based on LSDV errors would be unbiased. Figure 3.4 shows that as the percent of zeros increases, the expected value of $\hat{\lambda}$ under log-normality departs from the true value of two. LSDV estimates give $\hat{\lambda}$ average 2.1 when the DGP is set to reflect the percent of zeros in the DOTS data (25%). Figure 3.4 shows that, when the error term follows the CVMR DGP, $\hat{\lambda}$ is even worse at estimating the true λ, which in that case is one. This is not due to zeros, as we found similar bias in the previous section without zeros. Indeed, raising the share of zeros brings LSDV-based $\hat{\lambda}$ closer to the true value. More importantly, as shown by the 99% confidence intervals on point, there is no overlap in the $\hat{\lambda}$ from the two error structures.

In sum, while the MaMu test delivers biased estimates of λ under both DGPs, it can nevertheless be used to distinguish between log-normal and CVMR with perfect accuracy in 1000 repetitions of each DGP. A finding that $\hat{\lambda} \geq 2$ suggests EK Tobit is the estimator best matched to data where zeros are generated by bilateral fixed entry costs. In contrast a $\hat{\lambda}$ significantly less than 2 militates for Poisson or Multinomial PML, with the preference going to MNPML since its worst-case performance is better than Poisson.

A puzzle illustrated in Figure 3.4 is that the $\hat{\lambda}$ we obtain from regression on real data lie between the simulation predictions. Export data on all goods from the 2006 IMF Direction of Trade Statistics (DOTS) and manufactured goods in 2000 from the Trade and Production (TAP) data yield strikingly similar estimates of 1.77 and 1.79, respectively. This could point to a mixture distribution or to a process, such as the multinomial, that gives intermediate results.

5.3. Firm-Level Gravity, Extensive and Intensive Margins

With the simultaneous emergence of the heterogeneous firms modeling framework and firm-level trade data, questions about the margins of adjustment to trade shocks have become important in the literature. Researchers became interested in whether, after a rise in trade costs, or a fall in final demand for instance, the global trade fall comes from all firms reducing their individual flows, or on the contrary from exit of the smallest exporters. A recent example is Bricongne et al. (2012), who apply this decomposition to the 2008–2009 trade collapse to find that most of the adjustment came from existing firms cutting their shipments rather than from exit. There are however different ways to decompose aggregate exports. To determine the most useful way, we need keep in mind what different models predict.

The first extensive/intensive margin definition was proposed by Eaton et al. (2004), Hillberry and Hummels (2008), and Bernard et al. (2007) and uses the identity that total exports equals the number of active exporters multiplied by average shipments: $X_{ni} = N_{ni}\bar{x}_{ni}$. The total elasticity with respect to trade costs is therefore the sum of the elasticities of these two factors[44]:

$$\frac{\partial \ln X_{ni}}{\partial \ln \tau_{ni}} = \frac{\partial \ln N_{ni}}{\partial \ln \tau_{ni}} + \frac{\partial \ln \bar{x}_{ni}}{\partial \ln \tau_{ni}}. \tag{40}$$

This decomposition respects the traditional use of the extensive margin terminology as being the change in the number of exporters, but the use of the intensive margin is unconventional. It seems more in keeping with traditional usage to limit "intensive margin" changes to the individual responses of firms following the change in trade

[44] Bernard et al. (2007), Mayer and Ottaviano (2007), and Bernard et al. (2011) have analyzed finer decompositions, taking into account multiproduct firms. Models of such firms are covered in this handbook by Melitz and Redding (see Chapter 1).

costs. In (40), $\frac{\partial \ln \bar{x}_{ni}}{\partial \ln \tau_{ni}}$ contains this effect, but confounds it with the change in average shipments that comes from the changing composition of exporters. We therefore want to split this term itself into two margins, the intensive and compositional. Using $\bar{x}_{ni} = \frac{1}{G(\alpha^*_{ni})} \int_0^{\alpha^*_{ni}} x_{ni}(\alpha) g(\alpha) d\alpha$, and using the Leibniz rule as in Chaney (2008), it can be shown that[45]:

$$\frac{\partial \ln X_{ni}}{\partial \ln \tau_{ni}} = \underbrace{\frac{\partial \ln N_{ni}}{\partial \ln \tau_{ni}}}_{\text{ext. margin}} + \underbrace{\frac{1}{\bar{x}_{ni}} \left(\int_0^{\alpha^*_{ni}} \frac{\partial \ln x_{ni}(\alpha)}{\partial \ln \tau_{ni}} x_{ni}(\alpha) \frac{g(\alpha)}{G(\alpha^*_{ni})} d\alpha \right)}_{\text{int. margin}} + \underbrace{\frac{\partial \ln G(\alpha^*_{ni})}{\partial \ln \alpha^*_{ni}} \frac{\partial \ln \alpha^*_{ni}}{\partial \ln \tau_{ni}} \left(\frac{x_{ni}(\alpha^*_{ni})}{\bar{x}_{ni}} - 1 \right)}_{\text{compos. margin}}. \quad (41)$$

This three-way decomposition nests the one proposed by Eaton et al. (2004), Hillberry and Hummels (2008), and Bernard et al. (2007). In their decomposition, they simply add up the intensive and compositional ones and call it intensive. It also nests the alternative decomposition proposed by Chaney (2008), which is obtained when summing up our extensive and compositional and calling it the extensive.[46] The extensive and intensive margins in (41) have the classical respective interpretations. The compositional margin is caused by the fact that new entrants/exitors do not have the same productivity as the existing exporters. This margin is a function of the difference between the marginal firm, with shipments $x_{ni}(\alpha^*_{ni})$ and the average shipment before the shock, $\bar{x}_{ni}$. This percentage difference is weighted in the overall effect by the change in the distribution of firms associated with changes in trade costs (through changes in the cutoff).

Up to this point, the decomposition is purely definitional and does not depend on specifics of the model nor on the assumed distribution of heterogeneity. Also important is that the two first margins can be measured directly. The extensive margin is the elasticity of the number of exporters (from i to n) with respect to trade costs, and the intensive margin is the elasticity of the *average shipments of the incumbent firms*, that is the firms that

[45] The web appendix provides the derivation.

[46] Chaney (2008) starts from aggregate trade $X_{ni} = N_i \int_0^{\alpha^*_{ni}} x_{ni}(\alpha) dG(\alpha)$, and proceeds to decomposing between the elasticity of shipments due to incumbent exporters, and the one caused by entrants/exitors:

$$\frac{\partial \ln X_{ni}}{\partial \ln \tau_{ni}} = \frac{\tau_{ni}}{X_{ni}} \left(N_i \int_0^{\alpha^*_{ni}} \frac{\partial x_{ni}(\alpha)}{\partial \tau_{ni}} dG(\alpha) \right) + \frac{\tau_{ni}}{X_{ni}} \left(N_i x_{ni}(\alpha^*_{ni}) g(\alpha^*_{ni}) \frac{\partial \alpha^*_{ni}}{\partial \tau_{ni}} \right).$$

were exporting before the shock and still do afterwards.[47] One can therefore calculate these two margins and back out the third one as a residual to quantify the share of each. This "margins accounting" can in principle be done independently of the underlying foundation for gravity, or even with models that do not have closed-form solutions for those margins.

What should we expect for the value of the different margin elasticities? This will depend on modeling assumptions naturally. There are two types of such assumptions that are usually made: one has to do with the underlying constant price elasticity (CES + iceberg) modeling, the other imposes the heterogeneity in productivity to be distributed Pareto. Let us proceed by imposing those sequentially, and in that order.

5.3.1. Margins with a CES-Iceberg (Constant Price Elasticity) Model

If the price elasticity is constant, the intensive margin term simplifies to $\frac{\partial \ln x_{ni}(\alpha)}{\partial \ln \tau_{ni}}$ (which factors out of the integral in (41)). In the Melitz/Chaney model of heterogeneous firms exporting to multiple countries, a firm located in i and indexed by its unitary input coefficient α exports the following value to country n:

$$x_{ni}(\alpha) = \left(\frac{\sigma}{\sigma - 1}\right)^{1-\sigma} (\alpha w_i \tau_{ni})^{1-\sigma} \frac{X_n}{\Phi_n}. \tag{42}$$

The intensive margin will therefore be $1 - \sigma$.[48]

To calculate what the theory predicts for the extensive margin, we need to write equilibrium N_{ni}. Since $N_{ni} = G(\alpha^*_{ni})N_i$,

$$\frac{\partial \ln N_{ni}}{\partial \ln \tau_{ni}} = \frac{\partial \ln G(\alpha^*_{ni})}{\partial \ln \alpha^*_{ni}} \frac{\partial \ln \alpha^*_{ni}}{\partial \ln \tau_{ni}}. \tag{43}$$

The first elasticity in this product requires an assumption on the distribution of heterogeneity which we will turn to below. As can be seen in (17), the second elasticity is -1 regardless of distributions and follows from the iceberg assumption. Since profits in a given market depend on the product $\alpha w \tau$, to hold profits equal to zero, any increase in τ must be matched by an exactly proportionate decrease in α. Using this result also allows to simplify the compositional margin, such that we have now the following decomposition:

$$\frac{\partial \ln X_{ni}}{\partial \ln \tau_{ni}} = \underbrace{-\frac{\partial \ln G(\alpha^*_{ni})}{\partial \ln \alpha^*_{ni}}}_{\text{ext. margin}} + \underbrace{1 - \sigma}_{\text{int. margin}} + \underbrace{\frac{\partial \ln G(\alpha^*_{ni})}{\partial \ln \alpha^*_{ni}} \left(1 - \frac{x_{ni}(\alpha^*_{ni})}{\bar{x}_{ni}}\right)}_{\text{compos. margin}}. \tag{44}$$

[47] Incumbents is a slight abuse of language here. Strictly speaking, the relevant set of firms in the model is the one of firms that fall below the cost cutoff both before and after the trade cost shock, and therefore is defined in terms of productivity draw, rather than on initial presence in the market.

[48] These elasticities reflect the partial trade impact of a change in trade costs, defined as PTI above, since they hold Φ_n, X_n, and w_i constant when changing τ_{ni}. This is the natural partial effect to consider since fixed effects for each i and n effectively hold those attributes constant.

5.3.2. Margins with a CES-Iceberg Model and Pareto

Any progress on evaluating the expected values of the three elasticities in (44) requires an assumption on $G()$, the distribution of productivity. The literature almost universally uses the Pareto, which offers the very convenient feature of a constant elasticity of the CDF with respect to the cutoff, $\frac{\partial \ln G(\alpha_{ni}^*)}{\partial \ln \alpha_{ni}^*} = \theta$. The web appendix shows that in that case, the deviation of the marginal firms' exports from the average exports is inversely related to θ. The two θ cancel, leaving $\sigma - 1$ as the compositional margin:

$$\frac{\partial \ln X_{ni}}{\partial \ln \tau_{ni}} = \underbrace{-\theta}_{\text{ext. margin}} + \underbrace{1-\sigma}_{\text{int. margin}} + \underbrace{\sigma - 1}_{\text{compos. margin}} . \tag{45}$$

Hence the overall elasticity is $-\theta$, which comes from the fact that the compositional margin exactly compensates the intensive margin, so that the effect of a change in trade costs on average shipments is zero. The intuition is that a rise in trade costs should reduce export flows by all incumbent exporters (the intensive margin), which reduces the average exports. However, the same rise in trade costs causes the weakest firms to exit, which in turn raises average exports (the compositional margin). The fact that the second effect exactly compensates the first is an artifact of the Pareto distribution. We speculate that under other distributions than Pareto, the distributional margin would not be so strong as to compensate fully the intensive margin. We will return to empirical evidence of this below.

Equation (45) also sheds new light on the traditional practice of calculating the margins using (40), i.e. the impact of gravity variables on the number of exporters and the average shipments. Indeed, since this second impact is predicted to be zero in a strict version of the Melitz/Chaney model, one should actually obtain that the extensive margin is systematically 100% of the total effect.

Using data collected in the EU-funded project EFIGE, we have calculated average exports and number of exporters of three origin countries (France, Belgium, and Norway) to each destination country and regress those on the most traditional gravity proxies, GDP and distance to obtain an idea of those margins.[49] Results (available on the chapter companion website) show that the extensive margin is a dominant part of the overall effects in all samples, and for both variables. This is not an isolated finding. Using the same method of decomposition, Bernard et al. (2007), Mayer and Ottaviano (2007), Hillberry and Hummels (2008), and Lawless (2010) all point to the extensive margin accounting for most of the total elasticities of most gravity variables. However it is not 100% as the strict version of the theory would predict. Eaton et al. (2011a) show that under Pareto-heterogeneity, average exports are proportional to fixed cost of market

[49] All elasticities with respect to X_n (proxied by GDP) have theoretical predictions that are more complicated than the ones on τ_{ni} (proxied by distance). The main issue is that it is not tenable to use the PTI for those, holding Φ_n constant when changing GDP of n.

entry. Thus, one interpretation of the margins regressions is that such costs are rising in GDP and declining in distance. While plausible for GDP, it would be strange indeed for distance to raise variable trade costs but *lower* fixed entry costs. An alternative inference is that heterogeneity is not Pareto. In that case the intensive margin effects of GDP (positive) and distance (negative) are not completely compensated by opposite effects of the compositional margin. This alternative explanation strikes us to be at least as plausible.

Another advantage of the three-way decomposition (45) over the two-way (40) is that it is more handy if one wants to estimate structural parameters of the model. For instance, with firm-level exports and trade costs data, one can estimate the elasticity of the number of exporters to recover θ. Then change the dependent variable to the average shipments of firms that remain exporters over the whole sample to estimate σ. The two-way decomposition by Chaney (2008) offers the same advantage, and permits the same structural estimations except that one needs to estimate the overall elasticity to recover θ, and aggregate rather than average exports to recover σ.[50] Crozet and Koenig (2010) use firm-level regressions of the same theoretical setup to estimate the structural parameters from the equations for export values, productivity distributions and export probabilities, so as to calculate the Chaney (2008) margins. Interestingly, and in line with the arguments above, Crozet and Koenig (2010) find the share of the extensive margin using Chaney's method to be much smaller than what the literature has found using the first method. Also, they do find a large variance in the shares of the two margins across sectors, a finding hard to reconcile with the decomposition method using (40).

While the intensive margin using the margins decomposition is one method to estimate the parameter σ, there is a more direct way, using firm-level shipments. Firm-level trade data typically takes the form of exports values reported by the national customs administration for each firm over a certain number of years. While it would be very valuable to be able to put together several of those national datasets, confidentiality issues make it very unlikely to happen any time soon. Taking logs of (42), dropping the source country index, and adding a time dimension and a properly behaved error term, one obtains

$$\ln x_{nt}(\alpha) = (1-\sigma)\ln\left(\frac{\sigma}{\sigma-1}\right)+(1-\sigma)\ln(\alpha_t w_t)+(1-\sigma)\ln\tau_{nt}+\ln(X_{nt}/\Phi_{nt})+\varepsilon_{nt}(\alpha). \tag{46}$$

The first point to note is that there are two sources of identification for $1-\sigma$: one from the cost component of the firm ($\alpha_t w_t$), the other one from international price shifters (τ_{nt}). Let us focus on τ_{nt} first. The regression will need to capture both some firm-time level determinant and some destination-time one. It is quite clear from equation (46) that no ideal structure of fixed effects will work, since τ_{nt} and X_{nt}/Φ_{nt} vary along the

[50] Berman et al. (2012) use a related approach to evaluate the margins in a model with variable markups where they don't have closed-form solutions for the margins.

same dimensions. One path is to introduce firm-destination effects, that capture the time invariant determinants of αw and X_n/Φ_n, but also any part of $\varepsilon_{nt}(\alpha)$ that does not change over time. The regression can then identify the effect of τ_{nt} from the variation over time (the regression should also include proxies for changes in demand of the destination and efficiency of the firm). Such changes in trade costs can come from trade policy naturally, and there are databases (listed in the web appendix) which can be used to measure changes in applied tariffs by different destination countries. Moreover, any bilateral price shifter would in theory have the same impact: freight rates for instance also reveal the trade cost elasticity. Fitzgerald and Haller (2012) and Berman et al. (2012) estimate this elasticity using firm-level shipments for Irish and French exports respectively. The price shifters in Fitzgerald and Haller (2012) are the real exchange rate, and tariff changes from 2000 to 2004. The impact of tariffs seems to be of the same order of magnitude as the aggregate literature, with an elasticity around −5. Interestingly enough, the coefficient on the exchange rate is much lower, between 0.8 and 1, which is very similar to what Berman et al. (2012) find for French firms. This discrepancy is reminiscent of findings in the aggregate literature.

An interesting case to consider for firm-level exports is when exports of a certain good originate from one country of production only (Scotch whisky would be an example). We can then write

$$x_n(\alpha) = \frac{(\alpha\tau_n)^{1-\sigma}}{N_i \int_0^{\alpha_n^*} (\alpha\tau_n)^{(1-\sigma)} dG(\alpha)} X_n = \frac{\alpha^{1-\sigma}}{N_i V_n} X_n, \quad \text{with} \quad V_n = \int_0^{\alpha_n^*} \alpha^{1-\sigma} dG(\alpha). \tag{47}$$

The trade costs affects all competitors equally in the destination market, and therefore drops out of the export value equation. In that extreme case the predicted response of trade flows to trade costs is just zero, even though the true price elasticity is $1-\sigma$. The only case where the trade elasticity of individual exporters with respect to trade costs will be $1-\sigma$ is when the exporting country considered does not affect Φ_n, and is therefore a marginal player in the considered industry. This is not only true for firm-level exports, but also for industry-level gravity equations. Therefore when trying to estimate the trade elasticity with respect to trade costs, one should be careful about the degree of monopoly that different exporting countries have on world markets. In the limit if a country is the only exporter of a given good, rising tariffs cannot affect its market share. As a consequence, different coefficients on tariffs across industries can be a noisy estimate of different values of σ or θ characterizing the sectors. The difference in coefficients might come from differences in the concentration of supply across industries.

The other source of identification of the trade elasticity is the coefficient on α, the inverse of the firm's productivity. More generally, any cost-shifter in this model is entirely transmitted in the delivered price of the firm, and cuts sales by $1-\sigma$ percent. Pure cost shifters are however rarely measured at the level of the firm. Let us be as general as possible,

and index firms by a "performance" variable s, that shifts utility by a factor γ and raises marginal costs with elasticity λ. Crozet et al. (2012) show that s then impacts individual shipments with elasticity $(\gamma - \lambda)(\sigma - 1)$. The demand parameter σ is now grouped with the quality elasticities. Even estimating the compound parameters poses a challenge because of a selection bias inherent to this whole class of models involving selection into export markets. To see this, we need to add the error term to the estimated model. To simplify exposition, let us continue with firms originating from one country only:

$$\ln x_n(s) = (\gamma - \lambda)(\sigma - 1)\ln s + \ln (X_n/N_i V_n) + \varepsilon_n(s). \tag{48}$$

Crozet et al. (2012) model $\varepsilon_n(s)$ as a firm-destination demand shifter. The econometrician does not observe the quality of the match between a firm's variety and the destination consumer's tastes, which is what $\varepsilon_n(s)$ is capturing. Since only firms with $x_n(s) \geq f_n\sigma$ can enter country n profitably, it is clear from (48) that the firms that are active in n despite a small observed s must have a high $\varepsilon_n(s)$ and vice-versa. This creates a negative correlation between s and $\varepsilon_n(s)$, hence a downward selection bias on s. This issue can be resolved using the EK Tobit method described in Section 5.2. Crozet et al. (2012) show that for the case of exports by Champagne producers, the bias is quite large. They also use Monte Carlo simulations to show that the magnitude of the expected bias is actually very similar when assuming alternative error structures (logistic, gumbel, and exponential) and in line with the bias found in the data.

6. DIRECTIONS FOR FUTURE RESEARCH

Predicting which topics will turn out to be fertile for future research is never easy. However, based on our assessment of the current set of problems and unresolved issues we offer three suggestions.

First, the underlying determinants of trade costs remain poorly understood. We are comfortable with transport costs and tariffs yet we have reason to believe that neither are the most important determinants of trade costs. First, distance effects are too large and have the wrong functional form to be determined by freight costs. Second, border effects are large even along borders where tariffs are very small. Other variables such as language and common currency have impacts on trade that seem very large compared to any reasonable accounting of the costs that different languages or different currencies impose. We believe that authors need to dig deeper to understand what underlies these impacts.

The second topic that is attracting growing interest is the dynamics of trade. All the micro-foundations of gravity that we examined are static models. They provide a derivation for a cross-section but are questionable bases for panel estimation. This raises the econometric problem of how to handle the evolution of trade over time in response to changes in trade costs. More fundamentally, we need to think more about how to

incorporate short-run capacity effects, learning, sunk costs, and other dynamic phenomena into the gravity equation framework.

The final topic has been lurking throughout our derivations of the micro-foundations. In every model there came a point where very specific functional forms were imposed in order to maintain tractability. The constant elasticity of substitution model for preferences is nearly ubiquitous. Where it is less important, specific forms for heterogeneity (Fréchet, Pareto) are often essential. Finally theorists have often resorted to modeling firms using a continuum. Given the immense size of firms like Airbus or Boeing, it is an embarrassment to stipulate that all firms have zero mass and act as if they had no influence on the price index. Future research will need to devise ways to investigate the consequences of departing from these assumptions and ways to test whether the data clearly reject the current set of restrictions customarily imposed mainly for tractability rather than realism.

7. CONCLUSION

The use of gravity equations to understand bilateral trade patterns exemplifies the beneficent roles of empirical regularities in guiding theory development and theory in guiding estimation. Our graphic displays of the systematic distance and size effects in trade data show the empirical appeal of the gravity equation. We have cataloged the diverse set of microfoundations that deliver "structural gravity," our label for a formulation that matches stylized facts while calling for a more sophisticated estimation approach than the one initially employed. After a quantitative summary of 1000s of prior estimates, we illustrate the use of the structural form to determine the complete trade and welfare impacts of policy changes. Our selective survey of topics at the frontier of current research suggests that a great deal of interesting work lies ahead.

REFERENCES

Abowd, J., Kramarz, F., Margolis, D., 1999. High wage workers and high wage firms. Econometrica 67 (2), 251–333.

Ahlfeldt, G., Redding, S., Sturm, D., Wolf, N., 2012. The Economics of Density: Evidence from the Berlin Wall, manuscript.

Anderson, J., 1979. A theoretical foundation for the gravity equation. The American Economic Review 69 (1), 106–116.

Anderson, J., 2011. The gravity model. The Annual Review of Economics 3 (1), 133–160.

Anderson, J., Marcouiller, D., 2002. Insecurity and the pattern of trade: an empirical investigation. Review of Economics and Statistics 84 (2), 342–352.

Anderson, J.E., van Wincoop, E., 2003. Gravity with gravitas: a solution to the border puzzle. The American Economic Review 93 (1), 170–192.

Anderson, J., Yotov, Y., 2010a. The changing incidence of geography. American Economic Review 100, 2157–2186.

Anderson, J., Yotov, Y., 2010b. Specialization: Pro-and Anti-Globalizing, 1990–2002. Working Paper 16301, NBER.

Anderson, J.E., Yotov, Y.V., February 2012. Gold Standard Gravity. Working Paper 17835, NBER.

Anderson, S., De Palma, A., Thisse, J., 1992. Discrete Choice Theory of Product Differentiation. MIT Press.

Arkolakis, C., Costinot, A., Donaldson, D., Rodríguez-Clare, A., 2012a. The Elusive Pro-Competitive Effects of Trade, manuscript.
Arkolakis, C., Costinot, A., Rodriguez-Clare, A., 2012b. New trade models, same old gains? American Economic Review 102 (1), 94–130.
Armington, P.S., 1969. A theory of demand for products distinguished by place of production. Staff Papers, International Monetary Fund 16 (1), 159–178.
Baier, S.L., Bergstrand, J.H., 2001. The growth of world trade: tariffs, transport costs, and income similarity. Journal of International Economics 53 (1), 1–27.
Baier, S., Bergstrand, J., 2007. Do free trade agreements actually increase members' international trade? Journal of International Economics 71 (1), 72–95.
Baier, S., Bergstrand, J., 2009. Bonus vetus OLS: a simple method for approximating international trade-cost effects using the gravity equation. Journal of International Economics 77 (1), 77–85.
Baier, S., Bergstrand, J.H., 2010. Approximating general equilibrium impacts of trade liberalizations using the gravity equation. In:Van Bergeijk, P., Brakman, S. (Eds.), The Gravity Model in International Trade: Advances and Applications. Cambridge University Press, pp. 88–134 (Chapter 4).
Baker, M., Fortin, N.M., 2001. Occupational gender composition and wages in Canada, 1987–1988. The Canadian Journal of Economics 34 (2), 345–376.
Baldwin, R., 2006. The Euro's Trade Effects. Technical Report, European Central Bank.
Baldwin, R., Harrigan, J., 2011. Zeros, quality, and space: trade theory and trade evidence. American Economic Journal: Microeconomics 3 (2), 60–88.
Baldwin, R., Taglioni, D., 2007. Trade effects of the euro: a comparison of estimators. Journal of Economic Integration 22 (4), 780–818.
Behrens, K., Mion, G., Murata, Y., Südekum, J., 2009. Trade, Wages, and Productivity. Technical Report 7369, CEPR.
Bergstrand, J., 1985. The gravity equation in international trade: some microeconomic foundations and empirical evidence. The Review of Economics and Statistics 67 (3), 474–481.
Bergstrand, J.H., Egger, P., Larch, M., 2013. Gravity redux: estimation of gravity-equation coefficients, elasticities of substitution, and general equilibrium comparative statics under asymmetric bilateral trade costs. Journal of International Economics 89 (1), 110–121.
Berman, N., Martin, P., Mayer, T., 2012. How do different exporters react to exchange rate changes? The Quarterly Journal of Economics 127 (1), 437–492.
Bernard, A., Eaton, J., Jensen, J., Kortum, S., 2003. Plants and productivity in international trade. American Economic Review 93 (4), 1268–1290.
Bernard, A.B., Jensen, J.B., Redding, S.J., Schott, P.K., 2007. Firms in international trade. Journal of Economic Perspectives 21 (3), 105–130.
Bernard, A.B., Redding, S.J., Schott, P.K., 2011. Multiproduct firms and trade liberalization. The Quarterly Journal of Economics 126 (3), 1271–1318.
Berthou, A., Fontagné, L., 2013. How do multi-product exporters react to a change in trade costs? Scandinavian Journal of Economics 115 (2), 326–353.
Boulhol, H., Bosquet, C., 2013. Applying the GLM variance assumption to overcome the scale-dependence of the Negative Binomial QGPML Estimator. Econometric Reviews, posted online: 14 June 2013.
Bricongne, J.-C., Fontagné, L., Gaulier, G., Taglioni, D., Vicard, V., 2012. Firms and the global crisis: French exports in the turmoil. Journal of International Economics 87 (1), 134–146.
Broda, C., Weinstein, D.E., 2006. Globalization and the gains from variety. Quarterly Journal of Economics 121 (2), 541–585.
Caliendo, L., Parro, F., 2012. Estimates of the Trade and Welfare Effects of NAFTA. Technical Report 18508, NBER.
Cameron, A., Trivedi, P., 1998. Regression Analysis of Count Data. Cambridge University Press.
Chaney, T., 2008. Distorted gravity: the intensive and extensive margins of international trade. American Economic Review 98 (4), 1707–21.
Charbonneau, K.B., 2012. Multiple Fixed Effects in Nonlinear Panel Data Models Theory and Evidence, Princeton mimeo, November.
Chen, H., Kondratowicz, M., Yi, K.-M., 2005. Vertical specialization and three facts about U.S. international trade. The North American Journal of Economics and Finance 16 (1), 35–59.

Chen, N., Novy, D., 2011. Gravity, trade integration, and heterogeneity across industries. Journal of International Economics 85 (2), 206–221.
Cipollina, M., Salvatici, L., 2010. Reciprocal trade agreements in gravity models: a meta-analysis. Review of International Economics 18 (1), 63–80.
Coeurdacier, N., Martin, P., 2009. The geography of asset trade and the euro: insiders and outsiders. Journal of the Japanese and International Economies 23 (2), 90–113.
Combes, P.-P., Lafourcade, M., Mayer, T., 2005. The trade-creating effects of business and social networks: evidence from France. Journal of International Economics 66 (1), 1–29.
Costinot, A., Donaldson, D., Komunjer, I., 2012. What goods do countries trade? A quantitative exploration of Ricardo's ideas. Review of Economic Studies 79 (2), 581–608.
Crozet, M., Koenig, P., 2010. Structural gravity equations with intensive and extensive margins. Canadian Journal of Economics/Revue Canadienne d'Économique 43 (1), 41–62.
Crozet, M., Head, K., Mayer, T., 2012. Quality sorting and trade: firm-level evidence for French wine. Review of Economic Studies 79 (2), 609–644.
Deardorff, A., 1984. Testing trade theories and predicting trade flows. In: Jones, R., Kenen, P.B. (Eds.), Handbook of International Economics, vol. 1. Elsevier, pp. 467–517.
De Benedictis, L., Taglioni, D., 2011. The gravity model in international trade. In: De Benedictis, L., Salvatici, L. (Eds.), The Trade Impact of European Union Preferential Policies: An Analysis Through Gravity Models. Springer, pp. 55–90 (Chapter 4).
Dekle, R., Eaton, J., Kortum, S., 2007. Unbalanced trade. American Economic Review 97 (2), 351–355.
Dekle, R., Eaton, J., Kortum, S., 2008. Global rebalancing with gravity: measuring the burden of adjustment. IMF Staff Papers 55 (3), 511–540.
de Sousa, J., Lochard, J., 2011. Does the single currency affect foreign direct investment? The Scandinavian Journal of Economics 113 (3), 553–578.
De Sousa, J., Mayer, T., Zignago, S., 2012. Market access in global and regional trade. Regional Science and Urban Economics 42 (6), 1037–1052.
Disdier, A.-C., Head, K., 2008. The puzzling persistence of the distance effect on bilateral trade. The Review of Economics and Statistics 90 (1), 37–48.
Eaton, J., Kortum, S., 2001. Trade in capital goods. European Economic Review 45 (7), 1195–1235.
Eaton, J., Kortum, S., 2002. Technology, geography, and trade. Econometrica 70 (5), 1741–1779.
Eaton, J., Tamura, A., 1994. Bilateralism and regionalism in Japanese and U.S. trade and direct foreign investment patterns. Journal of the Japanese and International Economies 8 (4), 478–510.
Eaton, J., Kortum, S., Kramarz, F., 2004. Dissecting trade: firms, industries, and export destinations. The American Economic Review 94 (2), 150–154.
Eaton, J., Kortum, S., Kramarz, F., 2011a. An anatomy of international trade: evidence from French firms. Econometrica 79 (5), 1453–1498.
Eaton, J., Kortum, S., Neiman, B., Romalis, J., 2011b. Trade and the Global Recession. Technical Report, NBER.
Eaton, J., Kortum, S., Sotelo, S., 2012. International Trade: Linking Micro and Macro. Technical Report, NBER.
Egger, P., Larch, M., 2011. An assessment of the Europe agreement's effects on bilateral trade, GDP, and welfare. European Economic Review 55 (2), 263–279.
Egger, P., Larch, M., Staub, K.E., Winkelmann, R., 2011. The trade effects of endogenous preferential trade agreements. American Economic Journal: Economic Policy 3 (3), 113–43.
Erkel-Rousse, H., Mirza, D., 2002. Import price elasticities: reconsidering the evidence. Canadian Journal of Economics 35 (2), 282–306.
Ethier, W.J., 1982. National and international returns to scale in the modern theory of international trade. The American Economic Review 72 (3), 389–405.
Fajgelbaum, P., Grossman, G., Helpman, E., 2011. Income distribution, product quality, and international trade. Journal of Political Economy 119 (4), 721–765.
Fally, T., 2012. Structural Gravity and Fixed Effects. University of Colorado note, June.
Feenstra, R.C., 1994. New product varieties and the measurement of international prices. The American Economic Review 84 (1), 157–177.

Feenstra, R.C., 2003. A homothetic utility function for monopolistic competition models, without constant price elasticity. Economics Letters 78 (1), 79–86.
Feenstra, R.C., 2004. Advanced International Trade: Theory and Evidence. Princeton University Press, Princeton, New Jersey.
Feenstra, R.C., Markusen, J.R., Rose, A.K., 2001. Using the gravity equation to differentiate among alternative theories of trade. Canadian Journal of Economics 34 (2), 430–447.
Feenstra, R., Obstfeld, M., Russ, K., 2010. In Search of the Armington Elasticity. University of California-Davis, mimeo.
Fitzgerald, D., Haller, S., 2012. Exporters and Shocks, manuscript.
Frankel, J., 2010. The estimated trade effects of the euro: why are they below those from historical monetary unions among smaller countries? In: Alesina, A., Giavazzi, F. (Eds.), Europe and the Euro. University of Chicago Press, pp. 169–212 (Chapter 5).
Frankel, J., Stein, E., Wei, S., 1997. Regional Trading Blocs in the World Economic System. Institute for International Economics, Washington, DC.
Glick, R., Rose, A.K., 2002. Does a currency union affect trade? The time-series evidence. European Economic Review 46 (6), 1125–1151.
Glick, R., Taylor, A.M., 2010. Collateral damage: trade disruption and the economic impact of war. Review of Economics and Statistics 92, 102–127.
Guimaraes, P., Portugal, P., 2010. A simple feasible alternative procedure to estimate models with high-dimensional fixed effects. Stata Journal 10 (4), 628–649.
Hallak, J.C., 2006. Product quality and the direction of trade. Journal of International Economics 68 (1), 238–265.
Hanson, G., 2005. Market potential, increasing returns and geographic concentration. Journal of International Economics 67 (1), 1–24.
Harrigan, J., 1996. Openness to trade in manufactures in the OECD. Journal of International Economics 40 (1–2), 23–39.
Haveman, J., Hummels, D., 2004. Alternative hypotheses and the volume of trade: the gravity equation and the extent of specialization. The Canadian Journal of Economics/Revue Canadienne d'Économique 37 (1), 199–218.
Head, K., Mayer, T., 2000. Non-Europe: the magnitude and causes of market fragmentation in the EU. Review of World Economics 136 (2), 284–314.
Head, K., Mayer, T., 2004a. The empirics of agglomeration and trade. Handbook of Regional and Urban Economics 4, 2609–2669.
Head, K., Mayer, T., 2004b. Market potential and the location of Japanese investment in the European Union. Review of Economics and Statistics 86 (4), 959–972.
Head, K., Ries, J., 2001. Increasing returns versus national product differentiation as an explanation for the pattern of US–Canada trade. American Economic Review 91 (4), 858–876.
Head, K., Ries, J., 2008. FDI as an outcome of the market for corporate control: theory and evidence. Journal of International Economics 74 (1), 2–20.
Head, K., Mayer, T., Ries, J., 2009. How remote is the offshoring threat? European Economic Review 53 (4), 429–444.
Head, K., Mayer, T., Ries, J., 2010. The erosion of colonial trade linkages after independence. Journal of International Economics 81 (1), 1–14.
Helliwell, J., 1998. How Much Do National Borders Matter? Brookings Institution Press, Washington, D.C.
Helpman, E., Melitz, M., Rubinstein, Y., 2008. Estimating trade flows: trading partners and trading volumes. Quarterly Journal of Economics 123 (2), 441–487.
Hillberry, R., Hummels, D., 2008. Trade responses to geographic frictions: a decomposition using micro-data. European Economic Review 52 (3), 527–550.
Hortaçsu, A., Martinez-Jerez, F.A., Douglas, J., 2009. The Geography of trade in online transactions: evidence from eBay and Mercado Libre. American Economic Journal: Microeconomics 1 (1), 53–74.
Hummels, D., 1999. Towards a Geography of Trade Costs. Technical Report 17, GTAP Working Paper.
Imbs, J., Méjean, I., 2009. Elasticity Optimism. Technical Report, CEPR.
Isard, W., Peck, M., 1954. Location theory and international and interregional trade theory. The Quarterly Journal of Economics 68 (1), 97–114.

Jacks, D., Meissner, C., Novy, D., 2008. Trade costs, 1870–2000. The American Economic Review 98 (2), 529–534.
Krugman, P., 1979. Increasing returns, monopolistic competition, and international trade. Journal of International Economics 4, 469–479.
Krugman, P., 1995. Increasing returns, imperfect competition and the positive theory of international trade. In: Grossman, G.M., Rogoff, K. (Eds.), Handbook of International Economics, vol. 3. Elsevier, pp. 1243–1277.
Krugman, P., 1997. Development, Geography, and Economic Theory, vol 6. The MIT Press.
Lai, H., Trefler, D., 2002. The Gains from Trade with Monopolistic Competition: Specification, Estimation, and Mis-Specification. Working Paper 9169, NBER, September.
Lawless, M., 2010. Deconstructing gravity: trade costs and extensive and intensive margins. Canadian Journal of Economics/Revue Canadienne d'Économique 43 (4), 1149–1172.
Leamer, E., Levinsohn, J., 1995. International trade theory: the evidence. In: Grossman, G.M., Rogoff, K. (Eds.), Handbook of International Economics, vol. 3. Elsevier, pp. 1339–1394.
Manning, W., Mullahy, J., 2001. Estimating log models: to transform or not to transform? Journal of Health Economics 20 (4), 461–494.
Martin, P., Mayer, T., Thoenig, M., 2008. Make trade not war? Review of Economic Studies 75 (3), 865–900.
Martin, P., Mayer, T., Thoenig, M., 2012. The geography of conflicts and free trade agreements. American Economic Journal: Macroeconomics 4 (4), 1–35.
Martin, W., Pham, C.S., 2011. Estimating the Gravity Model When Zero Trade Flows Are Frequent. Technical Report, World Bank.
Martin, P., Rey, H., 2004. Financial super-markets: size matters for asset trade. Journal of International Economics 64 (2), 335–361.
Mayer, T., Ottaviano, G., 2007. The Happy Few: The Internationalisation of European Firms. Bruegel Blueprint Series.
McCallum, J., 1995. National borders matter: Canada–US regional trade patterns. The American Economic Review 85 (3), 615–623.
Melitz, M.J., 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71 (6), 1695–1725.
Melitz, M., Ottaviano, G., 2008. Market size, trade, and productivity. Review of Economic Studies 75 (1), 295–316.
Novy, D., 2013. International trade without CES: estimating translog gravity. Journal of International Economics 89 (2), 271–282.
Okawa, Y., Van Wincoop, E., 2010. Gravity in International Finance. Working Paper 7, Hong Kong Institute for Monetary Research.
Okawa, Y., van Wincoop, E., 2012. Gravity in international finance. Journal of International Economics 87 (2), 205–215.
Ossa, R., 2011. A New trade theory of GATT/WTO negotiations. Journal of Political Economy 119 (1), 122–152.
Ossa, R., 2012. Why Trade Matters After All. Working Paper 18113, NBER.
Ottaviano, G., Tabuchi, T., Thisse, J., 2002. Agglomeration and trade revisited. International Economic Review 43 (2), 409.
Portes, R., Rey, H., 2005. The determinants of cross-border equity flows. Journal of International Economics 65 (2), 269–296.
Portes, R., Rey, H., Oh, Y., 2001. Information and capital flows: the determinants of transactions in financial assets. European Economic Review 45 (4–6), 783–796 (15th Annual Congress of the European Economic Association).
Rauch, J.E., Trindade, V., 2002. Ethnic Chinese networks in international trade. The Review of Economics and Statistics 84 (1), 116–130.
Redding, S., Venables, T., 2004. Economic geography and international inequality. Journal of International Economics 62 (1), 53–82.
Romalis, J., 2007. Nafta's and Cusfta's impact on international trade. Review of Economics and Statistics 89 (3), 416–435.

Rose, A., 2000. One money, one market: the effect of common currencies on trade. Economic policy 15 (30), 7–46.
Rose, A., 2004. Do we really know that the WTO increases trade? The American Economic Review 94 (1), 98–114.
Santos Silva, J., Tenreyro, S., 2006. The log of gravity. The Review of Economics and Statistics 88 (4), 641–658.
Santos Silva, J., Tenreyro, S., 2010. Currency unions in prospect and retrospect. Annual Review of Economics 2, 51–74.
Santos Silva, J., Tenreyro, S., 2011. Further simulation evidence on the performance of the Poisson-PML estimator. Economics Letters 112 (2), 220–222.
Savage, I.R., Deutsch, K.W., 1960. A statistical model of the gross analysis of transaction flows. Econometrica 28 (3), 551–572.
Simonovska, I., Waugh, M.E., 2011. The Elasticity of Trade: Estimates and Evidence. Working Paper 16796, NBER, February.
Tinbergen, J., 1962. Shaping the World Economy: Suggestions for an International Economic Policy. Twentieth Century Fund, New-York.
Trefler, D., December 1995. The case of the missing trade and other mysteries. The American Economic Review 85 (5), 1029–1046.
Wei, S.-J., April 1996. Intra-National versus International Trade: How Stubborn are Nations in Global Integration? Working Paper 5531, NBER.
Wooldridge, J., 2010. Econometric Analysis of Cross Section and Panel Data, second ed. The MIT press.