Chapter 4
The Gravity Model in International Trade

Luca De Benedictis and Daria Taglioni

Abstract Since Jan Tinberben's original formulation (Tinbergen 1962, *Shaping the World Economy*, The Twentieth Century Fund, New York), the empirical analysis of bilateral trade flows through the estimation of a gravity equation has gone a long way. It has acquired a solid reputation of good fitting; it gained respected micro foundations that allowed it to move to a mature stage in which the "turn-over" gravity equation has been replaced by a gravity model; and it has dominated the literature on trade policy evaluation. In this chapter we show how some of the issues raised by Tinbergen have been the step stones of a 50-year long research agenda, and how the numerous empirical and theoretical contributions that followed dealt with old problems and highlighted new ones. Some future promising research issues are finally indicated.

4.1 Introduction

When in 1962 Jan Tinbergen, the future winner of the first 1969 Alfred Nobel Memorial Prize for economics, was sketching the empirical analysis for a report financed by a New York-based philanthropic foundation, his mind was back at his college years. In 1929, he had received his PhD in physics from Leiden University, the Netherlands, with a thesis entitled *Minimum Problems in Physics and Economics* under the supervision of Paul Ehrenfest, a close friend of Albert Einstein's (Szenberg 1992, p. 276; Leen 2004). Theoretical physics was his bread and butter,

L. De Benedictis
Department of Economic and Financial Institutions, University of Macerata, Via Crescimbeni 20, 62100 Macerata, Italy
e-mail: debene@unimc.it

D. Taglioni
World Bank, Washington DC, USA and Centre for Economic and Trade Integration at The Graduate Institute of International and Development Studies, Rue de Lausanne 132, 1211 Geneva, Switzerland
e-mail: dtaglioni@worldbank.org

L. De Benedictis and L. Salvatici (eds.), *The Trade Impact of European Union Preferential Policies*, DOI 10.1007/978-3-642-16564-1_4,

before the concern for the causes of poverty of the local working class pressed him to switch to economics. Therefore, it must not come as a surprise that, when he had to propose to the team of fellow colleagues of the Netherlands Economic Institute an econometric exercise "to determine the normal or standard pattern of international trade that would prevail in the absence of trade impediments," he came out with the idea of an econometric model formulated along the lines of Newton's law of universal gravitation[1], where trade flows are directly related to the economic size of the countries involved, and inversely related to the distance between them.

All simple and successful ideas have a life of their own, and their paternity can be attributed to multiple individuals. Before Tinbergen, Ravenstein (1885) and Zipf (1946) used gravity concepts to model migration flows. Independently from Tinbergen, Pöyhönen (1963), inspired by Leo Tornqvist,[2] published a paper using a similar approach.[3] Tinbergen's student and team-member of the Netherlands Economic Institute, Hans Linnemann, published a follow-up study (Linnemann 1966) which extended the analysis and discussed the theoretical basis of the gravity equation using the Walrasian model as a benchmark.[4] By the 1970s the gravity equation was already a must. The famous international trade book by Edward Leamer and Robert Stern included almost an entire chapter on it (Leamer and Stern 1970, pp. 157–170), based on the contribution of Savage and Deutsch (1960). Leamer and Stern's book introduced trade economists to the term resistance, that entered their glossary as a synonym for distance and other trade impediments. To make a long story short, from the first conceptualisation of Tinbergen (1962) the gravity equation has been used time and again to empirically analyse trade between countries. It has been defined as the workhorse of international trade and has been considered as a "fact of life" in this field of research (Deardorff 1998). The gravity equation's ability to correctly approximate bilateral trade flows makes it one of the most stable empirical relationships in economics (Leamer and Levinsohn 1995).

In Tinbergen's version of the gravity equation, X_{ij}, the size of the trade flow between any pair of countries is stochastically *determined*[5] by: (1) M_i, the amount

[1]The description of the econometric analysis was included in Appendix IV to the *Shaping the World Economy* report (Tinbergen 1962, pp. 262–293). Tinbergen himself described the summary of the results in Chapter 3 of the same report (Tinbergen 1962, pp. 59–66).

[2]Leo Tornqvist, was a famous Finnish statistician teaching at the University of Helsinki and father of the Tornqvist Price Index.

[3]Describing the exchange of goods between countries in matrix form, Pöyhönen (1963) makes it evident how international trade flows also depend on *internal trade*, a point also briefly covered by Tinbergen in the main text of his book (Tinbergen 1962, pp. 60–61).

[4]Linnemann quotes Zipf's work (Zipf 1946) and referring to Isard and Peck (see the impressive figure 1 on page 101 of Isard and Peck (1954)) surprisingly states that "Some authors emphasize the analogy with the gravitation law in physics . . . we fail to see any justification for this." He was not prophetic, but he was basing this statement on the fact that the elasticity of trade flows to distance were never found equal to 2.

[5]All words and phrases in *italics* are Tinbergen's. We will use them as milestones in our selective grand tour of the gravity model in international trade. This does not mean that all the main issues in this field of research were already pointed out by the author of the first path breaking contribution.

of exports a country i is able to supply to country j, depending on its economic size measured in terms of GNP *converted* in US dollars; (2) M_j, the size of the importing market, measured by its GNP, also converted in US dollars; (3) ϕ_{ij}, the geographical distance between the two countries in 1,000 nautical miles, as *a rough measure* of transportation costs or an index of information about export markets. The *model* was expressed in a *log–log* form, so that the elasticity of the trade flow was a constant (a_1, a_2, and a_3) with respect to the three explanatory variables. Actually, trade flows were *measured* both in terms of exports and imports of *commodities* and only *non-zero trade flows* were included in the analysis.[6] Results turn out to be not much different using exports or imports. *Adjacent countries* were assumed to have a more intense trade than what distance alone would predict; the adjacency was indicated by the dummy variable N_{ij}, that took the value 1 if the two countries were sharing a common land border. Finally, the equation was augmented with political or *semi-economic* factors: a dummy variable V_{ij} indicated that goods traded received a preferential treatment in the importing country if they belonged to the British Commonwealth system of preferences.[7] As customary, a gravitational constant G and a i.i.d. stochastic term ε_{ij} were also included. In equation-form:

$$\ln X_{ij} = \underbrace{\ln G}_{a_0 \equiv \text{constant}} + \underbrace{a_1 \ln M_i + a_2 \ln M_j}_{\text{economic attractors}} + \underbrace{a_3 \phi_{ij} + a_4 N_{ij}}_{\text{distance}} + \underbrace{a_5 V_{ij}}_{\text{policy}} + \underbrace{\varepsilon_{ij}}_{\text{iid}} \tag{4.1}$$

Elasticities were estimated by means of an Ordinary Least Squares (OLS) *cross-country regression* on 1,958 trade flows data for 18 countries, as a first trial, and for 42 countries, as a robustness check.[8]

The relationship between trade and the dummy policy variable V_{ij} can be seen in a simple graphical illustration of this relationship, conditional on distance, as in Fig. 4.1. The linear prediction for trade flows reported in the chart is obtained by replicating Tinbergen's first exercise with data on trade, Free Trade Agreements

However, many open questions were already intriguing researchers fifty years ago. A surprising persistence that we think is worth pointing out.

[6]For an early discussion of the zero trade flows see Linnemann (1966, p. 64).

[7]A dummy variable was also included for Benelux and, in a larger subsample to a broad variable identifying preferential agreements. The strategy of considering the effect of Preferential Trade Agreements (PTA) through the use of dummy variable has been prominent in the literature. Only recently the alternative strategy of explicitly including the preferential margin guaranteed by the agreement has been taken into account (see Chapter 3). We will come back to this issue in Section 4.4.2.4.

[8]The countries included in the first exercise were mainly developed countries: Brazil, Venezuela, South Africa, Japan, Canada, USA, Austria, Belgium-Luxembourg, Denmark, France, Germany (FR), Italy, Netherlands, Norway, Sweden, Switzerland, UK, and Australia. For a complete list of the 42 countries included in the second exercise see Tinbergen (1962, p. 274). The Benelux preference (between Belgium, Luxembourg and the Netherlands) was also represented by a dummy variable.

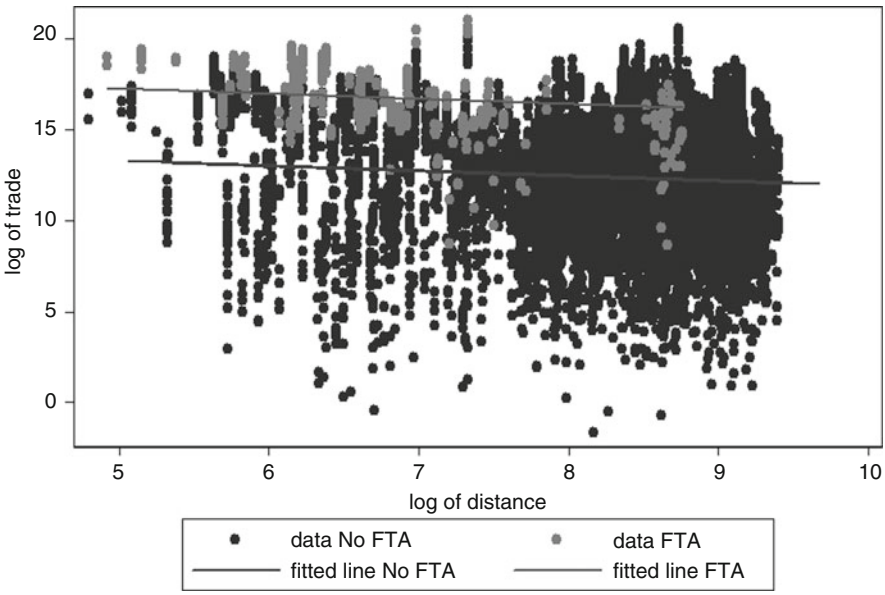

Fig. 4.1 Distance and trade preferences

(FTA) and distance, from Subramanian and Wei (2007). It reproduces the negative marginal effect of distance, conditional on the preferential treatment granted by FTA.[9] The positive effect of trade preferences is visible in Fig. 4.1 as the vertical distance between the two parallel regression lines. Things get less clear cut when we also include other covariates in the regression.

In the original estimation by Tinbergen (1962), the coefficients of GNP and distance had what became "the expected sign" in all subsequent analyses – the coefficients of the economic attractors were positive and the one of distance was negative – and resulted relevant and significant.[10] Moreover, the fit of the estimation was found to increase when the data sample was increased from 18 to

[9]The resulting estimation reproduces fairly well Tinbergen's original one. We did not have data on Benelux and also the trade data for South Africa was largely missing. We used data for 1960 and replaced GNP with GDP.

[10]In his comments to the regression's functional form, Tinbergen explained that in his view the economic size (GNP) of the importing country played a twofold role: it indicates its demand – external and internal – and its degree of diversity of production. In principle, the sign of the coefficient could have been positive (demand) or negative (self-sufficiency). For Tinbergen it was a surprise that the coefficient was positive. It was also a surprise to observe that countries "trading less than normally" (below the regression line) were the bigger and the richer countries. Though the second evidence – small countries trade more with the rest of the world – has been explored theoretically (Anderson and Yotov 2010) and empirically (Alesina et al. 2005; Rose 2006), the role played by self-sufficiency has been largely neglected by the literature.

42 countries; on the other hand, the coefficient for adjacency was never significant and the one for trade preference was borderline. Although its functioning wasn't perfect, Tinbergen, who was a *correlation hunter* (Szenberg 1992, p. 278), succeeded in identifying a specification whose key variables explained a very high percentage of variability in the data, with a multiple correlation coefficient, R^2, of 0.82. This result led the way to the application of the log-linearized version of Newton's universal law of gravity to social and economic activities. Since then, the equation was viewed as a big success in enlightening "... the dominant role played by ... exporters' and importers' GNP and distance in explaining trade flows" (Tinbergen 1962, p. 266).

The specification however, left room for improvement, and the positive but relatively small role of trade preferences was an issue that stimulated further inquiry. In this chapter we will address one at the time some of the main open issues – associated to Tinbergen's original wording, that we have highlighted by marking the text in *italics*. We will review, briefly, the theoretical and, more extensively, the empirical trade literature on the gravity equation and we will indicate some of the promising avenues for future research.

4.2 Estimating Gravity

Let's start from the first term highlighted in the introduction: *determined*. Bilateral trade flows are determined by the variables included in the right-hand-side of the gravity equation. This implies a clear direction of causality that runs from income and distance to trade. This direction of causality is nowdays largely theory-driven and based on the assumption that the gravity equation is derived from an micro-economic model where income and tastes for differentiated products are given. Empirically, the causality (as if in a randomized quasi-experimental setting[11] *à la* Rubin) of the gravity equation, as described in (4.1), is more difficult to establish: the equation as it stands represents a regression of endogenous variables on endogenous variables. As a consequence, the parameter of the gravitational constant G is not constant: it varies by trade partner and over time and is correlated with many, if not all, policy variables affecting trade (which are rarely considered as the equivalent of a treatment in a random trial experimental setting). Failure to acknowledge this leads to an estimated impact of the policy variables likely to be biased and often severely so.

[11]In this setting researchers are interested in the causal effect of a treatment that takes the form of binary trade policy intervention (when the treatment is a dummy variable) or an ordered or continuous trade policy intervention (when considering trade preferential margins). Units, in this case countries or specific sectors of a country, are either exposed or not exposed to the treatment. Even if the effect of the treatment can be potentially heterogeneous across units, usually researchers focus on the identification of an average treatment effect (see Angrist and Pischke 2008 for a discussion of quasi-experimental settings).

We are in the realm of omitted variable biases. To simplify, let's assume away GDP and distance and focus on the policy variables. The estimated gravity equation will have the following structure:

$$\ln X_{ij} = \underbrace{\ln G}_{a_0 \equiv \text{constant}} + \underbrace{a_5 V_{ij}}_{\text{policy}} + \underbrace{\varepsilon_{ij}}_{\text{iid}} \tag{4.2}$$

while the true structural model is:

$$\ln X_{ij} = \underbrace{\ln G}_{a_0 \equiv \text{constant}} + \underbrace{a_5 V_{ij}}_{\text{policy}} + \underbrace{a_6 \ln Z_{ij}}_{\text{omitted varible}} + \underbrace{\varepsilon_{ij}}_{\text{iid}} \tag{4.3}$$

We can write Z_{ij} as a function of V_{ij} in an auxiliary regression:

$$\ln Z_{ij} = \underbrace{b_0}_{\text{constant}} + \underbrace{b_1 V_{ij}}_{\text{policy}} + \underbrace{u_{ij}}_{\text{iid}} \tag{4.4}$$

Without being aware of it, we have estimated the following equation:

$$\ln X_{ij} = \underbrace{(a_0 + b_0 a_6)}_{\text{constant}} + \underbrace{(a_5 + a_6 b_1) V_{ij}}_{\text{policy}} + \underbrace{(\varepsilon_{ij} + a_6 u_{ij})}_{\text{iid}} \tag{4.5}$$

Therefore, unless $b_1 = 0$, $\mathrm{E}(\hat{a}_5) = a_5 + \underbrace{a_6 \left[\frac{\sum V_{ij} Z_{ij}}{\sum V_{ij}^2} \right]}_{\text{bias}}$. Accordingly, the bias depends on the correlation between the policy variable and the omitted variable, and can have a positive or negative sign. Furthermore, the mis-specification also affects the standard errors, which would result in a positive bias (Wooldridge 2002, Chapter 4).

The omitted variable problem in the gravity equation has been dealt through different approaches. The first has been to include in the equation one or more proxy variables correlated with the omitted variable. We will discuss this strategy in the context of the effect of distance on trade. A second approach has been to include a time-dimension in the analysis and to move from cross-country analysis to panel data analysis, since one of the most likely sources of omitted variables is country heterogeneity, an issue that is not likely or easy to account for in a cross-country setting. While we will tackle the aspects related to a correct specification in the following example, where we show that the biases from mispecification are non-trivial, here we would like to focus on the choice between cross-section and panel estimations. Even though elements such as distance and size are best captured by cross sections with the panel not adding much content in short horizons, in most cases panel specifications should be preferred to cross-section specifications

because of the inability of the latter to properly account for the omitted variables bias. On the other hand, policy effects, such as the trade promotion of free trade agreements or custom unions, are always better identified in panels, through the time series dimension. Indeed, in the cross-section specification they are highly collinear with distance.

With these issues in mind, in the next two sections we first empirically show the potential biases from a bad specification and then provide a synthetic discussion of how to specify a theoretically sound gravity equation. The aim is to give the reader an informed perspective of what theory-based specifications can be applied to address the various empirical questions posed to the gravity equation.

4.2.1 How Big Are the Biases?

In order to show how big are the biases from mis-specifying the gravity equation, we re-run Tinbergen's regression as a benchmark, for the same subset of 42 countries and for data taken at intervals of five years, from 1960 to 2005. We will show that the trade policy variable coefficient is very sensitive to the specification. In particular, we show the effect of introducing different types of fixed effect controls and of using real-vs-nominal GDP.[12] Results are reported in Table 4.1 below.

Columns (1) and (2) report the base regression as in Tinbergen (1962), with only two differences. First, instead of GNP we use GDP (in column (1) real GDP and in column (2) nominal GDP). Second, our policy variable of interest is whether a country pair is in an FTA relationship. Columns (3) and (4) reports results where time dummies are added to the regression, to account for the changing nature of the relationship over time, with the difference between column (3) and (4) being the real-vs-nominal GDP choice. Column (5) and (6) report results with time invariant importer and exporter fixed effects on top of the time dummies. Column (7) shows results for time varying exporter and importer fixed effects. Lastly, column (8) presents a specification where time invariant pair effects are also added.

In spite of Fig. 4.1, the baseline Tinbergen-like specification seems to suggest that being in an FTA does not have any statistically significant effect on trade if we use real GDP, but a positive and statistically robust effect if we use nominal GDP (columns 1 and 2). Similarly, adjacency (i.e. sharing a border) does not seem to be trade-enhancing when we use real GDP figures, and positive and significant when we use nominal GDP figures. All other variables have the expected sign and are statistically significant, with both GDP specifications. Adding time fixed effects (columns 3 and 4) and time-invariant importer and exporter fixed effects (columns 5 and 6) however has the surprising effect of reversing the sign of the FTA

[12]The use of nominal GDP (instead of real GDP) is theoretically more sound. We will come back to this issue in Section 4.5.3.

Table 4.1 The gravity equation with different fixed effects

Variables	(1) *ln imports*	(2) *ln imports*	(3) *ln imports*	(4) *ln imports*	(5) *ln imports*	(6) *ln imports*	(7) *ln imports*	(8) *ln imports*
ln $rgdp_i$	1.027***		1.077***		1.449***			
	(0.01)		(0.01)		(0.08)			
ln $rgdp_j$	1.115***		1.169***		1.375***			
	(0.01)		(0.01)		(0.08)			
ln gdp_i		0.704***		0.957***		0.803***		
		(0.01)		(0.01)		(0.05)		
ln gdp_j		0.777***		1.051***		1.003***		
		(0.01)		(0.01)		(0.06)		
ln distance	−1.227***	−0.897***	−1.225***	−1.014***	−1.045***	−1.067***	−1.089***	
	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	
sharing a border dummy	0.0780	0.406***	0.115	0.537***	0.608***	0.587***	0.564***	
	(0.09)	(0.10)	(0.09)	(0.08)	(0.09)	(0.09)	(0.09)	
FTA dummy	−0.0490	0.146**	0.214***	−0.172***	−0.396***	−0.632***	−0.717***	0.266***
	(0.06)	(0.07)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)	(0.06)
Constant	−17.61***	−16.64***	−18.37***	−25.54***	−32.26***	−19.85***	28.56***	13.02***
	(0.346)	(0.384)	(0.335)	(0.355)	(2.536)	(2.241)	(0.319)	(0.377)
Observations	10,781	10,831	10,781	10,831	10,781	10,831	10,831	10,831
R-squared	0.656	0.550	0.692	0.732	0.785	0.785	0.815	0.928
Time effects	No	No	Yes	Yes	Yes	Yes	No	No
Exporter and importer time invariant fe	No	No	No	No	Yes	Yes	No	No
Exporter and importer time-varying fe	No	No	No	No	No	No	Yes	Yes
Time invariant pair fe	No	No	No	No	No	No	No	Yes

Note: robust standard errors in parentheses. (*) significant at 10% level; (**) significant at 5% level; (***) significant at 1% level. fe stands for fixed effect(s).

coefficient in three out of the four cases. Notwithstanding the sign of the coefficient, the fact that the FTA coefficient acquires statistical significance and that its point estimates increase with the inclusion of time dummies, suggests the existence of a significant time trend non-orthogonal to the FTA dummy. Interestingly, while the FTA coefficient is negative, the coefficient for sharing a border is positive and strongly significant. The two results in combination lead us to formulate the hypothesis that the two variables might be correlated with each other. If this the case, entering the exporter and importer fixed effects in a time-varying way does not help achieving a sound specification.[13] Hence the only solution remains changing slightly the focus of our research question, by asking, what is the effect of entering in a FTA relationship for bilateral trade? With this different angle, we can formulate a gravity specification where we add time invariant pair effects on top of time-varying importer and exporter fixed effects to address pair-specific invariant omitted variables. The outcome is an FTA coefficient positively signed and statistically significant. The coefficient is now to be interpreted as the effect of entering in an FTA instead of being part of it, i.e. with this specification a country-pair that was part of a bilateral agreement throughout the period of observation would not be picked up by the FTA dummy.[14]

Given the evidence of how important it is to properly specify the gravity equation to account for country heterogeneity, we now turn to provide the reader with an informed perspective on the empirical issues associated with the estimation of the gravity equation. We do this by discussing how to achieve theoretically sound gravity specifications. In other words, abandoning for a while Tinbergen's wording, we link the gravity equation to the gravity model.

4.3 Theory-Based Specifications for the Gravity Model

For Tinbergen (1962, p. 263) the gravity *equation* was a "turnover relation," where no separate demand and supply were considered, no prices were specified, and no dynamics was taken into account. This doesn't mean that there was no model under the equation. The exporter's and importer's GNP captured, respectively the effect of production capacity and of demand and distance was a measure of the trade

[13]Another source of bias in the regression could come from self-selection, i.e. nations that choose to be in a given trade policy regime are not randomly chosen. Geographical proximity, common language, common border, former colonial status, size and wealth of a nation are likely to strongly influence the decision to enter or not in given policy regimes. This causes a selection problem. Matching methods have been used to control for self-selection (see Persson (2001) for an early application and Millimet and Tchernis (2009) for a discussion of the methodology). However, solving for self-selection needs to be done on a case by case basis.

[14]Fixed effects specifications require getting rid of RHS variables that are accounted for by the fixed effects. This explains why we have no entries for GDP, distance and border in columns (5) to (8).

feasibility set. Assumptions were not spelled out and restrictions were not explicitly imposed, but a model was already *in nuce*. Surprisingly, all developments up to the early 1980s concerned the empirics of the relationship, while the theoretical basis remained underdeveloped.[15] Since then, things have changed radically. Three decades of theoretical work has shown that the gravity equation can be derived from many different – and sometimes competing – trade frameworks. In 1979, James Anderson proposed a theoretical explanation of the gravity equation based on a demand function with Constant Elasticity of Substitution (CES) *à la* Armington (1969), where each country produces and sells goods on the international market that are differentiated from those produced in every other country. Later work has included the Armington structure of consumer preferences in (1) monopolistic competition frameworks (Krugman 1980; Bergstrand 1985, 1989; Helpman and Krugman 1985), (2) models *à la* Heckscher-Ohlin (Deardorff 1998), or (3) models *à la* Ricardo (Eaton and Kortum 2002). The catalyst of the more recent wave of theoretical contributions on gravity is the literature on models of international trade with firm heterogeneity, spearheaded by Bernard et al. (2003) and Melitz (2003).

Given the plethora of models available, the emphasis is now on ensuring that any empirical test of the gravity equation is very well defined on theoretical grounds and that it can be linked to one of the available theoretical frameworks. Accordingly, the recent methodological contributions brought to the fore the importance of defining carefully the structural form of the gravity equation and the implications of mis-specifying (4.1). In this context, two broad sets of key issues have been identified. A first important range of contributions is related to the multilateral dimension of the gravity model. Anderson and van Wincoop (2003) – building on Anderson (1979) – showed that the flow of bilateral trade is influenced by both the trade obstacles that exist at the bilateral level (Bilateral Resistance) and by the relative weight of these obstacles with respect to all other countries (what they called the Multilateral Resistance). After this contribution, the omission of a Multilateral Resistance term is considered a serious source of bias and an important issue every researcher should deal with in estimating a gravity equation. The second main area of methodological concern is related to the selection bias associated to the presence of heterogeneous firms operating internationally. Contrary to what is implied by models of monopolistic competition *à la* Krugman, not all existing firms operate on international markets. In fact, only a minority of firms serves foreign markets (Mayer and Ottaviano 2008; Bernard et al. 2007). Moreover, not all exporting firms export to all foreign markets as they are generally active only in a subset of countries.[16] The critical implication of firm heterogeneity for modeling the gravity

[15]Alan Deardoff refers to the gravity model as having "somewhat dubious theoretical heritage" (Deardorff 1998, p. 503). Similar assessments can be found in Evenett and Keller (2002) and Harrigan (2001).

[16]The heterogeneity in firm behavior is due to fixed costs of entry which are market specific and higher for international markets than for the domestic market. Hence, only the most productive firms are able to cover them. Firm productivity is furthermore correlated with a large array of other

equation is that the matrix of bilateral trade flows is not full: many cells have a zero entry. This is the case at the aggregate level and the more often this case is seen, the greater the level of data disaggregation. The existence of trade flows which have a bilateral value equal to zero is full of implications for the gravity equation because it may signal a selection problem. If the zero entries are the result of the firm choice of not selling specific goods to specific markets (or its inability to do so), the standard OLS estimation of the gravity equation would be inappropriate: it would deliver biased results (Chaney 2008; Helpman et al. 2008).

Irrelevant of the theoretical framework of reference, most of the modern mainstream foundations of the gravity equation are variants of the demand-driven model described in the appendix of Anderson (1979). Hence, in the following paragraphs, we summarise the key theoretical points of this common framework. We will mainly rely on the Anderson and van Wincoop (2003) and Baldwin and Taglioni (2006) derivations, using standard notation to facilitate the exposition. We will obviously mention where and in what way the supply-driven models *à la* Eaton and Kortum (2002) differ.

The starting point of Anderson and van Wincoop (2003) is a CES demand structure, with the assumption that each firm produces a unique variety of a unique good. Since trade data are collected in value terms it is convenient to work with the CES expenditure function rather than the CES demand function. The solution to the utility maximisation problem tells us that spending on an imported good that is produced in nation i and consumed in nation j is:

$$x_{ij} \equiv \left(\frac{p_{ij}}{P_j}\right)^{1-\sigma} M_j \quad \text{where} \quad \sigma > 1 \tag{4.6}$$

where x_{ij} is the expenditure in destination country j on a variety made in country i, P_j is nations-j's CES price index, σ is the elasticity of substitution among varieties assumed greater than one, and M_j is nation-j expenditure, and p_{ij} is the consumer price in nation j of goods produced in nation i

$$p_{ij} = \mu_{ij} p_i \phi_{ij} \tag{4.7}$$

In this equality, p_i is nation i's domestic price, μ_{ij} is the bilateral price mark-up (which depends on the assumed market structure) and ϕ_{ij} is the bilateral "trade costs," which is one plus the *ad valorem* tariff equivalent of all natural and manmade barriers, i.e. whatever cost-factor that introduces a wedge between domestic and foreign goods' prices, conditional on market structure. This is the pass-through equation. Combining this with (4.6) gives us the per-variety

observable firm characteristics. Hence firms that serve both domestic and foreign markets are not only more productive but also larger, more innovative and more intensive in human and physical capital. By contrast firms that only serve the domestic market are less productive, smaller, less innovative, and labor intensive.

relationship. Aggregating over all varieties exported from country i to country j (assuming that all varieties produced in nation i are symmetric) yields aggregate bilateral trade:

$$X_{ij} = \sum_i x_{ij} = n_{ij}\left(\mu_{ij} p_i \phi_{ij}\right)^{1-\sigma} \frac{M_j}{P_j^{1-\sigma}} \tag{4.8}$$

where X_{ij} indicates the value of the aggregate trade flows (measured in terms of the numeraire), and n_{ij} indicates the number of nation-i varieties sold in nation-j.[17]

Let us stress the point that our derivation of the gravity equation is based on an expenditure function. This explains two key factors. First, destination country's GDP enters the gravity equation (as M_j) since it captures the standard income effect in an expenditure function. Second, bilateral distance enters the gravity equation since it proxies for bilateral trade costs which get passed through to consumer prices and thus dampens bilateral trade, other things being equal. The most important insight from the above mathematical derivation is that the expenditure function depends on relative and not absolute prices. This allows factoring in firms' competition in market j via the price index P_j. Hence, (4.8) tells us that the omission of the importing nation's price index P_j from the original gravity equation described in (4.1) leads to a mis-specification. It should further be noted that the exclusion of dynamic considerations is problematic: Although we omitted time suffixes for the sake of simplicity, the reader should be aware that P_j is a time-variant variable, so it will not be properly controlled for if one uses time-invariant controls, unless the researcher is estimating cross-sectional data.

Having shown why destination-country GDP and bilateral distance enter the gravity equation, we turn next to explaining why the exporter's GDP should also be included. The explanation is Tinbergen's: it reflects the export capacity or the supply available on the side of the exporter. While the way it enters the equation is the same across theoretical frameworks, the interpretation of the role it plays depends on the specificities of the underlying theory. The Anderson-van Wincoop derivation is based on the Armington assumption of competitive trade in goods differentiated by country of origin. In other words, each country makes only one product, so all the adjustment takes place at the *price* level. This implies that nations with large GDPs export more of their product to all destinations, since their good is relatively cheap. This equates to saying that their good must be relatively cheap if they want to sell all the output produced under full employment. Helpman and Krugman (1985) make assumptions that prevent prices from adjusting (frictionless trade and factor price equalisation), so all the adjustment happens in the number of varieties that each nation has to offer. This implies that nations with large GDP export more to all destinations, since they produce many varieties. Since each firm produces one variety and each variety is produced only by one firm, stating that the

[17] Anderson and van Wincoop (2003, p. 174) assume that this number is equal to 1 for all origin and destination markets.

adjustment takes place at the level of varieties equates to stating that the number of firms in each country adjust endogenously. This is enough to lead to the standard gravity results.

Turning back to Anderson and van Wincoop and how the exporter's GDP should enter the gravity equation, the idea is that nations with big GDPs must have low relative prices so to sell all their production (market clearing condition). To determine the price p_i that will clear the market, we sum up nation i's sales over all markets, *including its own market*, as Tinbergen originally pointed out (Tinbergen 1962, pp. 60–61) and set it equal to overall production. This can be written as follows: $M_i = \sum_j n_{ij} x_{ij}$ which equates to

$$M_i = p_i^{1-\sigma} \sum_j n_{ij} \left[\left(\mu_{ij} \phi_{ij} \right)^{1-\sigma} \frac{M_j}{P_j^{1-\sigma}} \right], \tag{4.9}$$

where the second equality follows from the substitution of the expression for x_{ij}, that is produced in turn by the substitution of (4.7) into (4.6). Solving (4.9) for $p_i^{1-\sigma}$ yields:

$$p_i^{1-\sigma} = \frac{M_i}{\Omega_i}$$

$$\Omega_i = \sum_j \left[n_{ij} \left(\mu_{ij} \phi_{ij} \right)^{1-\sigma} \frac{M_j}{P_j^{1-\sigma}} \right] \tag{4.10}$$

where Ω_i represents the average of all importers' market demand – weighted by trade costs. It has been named in many different ways in the literature, including market potential (Head and Mayer 2004; Helpman et al. 2008), market openness (Anderson and van Wincoop 2003) or remoteness (Baier and Bergstrand 2009).

Using (4.10) in (4.8) yields a basic but correctly specified gravity equation

$$X_{ij} = n_{ij} \left(\mu_{ij} \phi_{ij} \right)^{1-\sigma} \frac{M_j}{P_j^{1-\sigma}} \frac{M_i}{\Omega_i} \tag{4.11}$$

If we suppose that each country only produces one product, as in Anderson and Van Wincoop (2003), i.e. n_{ij} (=1), and assume that the markup μ_{ij} depends upon the distance between the two trading partners, we arrive to the most familiar specification of the gravity equation:

$$X_{ij} = \phi_{ij}^{1-\sigma} \frac{M_j}{P_j^{1-\sigma}} \frac{M_i}{\Omega_i} \tag{4.12}$$

Hence, we just showed that origin country's GDP enters the gravity equation since large economies offer goods that are either relatively competitive or abundant in variety, or both. The derivation also shows that the exporting nation's market

potential Ω_i matters, and that the misspecification in the gravity equation would be more serious the bigger the asymmetry among countries.

Equation (4.12) is identical to (4.9) in Anderson and van Wincoop (2003, p. 175). But it is not identical to their final expression. As shown by Baldwin and Taglioni (2006), Anderson and van Wincoop (2003) assume that $\Omega_i = P_i^{1-\sigma}$ for all nations, since it is a solution to the system of equation that defines these two terms. There are three critical assumptions behind this. First, they assume that trade costs are two-way symmetric across all pairs of countries. This assumption however is automatically violated in the case of preferential trade agreements. Second, they assume that trade is balanced, i.e. $X_{ij} = X_{ji}$, also an hypothesis that is often violated in practice. Finally, they assume that there is only one period of data. Were the above three conditions verified, we could refer to the product of the two terms Ω_i and $P_i^{1-\sigma}$ as to a single country geography index, with the term of *multilateral resistance*; which can be empirically controlled for by a time-invariant country-fixed effect.[18] In fact, a more general case is that Ω_i and $P_i^{1-\sigma}$ are proportional, i.e. that $\alpha\Omega_i = P_i^{1-\sigma}$ and that there is a different α per year. If this point is acknowledged, it is simple to see that the gravity model in (4.1) is missing a time-varying dimension and that Ω_i and $P_i^{1-\sigma}$ must be accounted for with separate terms. An easy and practical solution to match the theory with the data is to introduce time-varying importer and exporter fixed effects. Obviously, in cross-sections, the Anderson van Wincoop specification is sufficient owing to the lack of time dimension. Often however, the need of correcting for omitted variables biases clashes with problems of collinearity with the other variables. Hierarchical Bayesian methods may be able to assist in reducing the resulting overparametrization problem (Guo 2009), but not in solving it. Alternatively, more sophisticated terms that account for Ω_i and $P_i^{1-\sigma}$ but that are orthogonal to the other variables in the equation must be computed, or strategies to control for potential collinearity have to be devised case-by-case.

A final aspect to consider is firm heterogeneity and the connected issue of zeroes in the trade matrix. In models with identical firms, in the absence of natural and man-made trade costs, countries either trade or they are in autarky. If they do trade, every firm in a country exports to every country in the world. Introducing firm heterogeneity in models of international trade however allows for a more realistic representation of reality, namely one where not all firms in a country export, not all products are exported to all destinations and not all countries in the rest of the world are necessarily served. Moreover, as trade barriers move around, the set of exporters will change, and this additional margin of adjustment – the extensive margin – will radically change the aggregate trade response to the underlying geographical and

[18]Obviously, some econometric fixes have been found. In particular, the practice introduced by Harrigan (2001) and popularized by Feenstra (2003), to control for Multilateral Resistance through the use of country fixed effects in the econometric estimation. Incidentally, the country fixed effect practice diverted the analysis from the causes of multilateral resistance to the effects of multilateral resistance. The latter remains a promising area of analysis, especially in the context of policy evaluation.

policy variables. Helpman et al. (2008), from a demand side, and Chaney (2008), from a supply side, have both introduced heterogeneity in gravity models, allowing for the more general derivation of gravity with heterogeneous firms.

Consider a world with many countries and same CES preferences across countries with elasticity of substitution $\sigma > 1$. Country i has a given number N_i of potential producers, i.e. entrants. These entrants draw their unit input requirement a from a distribution $G(a) = (a/\bar{a})^k$, where $k > \sigma - 1$ and $0 \leq a \leq \bar{a}$. The term k denotes the productivity distribution parameter that governs the entry and exit of firms into the export markets. Hence k indicates the degree of firm heterogeneity and σ the degree of differentiation across products. The same distribution $G(a)$ holds across countries, but the cost of the input bundle w_i is country-specific. Trade costs ϕ_{ij} for trade between countries i and j are composed of a variable and a fixed part. The variable component is $\tau_{ij} \geq 1$, a per-unit iceberg trade cost. The fixed component is $f_{ij} > 0$. These costs include also serving the domestic market where $i = j$ and where one can assume that $\tau_{ii} = 1$ and that f_{ij} includes overhead fixed costs.

If a producer in country i with unit cost a exports to j, it will set a price $p_{ij}(a)$ and generate export sales $x_{ij}(a)$ and export profits $\pi_{ij}(a)$:

$$p_{ij}(a) = \frac{\sigma}{\sigma - 1} w_i \tau_{ij} a \tag{4.13}$$

$$x_{ij}(a) = \frac{M_j}{P_j^{1-\sigma}} p_{ij}(a)^{1-\sigma} \tag{4.14}$$

$$\pi_{ij}(a) = \frac{1}{\sigma} x_{ij}(a) - w_i f_{ij} \tag{4.15}$$

As before M_j and $P_j^{1-\sigma}$ are expenditure and price index, respectively in importer country j. The cut-off for profitable exports from i to j which we define a_{ij} is determined by $\pi_{ij}(a_{ij}) = 0$. In other words, we assume that $\bar{a}$ is high enough to allow that $a \leq \bar{a}$ for every pair of countries i and j.

Given this, aggregate bilateral trade from i to j is then

$$X_{ij} = N_i \int_0^{a_{ij}} x_{ij}(a) dG(a) \tag{4.16}$$

If one defines $M_i = \sum_j X_{ij}$ as the value of country i's aggregate output, where trade with every country in the world including self is accounted for, then – after some algebraic transformations – the aggregate bilateral trade from i to j can be written as follows:

$$X_{ij} = {\tau_{ij}}^{-k} f_{ij}^{-\frac{k-\sigma+1}{\sigma-1}} \left(\frac{M_j}{P_j^{1-\sigma}} \right)^{\frac{k}{\sigma-1}} \frac{M_i}{\Omega_i}, \tag{4.17}$$

where $\Omega_i = {\tau_{ij}}^{-k} f_{ij}^{-\frac{k-\sigma+1}{\sigma-1}} \sum_j \left(\frac{M_j}{P_j^{1-\sigma}} \right)^{\frac{k}{\sigma-1}}$.

The gravity specification with firm-heterogeneity differs from previous specifications in two broad ways, which we summarise below. While some of the points we will make are already clear from (4.17), the interested reader is referred to Chaney (2008) which demonstrates explicitly each of the issues that we raise below. He does so by decomposing (4.17) by the two margins of trade, solving for each expression and expressing each margin in elasticities.

To start with, the per-unit trade costs are shown to affect both the intensive and the extensive margin of trade. However, they do so with some important differences. First, per-unit trade costs τ_{ij} are subject to firm heterogeneity (as indicated by the superscript k) and no longer to product differentiation (i.e. the parameter $1 - \sigma$ in 4.12). This is due to the fact that, with Pareto or Frechet distributed productivity shocks, the effect of σ on the intensive and extensive margin cancels out, so that in aggregate the elasticity of trade flows with respect to the per-unit trade costs only depends on k. Nevertheless, when per-unit trade costs move, both the intensive and the extensive margin of trade are affected and σ, the degree of competition in the market, plays an important role in the dynamics. The intensive margin of trade responds to changes in variable trade costs as in traditional specifications: i.e. the elasticity of incumbent exporters with respect to τ_{ij} is $(\sigma - 1)$, hence each firm faces a constant elasticity residual demand, and therefore when goods are very substitutable, the export of incumbents is very sensitive to trade costs. The extensive margin, on the other hand, behaves idiosyncratically. When per-unit trade costs move, some of the less productive firms start exporting, but their impact on aggregate flows is inversely proportional to σ. As goods become more substitutable (high σ), the market share of the least productive firms shrinks compared to the market share of the more productive firms and the change in trade costs has a decreasing impact on aggregate trade flows. Finally, fixed costs only matter for the extensive margin of trade, since those exporters that have already decided to enter a market are not going to change their decision. This effect is clearly visible with a first order approximation, as the derivative of trade flows to fixed costs posts zero elasticity for the intensive margin. A second important set of implications of firm-heterogeneity for gravity models arises because the importer CES market demand effect is amplified by the upshot of demand on the extensive margin of trade $k/(\sigma - 1) > 1$. By contrast, the exporter's market potential is computed as in previous models, given however differences in trade costs and the existence of importer fixed effects. Having shown how to handle firm heterogeneity in gravity models from a theoretical point of view,[19] in the following sections of the chapter we will now come back to Tinbergen's wording and discuss the empirical strategies that allow making use of the information contained in the trade model founding the gravity equation.

[19]From a practical point of view, it is not necessary to rely on firm-level data to consider the effect of firms heterogeneity. Given the productivity distribution of domestic firms, the aggregate volume of trade defines the volume of trade of the marginal exporting firm – the one with the productivity exactly equal to the cut-off point of the productivity distribution.

4.4 A Piecewise Analysis of the Gravity Equation

4.4.1 Dependent Variable

To put things in context, there are three issues associated with the left-hand side variable of the gravity equation. The first has to do with the issue of *conversion* of trade values denominated in domestic currencies and with the issue of deflating the time series of trade flows. The second is associated with the effect of the inclusion or exclusion of *zero-trade flows* from the estimation. Finally, the third issue is related with the *typology* of goods or economic activities to be included in the definition of trade flows: imports, exports, merchandise trade or any other possible candidate for a trade link between country *i* and country *j*. In the current section we will discuss the third and the first issues while leaving the problem of zero-trade flows for a more focused discussion in Section 4.5.1.

Starting with the issue of *typology*, in the large majority of studies the dependent variable is usually a measure of bilateral merchandise trade.[20] Three choices of trade flow measures are available to the researcher for the dependent variable of a classical gravity equation on goods trade: export flows, import flows or average bilateral trade flows. The choice of which measure to select should be driven first and foremost by theoretical considerations which mostly imply privileging the use of unidirectional import or export data. Sometimes however, considerations linked to data availability or differences in the reliability between exports and imports data may prevail. For example, a common fix to poor data is to average bilateral trade flows in order to improve point estimates. This is done because averaging flows takes care of three potential problems simultaneously: systematic under reporting of trade flows by some countries, outliers and missing observations. Although there are better ways of dealing with those problems,[21] it is common practice to justify the use of this procedure using the above arguments. This notwithstanding, caution should be applied in averaging bilateral trade. First of all, averaging is not possible in those cases where the direction of the flow is an

[20]Nevertheless, gravity models have also been employed for examining the determinants of trade in goods and services, other than merchandise. The gravity model offers a high probability of a good fit, but what we mentioned for trade in merchandise is also true for all other left-hand side variables: there is no reliable gravity equation without a supporting theoretical model. If one wants to explore a gravity model on Foreign Direct Investments (FDI), it is better to have a theory to refer to (as in Carr et al. 2001; or in Baltagi et al. 2007). The need for a theory is even more compelling if one wants to account for the many alternative strategies that heterogeneous firms have at their disposal to serve foreign markets, i.e. trade and FDI (and even differentiating further between offshoring or joint-ventures).

[21]It is true that reliability of the data varies significantly from country to country. But if this corresponds to a national characteristic that is considered to be constant along time, the country-specific quality of the data can be controlled for, as any other time-invariant country characteristic or country fixed effects.

important piece of information. Second, if carried out wrongly, averaging leads to mistakes.

Average bilateral trade is constructed by averaging the exports of country *i* to country *j* with the exports of country *j* to country *i*. Since each trade flow is observed as exports by the origin nation and imports by the destination country and most countries do both import and export from the same trade partner, typically four values are averaged to get the undirected bilateral trade that then needs to be log-linearised:[22]

$$T_{ij} = \mathrm{E}(x_{ij}, x_{ji}, m_{ij}, m_{ji}) \tag{4.18}$$

A bias may arise if researchers employ the log of the sum of bilateral trade as the left-hand side variable instead of the sum of the logs. Many published studies in the field of trade analysis, including some very recently published works, carry this bias. The mistake will create no bias if bilateral trade is balanced. However, if nations in the treatment group (i.e. the countries exposed to the policy treatment which average effect is being estimated) tend to have larger than usual bilateral imbalances – this is the case for trade between EU countries and also for North-South trade – then the misspecification leads to an upward bias of the treatment variable. The point is that the log of the sum (wrong procedure) overestimates the sum of the log (correct procedure). This leads to an overestimated treatment variable, as shown in Baldwin and Taglioni (2006). At any rate, the mistake implies that the researcher is working with overestimated trade flows within the sample.

Turning to *conversion*, the first item listed at the beginning of the section, trade should enter the estimation in nominal terms and it should be expressed in a common *numeraire*. This stems from the fact that the gravity equation is a modified expenditure equation. Hence, trade data should not be deflated by a price index. Deflating trade flows by price indices not only is wrong on theoretical grounds but it also leads to empirical complications and likely shortcomings, due to the scant availability of appropriate deflators. It is practically impossible to get good price indices for bilateral trade flows, even at an aggregate level. Therefore, approximations may become additional sources of spurious or biased estimation. For example, if there is a correlation between the inappropriate trade deflator and any of the right-hand side variables (the trade policy measures of interest), the coefficient will be biased, unless the measures are orthogonal to the deflators used.

[22]In constructing average trade, the researcher should make sure that the observations are statistically independent. Hence, if the two trade partners import and export from each other caution should be taken to cluster the four single observations in one single data point. We will come back to the issue of independence latter on.

As far as accounting conventions are concerned, trade data can be recorded either Free On Board (FOB) or gross, i.e. augmented with the Cost of Insurance and Freight (CIF).[23] Using CIF data may lead to simultaneous equation biases, as the dependent variable includes costs that are correlated with the right hand side variables for distance and other trade costs. If FOB data are not available, "mirror techniques," matching FOB values reported by exporting countries to CIF values reported by importing countries, can be used. These techniques however, remain to a large extent unsatisfactory due to large measurement errors (Hummels and Lugovskyy 2006). Hence, the suggestion as to this point is to be aware of whether CIF or FOB data are being used and interpret the results accordingly. If moreover the researcher is constructing a multi-country dataset, she should care for choosing data that are uniform, i.e. either all CIF or all FOB, controlling for measurement errors.

4.4.2 Covariates

As indicated above, a well specified gravity equation should include the "unconstant" terms Ω_i and $P_j^{1-\sigma}$. While several attempts at explicitly accounting for these terms have been made, including by means of structural assumptions on the underlying model and the use of non-linear methods of estimations, the practice has increasingly moved towards the use of simple-to-use fixed effects for these terms. As discussed earlier, however, fixed effects methods sometimes cannot be applied due to problems of overparametrization and correlation with the variable of interest.

[23]Most common sources of trade data include the following. International Monetary Fund (IMF) DOT statistics (http://www2.imfstatistics.org/DOT/) provides bilateral goods trade flows in US dollar values, at annual and monthly frequency. UN Comtrade (http://comtrade.un.org/) provides bilateral goods trade flows in US dollar value and quantity, at annual frequency and broken down by commodities according to various classifications (BEC, HS, SITC) and up to a relatively disaggregated level (up to 5 digit disaggregation). The CEPII offers two datasets CHELEM (http://www.cepii.fr/anglaisgraph/bdd/chelem.htm) and BACI (http://www.cepii.fr/anglaisgraph/bdd/baci.htm) which use UN Comtrade data but fill gaps. corrects for data incongruencies and CIF/FOB issues by means of mirror statistics. WITS by the World Bank provides joint access to UN Comtrade and data tariff lines collected by the WTO and ITC. The most timely annual, quarterly and monthly data are available from the WTO Statistics Portal. Similarly, the CPB provides data for a subset of world countries at the monthly, quarterly and annual frequency as indices. Series for values, volumes and prices are provided along with series for industrial production. Finally, regional or national datasets provide usually more detail. Notable examples are the US and EUROSTAT (EU27) bilateral trade data available in values and quantities up to the 10 digit and 8 digit level of disaggregation respectively. Australia, New Zealand and USA also collect consistent CIF and FOB values at disaggregate levels of bilateral trade. Interesting is also the case of China, It is interesting to note that China, besides providing SITC classifications also provides data series for processing trade.

4.4.2.1 Fixed Effects Specifications

The advantage of using fixed effect specifications lies in the fact that they represent by far the simplest solution to testing a gravity equation: they allow using OLS econometrics and do not require imposing ad-hoc structural assumptions on the underlying model. Specifications that make use of fixed effects are also very parsimonious in data needs: they only require data for the dependent variable and good bilateral values to estimate trade friction ϕ_{ij}.

Some caution however should be applied when using fixed effects on panel data. Importer and exporter fixed effects should be time-varying, as they capture time varying features of the exporter and importer, as discussed in the theory section above. Similarly, if data are disaggregated by industry, country-industry specific time-varying fixed effects should be applied. With very large panels, this may lead to computational issues. Whatever the solution the researcher devises, it is a necessary condition to control for the omitted time-varying terms Ω_i and $P_j^{1-\sigma}$ and to avoid large biases on the estimates of the other explanatory variables. Therefore, if computational complications arise, the researcher is recommended to find a way to solve the computational issues rather than giving up on properly specifying fixed effects. One final note of caution is in order: the use of exporter and importer fixed effects is suitable only if the variable of interest is dyadic, i.e. for ϕ_{ij}. If by contrast, the latter is exporter or importer specific, exporter and importer specific variables should be introduced explicitly and other means of avoiding the omitted variables bias (i.e. of controlling for Ω_i and $P_j^{1-\sigma}$) should be devised. Finally, pair (exporter–importer) fixed effects can also be used, if appropriate and if their introduction does not generate problems of collinearity with other explanatory variables.

4.4.2.2 Attractors

In line with the theoretical specification, attractors should reflect expenditure in the country of destination and supply in the country of origin. GDP, GNP and Population are all measures that have been used as proxies of the above terms. Per capita GDP (Frankel 1997) and measures for infrastructural development (Limao and Venables 2001) have also been used. Again, the appropriate measure should be selected on the basis of theoretical considerations.[24] As in the case of the dependent variable, these measures should enter in nominal terms. At any rate, deflating them would have no impact if one includes time fixed effects, which would swipe them away.

Many studies, the large part of them in a cross-sectional setting, augment the gravity equation with variables that could ease trade relations. Sharing a common

[24]This is true not only for variables to be included but also for restrictions on coefficients. From (4.14) the coefficient of M_i and M_j must be constrained to be one (this is why Anderson and van Wincoop (2003) estimated the gravity equation using $\frac{X_{ij}}{M_i \cdot M_j}$ as the left-hand side variable). With heterogeneous firm, as in (4.17), this is not required.

language, common historical events – such as colonial links, common military alliances or co-membership in a political entity – common institutions or legal systems, common religion, common ethnicity or nationality (through migration), similar tastes and technology, and input–output linkages enhance international trade. Many of those issues are of interest *per se* and are worth to be explored. An example in point is Head et al. (2010) who, while examining the effect over time of the independence of post-colonial trade between the colonized country and the former colonizer, conclude that trade flows are associated to some sort of relational capital that deteriorates with time if it is not renewed. They do so by showing that on average there is little short-run effect of the change in colony-colonizer relationship on trade: the reduction takes place progressively, over time, but trade does not stop suddenly, even in cases of hostile separation.

The researcher should be aware that most attractors have in general very low time variability. For this reason the researcher should pay particular caution in introducing them in fixed effects specifications. Should a specific attractor represent the core of the analysis, a safer option would be to avoid fixed effects estimations. This can be done by introducing measures of the exporter's market openness Ω_i and importer's CES price index, $P_j^{1-\sigma}$ along with the trade partners' GDPs. However, exporter's market openness and, even more so, importer's CES price index are difficult to construct. Once more, case-by-case solutions may be needed in controllig for the omitted variable bias.[25]

4.4.2.3 Trade Frictions

Distance matters! As Waldo Tobler's first law of geography states: "Everything is related to everything else, but near things are more related than distant things." The question is: why? As we emphasized in the introduction, Tinbergen's idea is that physical distance is *a rough measure* of transportation or information costs about foreign markets - already too many things for one single rough (and robust!) measure. Econometric estimates of the constant elasticity of trade to distance range within an interval of -0.7 and -1.2 (Disdier and Head 2008) and distance appears to be very persistent over time (Brun et al. 2005).

In the early years of the empirical analysis on bilateral trade flows, many researchers focused on producing better approximations for trade distance than simple Euclidean distance between the two poles of economic attraction of the two trade partners (respective capitals, main city in term of population or local production, main port or airport). To do so, some choose to estimate wedges between CIF

[25]It is difficult to give further details here, as the solutions should be devised case by case, based on the nature of the data at hand and on the research question. Nevertheless, options include introducing fixed effects at a different frequency than the attractors (Ruiz and Villarubia 2008; De Benedictis and Vicarelli 2009; Cardamone 2011) or to only look at entries and exits in a different (policy) regimes as we have done in Section 4.2.1.

and FOB data. Others used great-circle or orthodromic formulas.[26] Nowadays, all most common distance measures across virtually all country pairs in the world are freely available online[27] or can be obtained from the applets of the most important geo-representations available on the web. The issue is therefore not anymore how to calculate physical distance between two countries in the most appropriate way, but how to interpret the distance coefficient and if distance has a linear effect on trade.

Starting from the second issue, there is no reason to believe that distance should be related to trade in a linear manner. Trade costs are much dependent on the characteristics of specific goods, such as fragility, perishability, size or weight. In aggregate terms, trade cost would be country specific, depending on country's remoteness and sectoral specialization. In the absence of hard theoretical priors it is better to be agnostic and let the data speak. This is what has been done by Henderson and Millimet (2008). Using nonparametric techniques they found that the linearity assumption was supported by the data. We interpret this result as being clear evidence of variable trade costs being linear in distance for the average country. But what about fixed costs?

From the literature on heterogeneous firms and trade we know that fixed costs affect only the extensive margin of trade (Chaney 2008). Lawless (2010), extends the strategy proposed by Bernard et al. (2007), and decomposes the dependent variable of the gravity equation (export flows to each different foreign market) into the number of firms exporting (the extensive margin) and average export sales per firm (the intensive margin). Although, the proxy chosen for the intensive margin is not ideal in representing firm heterogeneity in exports, Lawless shows that distance has a negative effect on both margins, but the magnitude of the effect is considerably larger and significant for the extensive margin. Furthermore, the variables capturing the fixed cost (i.e. language, internal orography, infrastructure and import barriers) work through the extensive margin. Even Tinbergen (1962), in formulating the gravity equation as in (4.1) distinguished between variable costs (distance) and fixed costs. He approximated fixed costs by the cost-reducing effect of the *adjacency* dummy. We are therefore back to square one to the question of what lies behind the distance coefficient.

Let's tackle this issue from a very general point of view. In modern econometric terms, the concept of *distance as a rough measure of trade costs* (broadly defined as every cost that generates a conditional wedge between domestic and foreign prices) can be translated in the presence of a measurement error in the distance variable.

[26]The great-circle, or orthodromic, formula is the formula used for calculating the distance between longitude-latitude coordinates of the polar city of two countries is based on the spherical law of cosines is: $\phi_{ij} = a\cos\left(\sin(lat_i)\cdot\sin(lat_j) + \cos(lat_i)\cdot\cos(lat_2)\cdot\cos(long_j - long_i)\right)\cdot R$; where $R = 6{,}371$ is the radius of the earth, in km.

[27]CEPII generated a positive externality for all researchers by making freely available their measures of distance (see http://www.cepii.fr/anglaisgraph/bdd/distances.htm). Jon Haveman, Vernon Henderson and Andrew Rose were pioneers in this matter. Haveman's collection of International Trade Data and his "Useful Gravity Model Data" can be freely downloaded from, the FREIT. database http://www.freit.org/TradeResources/TradeData.html#Gravity.

It is well known that a measurement error in an explanatory variable, such as ϕ_{ij}, does result in a bias in the OLS estimates of a_3 in (4.1).

Following on that, we can write the measured value of ϕ_{ij} as the sum of the true unobserved value of the trade cost ϕ_{ij}^{*} plus a measurement error e_{ij} that is an i.i.d. normally distributed random variable:

$$\phi_{ij} = \underbrace{\phi_{ij}^{*}}_{\text{true unobserved measure}} + \underbrace{e_{ij}}_{\text{classical measurement error}}. \tag{4.19}$$

Consider now a simplified version of the gravity equation described in (1), where trade flows depend only on distance:

$$\ln X_{ij} = a_3\phi_{ij}^{*} + \varepsilon_{ij} = a_3\phi_{ij} + (\varepsilon_{ij} - a_3 e_{ij}). \tag{4.20}$$

The presence of e_{ij} in the error term generates a mechanical correlation between the error term, $(\varepsilon_{ij} - a_3 \cdot e_{ij})$, and the explanatory variable $\phi_{ij} = \phi_{ij}^{*} + e_{ij}$. It can be shown (Wooldridge 2002, p. 75) that $\hat{a}_3$ converges in probability to a fraction $\frac{\text{var}(\phi_{ij}^{*})}{\text{var}(\phi_{ij}^{*})+\text{var}(e_{ij})} < 1$ of the true a_3. This bias is called attenuation bias, since $\hat{a}_3$ is biased towards zero, irrespectively of whether a_3 is positive or negative. The magnitude of the attenuation bias is linked to the so called *signal-to-noise ratio* since $\text{var}(\phi_{ij}^{*})$ is the variance of the correct signal while $\text{var}(e_{ij})$ is the variance of the noise. The larger the latter relative to the former, the larger is the magnitude of the attenuation bias, i.e. if half the variance of $\text{var}(\phi_{ij})$ is noise, the bias would be 50%.[28]

If the distance variable is measured with error, we should expect an attenuation bias in the relevant coefficient. There is a general consensus that the distance coefficient is instead too high and the fact that it is highly persistent and also increasing over time (Disdier and Head 2008) is at odds with the evidence reported by Hummels and Lugovskyy (2006) of a decreasing pattern in freight costs. Many have offered possible explanations; we will point out to a simple mechanical one. If the error-in-variable is not of the classical kind but is instead positively correlated with the distance variable ϕ_{ij}, the bias would tend to be positive and the magnitude would still depend on the signal-to-noise ratio.

Many authors have implicitly worked on the minimization of the signal-to-noise ratio, better defining the relevant meaning of "distance." Some worked along the lines of distance as a proxy for transport costs, and it is surprising to observe (Anderson and van Wincoop 2004) how little is known on transport costs and their different modes, their magnitude and evolution, and their determinants. Hummels and Skiba (2004) focus on the implications of differences in transport

[28]It is worth noting that in a multivariate regression we do not have such a clear and simple result, but the bias will also depend on the correlation between ϕ_{ij} (measured with error) and other covariates. The problem is even more serious with estimates in first-differences, whose aim is to eliminate a possibly omitted fixed effect (Griliches and Hausman 1986). The traditional solution is to find an instrument correlated with distance but not with the error term.

costs across goods on trade patterns, challenging the conventional Samulson's iceberg assumption that transport costs are linear in distance. They show that actual transport costs are much closer to being per unit than iceberg, and they derive clear implications for trade: imports from more distant locations will have disproportionately higher FOB prices. Harrigan (2010) separates air and surface transport costs. Using a Ricardian model with a continuum of goods which vary by weight and hence transport cost, he shows that comparative advantage depends on relative air and surface transport costs across countries and goods. He tests the implication that the US should import heavier goods from nearby countries, and lighter goods from faraway countries, using detailed data on US imports from 1990 to 2003. Looking across US imported goods, nearby exporters have lower market share in goods that the rest of the world ships by air. Looking across exporters for individual goods, distance from the US is associated with much higher import unit values. The effects are significant and economically relevant. Jacks et al. (2008) work in the opposite direction, deriving distance measures from a Anderson-van Wincoop type gravity equation,[29] and finding that the decline in this inherent measure of trade cost explain roughly 55 percent of the pre–World War I trade boom and 33 percent of the post–World War II trade boom, while the rise in that very measure explains the entire interwar trade bust. This stream of research requires a leap of faith on the data-generating process of the trade cost measure and the acceptance that trade costs are the trade empirics equivalent of the Solow's residual: a measure of our ignorance.

Others have worked on Tinbergen's idea that distance could be more than transport costs, moving from spatial distance to economic distance. In analogy with the inclusion of further attractors as explanatory variables, the gravity equation has been therefore augmented with many dyadic variables that could reduce trade (trade policy aside). These variables are mainly associated with a common history of conflict, and are generally found to be highly significant (Martin et al. 2008a, b).

The border between two nations is an equilibrium concept. It is the remaining evidence of the solution of a bargaining process concluding an international conflict and is the fossil of historical events. Since the seminal works of McCallum (1995) and Helliwell (1998), trade economists have wondered how borders could generate a home bias in consumption. Using data on interprovincial and international trade by Canadian provinces for the period 1988–1990, McCallum (1995) showed that, other things being equal, the estimated interprovincial trade was more than 20 times larger than trade between Canadian provinces and US states. The result was striking and largely unbelievable. Anderson and van Wincoop (2003), controlling for multilateral resistance, reduced the border effect by half. Wei (1996), developing a procedure to calculate a country's *trade with self* – a measure rarely reported by official statistics, and relevant on a theoretical basis (being part of the consumer expenditure) – obtained the same reduction for OECD countries and much more for European countries. His estimate of the ratio of imports from self to imports from

[29]See also Novy (2010) for a distance measure derived from heterogeneous firms trade models.

other European countries was 1.7. But he was not controlling for multilateral resistance and was using aggregate data. Disregarding the role of sectoral specialization would attenuate the border effect. Head and Mayer (2000) found that in 1985, Europeans purchased 14 times more from domestic producers (for the average industry) than from equally distant foreign ones. The border effect varies from sector to sector and is related more to consumer tastes than to trade barriers.

We would like to conclude this section on distance by mentioning that over the years, the gravity equation has been applied with great success also to issues which are only marginally related to the cost of physical distance. Blum and Goldfarb (2006) show that gravity holds even in the case of digital goods consumed over the Internet and that do not have trading costs. This implies that trade costs cannot be fully accounted by the effects of distance on trade.[30] Using bilateral Foreign Direct Investment (FDI) data, Daude and Stein (2007) find that differences in time zones have a negative and significant effect on the location of FDI. They also find a negative effect on trade, but this effect is smaller than that on FDI. Finally, the impact of the time zone effect has increased over time, suggesting that it is not likely to vanish with the introduction of new information technologies. Portes and Rey (2005) show that a gravity equation explains international transactions in financial assets at least as well as goods trade transactions. In their analysis, distance proxies some information costs, information transmission, an information asymmetry between domestic and foreign investors. Tinbergen would have been happy to know it, since he proposed information as a possible further explanation of the role of distance (Tinbergen 1962, p. 263). Guiso et al. (2009) go even further, finding that lower bilateral trust leads to less trade between two countries, less portfolio investment, and less FDI. The effect strengthens as more trust-intensive goods are exchanged.

4.4.2.4 Trade Policy

As we pointed out in the introduction, the original use of the gravity equation by Tinbergen was "to determine the normal or standard pattern of international trade that would prevail in the absence of trade impediments," which resulted in the evaluation of the effect of the British Commonwealth and of other FTA. The wider use of the gravity equation has still remained the same: the ex post evaluation of the trade-enhancing effect of preferential trade policy.

The mainstream approach to preferential trade policy evaluation still follows Tinbergen's original strategy, defining the presence of FTA or Custom Unions (CU) or any specific preferential trade policy regime [i.e. Generalised System of

[30]Blum and Goldfarb (2006) also show that Americans are more likely to visit websites from nearby countries, even controlling for language, income, and immigrant stock. For taste-dependent digital products, such as music, games, and pornography, a 1% increase in physical distance reduces website visits by 3.25%. On the contrary, for non-taste-dependent products, such as software, distance has no statistical effect.

Preferences (GSP), African, Caribbean and Pacific (ACP) Partnership, Everything But Arms (EBA), in the case of the European Union (EU)] with positive realization of a Bernoully process. In all these cases, as in Fig. 4.1, the trade effect of the preferential trade policy is the marginal effect of a dummy variable that takes the value of one if the preferential trade policy affects the imports of country *i* from country *j* (in sector *s* at time *t*). The advantage of this strategy is in the ease of implementation. The list of existing FTA, CU, or specific preferential trade policies is generally available online[31] and subsets are included in many datasets used and made available by experts in the field.[32] The disadvantages are that the dummy identification for policy measures implies that all countries included in a treated group are assumed to be subject to the same dose of treatment, which may be correct in the case of non discriminatory policy (e.g. the Most Favored Nation (MFN) clause of the GATT/WTO agreement) but which is false in the case of non reciprocal preferential agreements. In addition, the treatment gets confounded with any other event that is specific to the country-pair and contemporaneous to the treatment (De Benedictis and Vicarelli 2009). Moreover, questions related to the effect of a gradual liberalization in trade policies cannot be answered using dummies, and the trade elasticity to trade policy changes cannot be estimated. Since this is the most common event (trade policy *non facit saltus*, at least not all the times shifts from zero to one) the use of a dummy for preferential trade policy can be a relevant shortcoming.

An alternative exists, and it is largely explored in this volume. It consists in switching from a dummies strategy to a continuous variables strategy, quantifying the preferential margin that the preferential agreement guarantees. This alternative strategy has been fruitfully used by Francois et al. (2006), Cardamone (2007) and Cipollina and Salvatici (2010). It opens an interesting research agenda and also offers some methodological challenges and some puzzling results.[33] These issues are discussed at length in Chapter 3.

Some issues are however worth discussing also in this context. The first is related to the choice of the dependent variable and its consequences. Generally, the stream of literature adopting a dummy strategy focuses on aggregate effects, uses aggregated data, while all papers adopting the alternative strategy of preferential margins

[31]The WTO collects all Trade Agreements that have either been notified, or for which an early announcement has been made, to the WTO (http://rtais.wto.org/UI/PublicMaintainRTAHome.aspx). The World Bank – Dartmouth College Tuck Trade Agreements Database can also be consulted at http://www.dartmouth.edu/~tradedb/trade_database.html.

[32]Andrew Rose's homepage (http://faculty.haas.berkeley.edu/arose/RecRes.htm) is a great example of data sharing. It has encouraged new research and promoted the good practice of replicability in empirical research.

[33]Francois et al. (2006) estimate of the trade policy elasticity has a huge variance and also include some *negative* cases. This result is by no means exclusive to this stream of literature. Also some dummy strategy papers find negative coefficients to preferential dummies (Martínez-Zarzoso et al. 2009).

variables focus on disaggregated data on trade.[34] This alternative strategy expands the panel data along the sectoral dimension, and is therefore more demanding in terms of specific knowledge required, data mining, accuracy in the derivation of the preferential margin,[35] and caution in the aggregation of tariff/products lines, from high level of product disaggregation (often at the 8th or even higher number of digits) to more aggregated data. Inaccurate aggregation could lead to a serious bias. But if precautions are taken on all the complications implicit in this approach, the higher level of information would increase the chance of more precise estimation of causal effect of trade policy. This is currently the most challenging problem of this literature.

The second issue is related to the exogeneity of trade policy. Baier and Bergstrand (2004, 2007) convincingly argue that the chance that the trade policy variable could be highly correlated with the error term is not irrelevant. The possible reverse causation between trade and trade policy could generate an endogeneity bias in the OLS estimates due to self-selection.[36] The same can happen if trade policy is measured with error (as certainly is in the dummy strategy case) or if it does not include relevant missing components (non-tariff barriers) that will end up in the error term. All this calls for an instrumental variable approach. And this is true for both the dummy and preferential margin strategies.

As suggested by Baier and Bergstrand (2007) and others, a possible solution to the omitted variable bias is the use of panel data techniques, that allow to control for time-varying unobserved country heterogeneity, and time-invariant country-pair unobserved characteristics. When instruments are rare this can be a proficuous alternative. On the other hand, the selection bias can be controlled for using a Heckman correction (Helpman et al. 2008; Martínez-Zarzoso et al. 2009).

We would like to conclude this section with a short mention of the role of counterfactuals and control groups in trade policy evaluation. While there is widespread consensus on the relevance of the modern literature on program evaluation (Imbens and Wooldridge 2009), its application to trade policy issues is still rare. Since the gravity equation appears to be appropriate to estimate the causal effect on trade volumes of an average trade policy treatment, some effort should be devoted to the appropriate definition of the treatment (especially in the case of preferential margin), the timing of the treatment, the suitable control group, the counterfactual and the share of the population affected by the treatment when an instrumental variable method is used to estimate average causal effects of the

[34]See Chapter 5 of this volume for a Meta Analysis of the literature on EU preferential trade policy.

[35]See Chapter 3 of this volume. Chapter 10 also shows that the different formula adopted to derive the preferential margins matter significantly for the assessment of the existence and extent of preference erosion for developing countries.

[36]It is difficult to argue that countries enter a preferential agreement at random. Whereas it is hard to observe the original motives that lead to the signing of the agreement, it is reasonable that those motives could be correlated with trade volumes. This gives rise to the selection bias. In particular, the estimated trade policy coefficient will be upward biased if the omitted variables guiding the selection and the trade policy variable are positively correlated.

treatment. Propensity score matching estimators have been used by Persson (2001) and, showing that, in both cases, the relevant policy coefficient is substantially reduced. This literature is still in an embryonic phase, and the one explored by Millimet and Tchernis (2009) through propensity score is by no means the only possible weighting scheme to apply to the gravity equation (Angrist and Pischke 2008). Future research along these lines is required, and from a policy point of view, any step from the analysis of the average treatment effect towards the identification of heterogeneous treatment effects among the countries in the treatment group has to be encouraged.

4.5 "New" Problems and New Solutions

Having described the main components of the gravity equation, there are still some issues – potentially problematic – that deserve mention before bringing this chapter to a close. Some of these issues are well known, others are less so. The literature offers some possible solutions, some of which are firmly established, others are still under debate. We list them for the sake of the reader that wants to explore them further.

4.5.1 The Zero Problem and the Choice of the Estimator

One well recognized problem in empirical trade is that trade datasets often contain zeroes: the cross-country trade matrix is sparse. The conceptual reason why this is the case is exposed at length in Section 4.3. From an empirical point of view, the number of zeroes in the matrix increases with the increase in the level of disaggregation of the data and with the inclusion of smaller and poorer countries. At the aggregate country level, for the year 2000, only about 50 percent of the trade cells had a positive entry. The traditional *log–log* form of the gravity equation calls for particular caution in dealing with zeroes. Since it is not possible to raise a number to any power and end up with zero, the log of zero is undefined, and zero-trade flows cannot be treated with logarithmic specifications. At the same time, they need to be dealt with since they are non-randomly distributed. They indicate absence of trade, hence suggesting that barriers to trade are prohibitive to allow a particular trade relationship to take place at a given demand and supply.

What to do with the zeroes? A number of methods have been explored and proposed by the literature. Here we provide a summary of the most popular of these methods. A first possibility is to ignore the zeroes. However, this would be acceptable only if zeros were the result of an approximation of small trade flows. In this case, the zero-value has no specific meaning and is not a symptom of a self-selection process, as in the presence of distortions due to heterogeneity in exports. By contrast, if the zeros are a sign of selection, a second solution is available to the

researcher: to replace them with a very small positive trade flow, i.e. replace all observations in the data-series by $x_{ij} + 1$. However, this apparently innocuous procedure leads to an inconsistent estimator. Third, assuming that the problem is not of selection but truncation (censored data), the Tobit estimator may be used, provided that the truncation value is known. If this is not the case, the inconsistency of the estimator cannot be avoided. Finally, one can control for the selection bias by means of a Heckman procedure. Indeed, the most popular way to correct for the selection bias is the Heckman 2-stages least squared estimation that introduces in the specification the inverse of the so-called Mills ratio (Heckman 1979).[37] However, in order to do so one needs variables that may explain the selection (zero or positive trade) but not the value of traded good, when this is positive. The exclusion restriction is crucially relevant in this case, and if the variable included in the selection equation also affects the outcome variable, it can lead to the researcher preferring simple OLS to the Heckman procedure (Puhani 2000). Helpman et al. (2008) for example, propose as selection variable the use of the regulation cost of firm's entry. This is a variable collected and analysed by Djankov et al. (2002). This choice is theory-driven, since, as aforementioned the fixed cost of entry only affects the extensive margin of trade under models of firm heterogeneity. Unfortunately, due to the limited data coverage, the costs in terms of sample size reduction are heavy. Hence, even Helpman et al. (2008) in their main results opt for an alternative measure: common religion. The problem with this choice however is that, from previous analyses we know that this type of attractor affects both the extensive and the intensive margin, so that the exclusion condition is violated. In conclusion, the question of the most appropriate selection variable is still open and more research on the topic is needed.

The evidence on the non-triviality of zero-trade flows in data and the growing importance of micro-foundations based on international trade models with firm heterogeneity have pushed researchers to seek solutions. Given the inability of log-linear models to efficiently account for zeroes, the emphasis has moved from OLS estimators to non-linear estimators. In an influential paper, Santos-Silva and Tenreyro (2006) propose an easy-to-implement strategy to deal with the inconsistency occurring when the gravity equation is estimated with OLS using a *log–log* functional form, in the presence of heteroskedasticity and zero trade flows. When the cross-country trade matrix is sparse, the assumption in (4.1) of a (log) normally distributed error term ε_{ij} is violated. In such cases, Santos-Silva and Tenreyro recommend the use of a Poisson Pseudo Maximum-Likelihood (PPML) estimator, using a log-linear function instead of log–log one. A sequel of contributions centered on the relative performance of different nonlinear estimators has followed. The econometric literature on count data (Cameron and Trivedi 2005), applied to non-negative integer values, offers different Poisson-family alternatives to PPML (Burger et al. 2009). How to choose among them is not always straightforward and

[37]The inverse Mills ratio, named after the statistician John Mills, is the ratio of the probability density function over the cumulative distribution function of a distribution.

the practitioner should always be guided by the structure of the data, the level of overdispersion and the assumptions she is willing to impose on the data. As an example, the Poisson model imposes some conditions on the moments of the distribution assuming equidispersion: the conditional variance of the dependent variable should be equal to its conditional mean (and equal to the mean occurrence rate). This is often a too strong assumption, mostly because it is equivalent to say that the occurrence of an event in one period of time (a zero in the trade flow matrix) is independent of its occurrence in the previous period. Is this reasonable?

If the data is characterized by overdispersion, it is possible to correct for between-subject heterogeneity using a Negative Binomial Regression Model (NBPML). NBPML is essentially a Poisson model with the same expected value of the dependent variable as before, but with a variance that takes the form of an additive (quadratic) function of both the conditional mean and a dispersion parameter capturing unobserved heterogeneity. Therefore, not correcting for overdispersion will still lead to consistent estimates of the dependent variable but to a downward bias in the standard errors of the variables of interest. By using NBPML and allowing the dispersion parameter to be different from zero, one can obtain correct standard errors and can properly test if a NBPML estimator is to be preferred to a PPML estimator.[38]

When the number of zeroes is much greater than what is predicted by a Poisson or Negative Binomial distribution (as it is often the case with disaggregated data) it is possible to rely on Zero-Inflated Poisson Model (ZIPPML) or Zero-Inflated Negative Binominal Model (ZINBPML). Both models assume that excess of zeros in the data is generated by a double-process (as in hurdle models), a count process (as in PPML and NBPML) supplemented by a binary process. If the binary process takes a value of zero then the dependent variable assumes a value zero. If the binary process takes a value one then the dependent variable takes count values 0, 1, 2, ... coming from a Poisson density or a negative binomial density. In both cases zeroes occur in two ways: as a realization of the binary process and as a realization of the count process when the binary random variable takes a value of one.

This choice is not harmless because the estimate of the first moment of the distribution changes between PPML and ZIPPML (as for the negative binomial case). The issue leads to a problem of inconsistency on top of the problem of efficiency. Using a count regression when the zero-inflated model is the correct specification implies a misspecification, which will lead to inconsistent estimates.

Opting for a ZIPPML or a ZINBPML estimation offers some advantages since it allows to study separately the probability of trade to take place, from the volume of trade, giving insights both into the intensive and the extensive margin of trade. At the same time, the two-part modeling, because of the form of the conditional mean specification, makes the calculation of marginal effects more complex.

[38]Cameron and Trivedi (2005 p. 676) suggest using a likelihood ratio test on the dispersion parameter to test whether it is equivalent to use a NBPML or a PPML estimator.

To conclude, the literature offers two main strategies to deal with the zeroes problem: a Heckman two-step procedure (controlling for heteroskedasticity) or count data (two-part) modeling. Both strategies need to take seriously the exclusion restriction. In both cases the researcher should pose herself a simple and difficult question: where are all those zeros coming from? Answering convincingly (Cipollina et al. 2010) is the prelude of a correct estimation strategy.

4.5.2 Dynamics

Dynamics is largely a missing piece in the gravity model story. Since Tinbergen (1962, p. 263) "... no attention is paid to the development of exports over time." By and large, this candid admission is still the norm (Eichengreen and Irwin (1995) are an exception). However, there are at least two good reasons to take dynamics into consideration. The first one is a direct consequence of deriving the gravity equation from a micro-founded trade model with heterogeneous firms. As shown in (4.17), if the decision of the firm to sell its products abroad (intensive margin) depends on the firm's ability to cover the sunk cost of entry in the foreign market, it would imply that the firm's decision today will be dependent on its past decisions. Therefore, the export process should be autoregressive. To put it differently, trade models with firm heterogeneity tell us that trade is essentially an entry and exit story. Firms enter and exit from the international markets as a consequence of a selection process on productivity, a learning mechanism, and according to the nature of exogenous shocks on the cost of distance. Some promising attempts (Costantini and Melitz 2008) are already underway.

The second reason is in the empirical counterpart of this proposition. Bun and Klaassen (2002), De Benedictis and Vicarelli (2005) and Fidrmuc (2009) all find strong persistence in aggregate trade data, and countries that trade with each other at time $t - 1$ also tend to trade at time t. This evidence has also been reframed by Felbermayr and Kohler (2006) and Helpman et al. (2008, p. 443) that emphasised that "... the rapid growth of world trade from 1970 to 1997 was predominantly due to the growth of the volume of trade among countries that traded with each other in 1970 (the intensive margin) rather than due to the expansion of trade among new trade partners (the extensive margin)."

The introduction of dynamics in a gravity panel setting raises serious econometric problems due to the inconsistency of the estimators generally used in static panel data. If country specific effects are unobserved, the inclusion of the lagged dependent variable on the right-hand side of the equation leads to correlation between the lagged dependent variable and the error term that makes least square estimators biased and inconsistent.[39]

[39]See De Benedictis and Vicarelli (2005) for a discussion of the issue in the context of the gravity model.

Dynamic panel data models offer different options to the practitioner (Matyas and Sevestre 2007). The ones explored so far are the Blundell-Bond system GMM estimator (De Benedictis and Vicarelli 2005; De Benedictis et al. 2005) and the full set of panel cointegration estimators (i.e. the Fully Modified OLS estimator or the Dynamic OLS) that control for the endogeneity of dependent variables (Fidrmuc 2009). Both kind of contributions are exploratory in nature, and much more can be done along these lines of research.

4.5.3 Interdependence and Networks

The last topic we want to raise in these pages is interdependence. Anderson and van Wincoop (2003) have clearly made the case that the role played by the multilateral dimension of trade in the analysis of bilateral trade flows should not be disregarded, due to both theoretical and empirical reasons. Empirically, it was already mentioned in Section 4.4.2.1 that Multilateral Resistance is controlled for by means of (time-varying) country fixed effects. This simple procedure is correct, but it has often diverted the attention of the empirical researcher from two related issues. First, the fact that country *i* and country *j* are not independent. Second, the role of the third-country in the choices of *i* and *j*, where the notion of the third-country can be extended to the complete structure of trade links in which *i* and *j* are involved.

Dealing with the first issue, we know from disciplines that make frequent use of relational data, such as sociology and psychology, that dyadic observations typically violate the assumption of independence of observations, i.e. *i* and *j* should be considered as being part of a group g. This implies that we cannot rely anymore, as in (4.1) on the assumption of an i.i.d. stochastic term ε_{ij} (Lindgren 2010).

The traditional robust standard errors procedure is not sufficient to correct the error structure and may lead to biased estimated errors and erroneous statistical inference. Recent work by Cameron et al. (2010) shows that the appropriate way to control for such interdependence is to consider the potential correlation within group g in the covariance matrix, clustering the errors around g. This practice has now become more frequent, and many recent empirical estimates of the gravity equation report standard errors clustered at the country-pair level. This implies that, when a cluster is identified, standard errors need to be clustered. Indeed, it is not sufficient to include in the regression a fixed-effect parameter for each cluster a country-pair dummy since the fixed-effect centers each cluster's residual around zero but it does not affect the intra-cluster correlation of errors.

While the concept is relatively simple to grasp, the practice is more complicated. In the gravity equation there may be several choices for clustering. As for more general cases, in a panel data framework each single country can be considered as a cluster along the time dimension. Therefore, we shall cluster the errors by *i* if we believe that countries have a memory of their past decisions and project it onto the future. Moreover, if the two countries belong to the same

preferential trade agreement V_{ij} we shall cluster by it as well. When average trade flows are used, as in eq. 4.18, exports of country i to country j are not anymore independent from exports of country j to country i. Hence, the two observation must be clustered in one single data point. If a hierarchical structure exists, we can nest the level of clustering choosing the most aggregate level. The caveat to keep in mind is that the number of clusters should be sufficiently large and sufficiently balanced. Researchers are therefore invited to use two-way or multi-way clustering, clearly discussing the adopted clustering structure instead of leaving the clustering procedure as a side note to the summary table of regression's results.

As far as the role of the third-country in the choices of i and j is concerned, their relevance has been widely recognized in trade theory, but only recently the empirical literature (see Baier and Bergstrand 2004 and Magee 2003, 2008 on FTA) has started considering how to include interdependence in the analysis. Baltagi, Egger and Pfaffermayr (2008) use a spatial lag panel data model to estimate the effect of regional trade agreements on inward FDI from Western European countries. They use the spatial weighting matrix to capture interdependence in a panel setting. The inclusion of this matrix is crucial for their results. First of all, they find spatial correlation in the data, leading to transmission effects of the 1990s preferential trade liberalization between the European Union and the Central and Eastern European Countries (CEEC). They find that the so-called Europe Agreements had a negative impact on Western Europe inward FDI, both in 1995 (when four agreements were ratified) and in 1999 (when only a single was ratified). At the same time, the CEEC experienced on average the strongest positive effects. Finally, they also find that the negative effects on FDI flowing into Western Europe is largely offset by the positive effects on FDI going to Central and Eastern European countries. This empirical work clearly shows that the analysis of the third-country effect is crucial in determining the relocation of FDI from Western European host countries to Eastern European host countries flowing from the Europe Agreements.

Both Egger and Larch (2008) and Chen and Joshi (2010) focus on the formation of an FTA given the existence of previous FTA. The general prediction of the three-country oligopolistic model of Chen and Joshi is that the role played by third countries is fundamental in understanding the formation of FTA. Just to give an example: if country i has an FTA with country j, but a third country, say k does not, country i and country k are more likely to establish an FTA when country i has a sufficiently large market size and high marginal cost of production relative to country j and the transport cost between the two is relatively low. This proposition is confirmed by the data on 78 countries between 1991 and 2005. The contribution by Chen and Joshi opens a promising research agenda on the role of third-country effects in trade policy, but the empirical analysis should be confirmed after a proper clustering of the error structure.

The relevant role played by the third-country in shaping the decision of country i and country j put in to question the fact that the role of countries interdependence could be relegated to the inclusion of the Multilateral Resistance term in the gravity equation.

In (4.10), Ω_i represents the average of all importers market demand – weighted by trade costs, while firms competition in market j is factored in via

the price index P_j, which is also an average value. Therefore the Multilateral Resistance term includes in the gravity equation the *average* third-country effect. Which is perfectly sound from a modeling perspective, given the strong symmetry at the sectoral and country level assumed in the gravity model. On the other hand, the strong asymmetry revealed by the data cast some doubts on the fact that the average third-country effect should be a sufficient statistic to capture complex interdependence. Along these lines, De Benedictis and Tajoli (2010) generalise the third-country effect using network analysis (Jackson 2008). They focus on the interconnected structure of trade flows describing the changing topological property of the world trade network along time. They further focus on the extensive margin of trade, considering trade flows as a binary variable that takes the value 1 when trade occurs, and zero otherwise. They also calculate some network statistics to measure the level of interconnection of each country and the level of relative centrality of a country with respect to the whole trade network. The inclusion of these statistics in a gravity equation turns out to be significant with a positive and economically relevant coefficient, even when the standard errors were bootstrapped to take into account the correlation in the error structure (Davison and Hinkley 1997). This approach to interdependence in trade relations is in its infancy, and it may be fruitful to delve into the full implications of the multilateral dimension of the gravity model.

4.6 Conclusions

This chapter has shown how the 50-year long progress in the research agenda on gravity equation revolves around issues that were already raised in Tinbergen's original formulation of the relationship. The numerous empirical and theoretical contributions however, have allowed over the years to bring new, more efficient solutions to the old problems and to generate consensus around some new key issues. For example, it is now widely accepted that nominal variables should be used. Similarly panel estimations are to be preferred to cross-section estimates in most cases and fixed effects should be selected not blindly but with a view at how to best isolate developments in the variable of interest. Moreover, it is now widely accepted that distance is only an imperfect proxy for trade costs, that its effect on the extensive and intensive margin of trade differs from each other and that zero values contain information that should not be neglected.

Despite the fact that the state of the art on gravity equation has become very sophisticated, there are still many areas where further research is warranted. The analysis of gravity models on firm data is a promising avenue of research, as shown among others by Bernard et al. (2007). The changing nature of trade relationships calls for a re-evaluation of gravity specifications in particular contexts. For example, Baldwin and Taglioni (2010) show that the gravity equation breaks down when trade in parts and components is important. Structural estimations of gravity are also becoming very popular. Egger et al. (2010) and Anderson and Yotov (2010) are

two important examples of this promising literature. Finally the increasing widespread availability of data is making quasi-natural experiments more common also in the evaluation of trade impediments (Feyrer 2009).

All the progress made from Tinbergen on to clarify the mysterious fitting power of the gravity equation is now at the disposal of a new generation of correlation hunters, wishing to move towards a better causal evaluation of trade enhancing policies.

Acknowledgments This essay has been written while Luca De Benedictis was visiting ARE at the University of California, Berkeley and EIEF in Rome. He gratefully acknowledges UCB and EIEF for the great hospitality. Luca De Benedictis also acknowledges the financial support received by the Italian Ministry of Education, University and Research (Scientific Research Program of National Relevance 2007 on "European Union policies, economic and trade integration processes and WTO negotiations" – PUE & PIEC). We are grateful to Davide Castellani, Michele Di Maio, Lucia Tajoli, Claudio Vicarelli and especially Luca Salvatici for the comments and the many conversations on the topic. Any remaining errors are solely our responsibility.