AGRODEP
African Growth & Development Policy
modeling consortium
FACILITATED BY IFPRI

AGRODEP Technical Note TN-04

March 2012

The Gravity Model in International Trade

Version 1

Luca Salvatici

Abstract Since Jan Tinberben's original formulation (Tinbergen 1962), gravity has long been one of the most successful empirical models in economics. Incorporating the theoretical foundations of gravity into recent practice has led to a richer and more accurate estimation and interpretation of the spatial relations described by gravity. Recent developments are reviewed here and suggestions are made for promising future research.

1 Introduction

This gravity guide provides a literature review and a methodological discussion about the gravity equation. 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 its ability to correctly approximate bilateral trade flows makes it one of the most stable empirical relationships in economics (Leamer and Levinsohn 1995).

Over the years there has been dramatic progress both in understanding the theoretical basis for the equation and in improving its empirical estimation (De Benedictis and Taglioni, 2011). This review cannot and does not intend to be a complete survey of a huge (and still increasing) literature. The aim is to provide the reader with an informed perspective on the empirical issues associated with the estimation of the gravity equation. To this end, we deliberately scant or omit some topics in order to have the possibility to discuss how to achieve theoretically sound gravity specifications. In the following, then, 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.

We organize our review into 5 parts. Section 2 discusses the theoretical general equilibrium foundations for the gravity equation for trade. Section 3 deals with the role of frictions inhibiting the flows of goods. While distance has long been recognized as a prominent friction impeding trade, there are numerous other impediments to these flows, some of which are “natural” – such as being landlocked – and some of which are “artificial” (or “man-made”) – such as trade policies. Section 4 discusses very recent developments in the theoretical foundations for the gravity equation, and econometric implications from the use of disaggregated data. Section 5 concludes.

2 Theory-based specifications for the gravity model

In Tinbergen's version of the gravity equation, X_{ij}, the size of the trade flow between any pair of countries is stochastically *determined* by: (*i*) E_i, the amount 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; (*ii*) E_j, the size of the importing market, measured by its GNP, also converted in US dollars; (*iii*) 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.[1] *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* 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. The strategy of con-

[1] Results turn out to be not much different using exports or imports.

sidering 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: we will come back to this issue in the following. As customary, a i.i.d. stochastic term ε_{ij} was also included:

$$\ln X_{ij} = \underbrace{a_0}_{\text{constant}} + \underbrace{a_1 \ln M_i + a_2 \ln M_j}_{\text{economicattractors}} + \underbrace{a_3 \ln \phi_{ij} + a_4 N_{ij}}_{\text{economicattractors}} + \underbrace{a_5 V_{ij}}_{\text{policy}} + \underbrace{\varepsilon_{ij}}_{\text{error term}} \quad (1)$$

In the original estimation by Tinbergen (1962), the coefficients of GNP and distance had what became "the expected signs" in all subsequent analyses – the coefficients of the economic attractors were positive and the one of distance was negative – and resulted relevant and significant. The specification however, left room for improvement, and the positive but relatively small role of trade preferences was an issue that stimulated further inquiry.

Let's start from the first highlighted term: *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 however theory-driven and based on the assumption that the gravity equation is derived from an microeconomic model where income and tastes for differentiated products are given. Three decades of theoretical work has shown that the gravity equation can be derived from many different – and sometimes competing – trade frameworks. The progress in modeling of gravity has yielded two distinct rationales for the same observationally equivalent model of the distribution of economic flows between origins and destinations, one based on a demand function with Constant Elasticity of Substitution (CES) *à la* Armington (Anderson, 1979), where each country produces and sells goods on the international market that are differentiated from those produced in every other country; the other based on the supply side through 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). Contrary to what is implied by models of monopolistic competition *à la* Krugman, not all existing firms operate on international markets. 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. The critical implication of firm heterogeneity for modeling the gravity 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).

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 mispecifying equation (1). Irrelevant of the theoretical framework of reference, most of the modern mainstream foundations of the gravity equation are variants of the demand-driven model firstly described in Anderson (1979). Here, we will mainly rely on the Anderson and van Wincoop (2003) and Baldwin and Taglioni (2006) derivations, using standard notation to facilitate the exposition.

2.1 THE BASIC MODEL

According to Anderson (2010), from a modeling standpoint, gravity is distinguished by its parsimonious and tractable representation of economic interaction in a many country world. This distinguishing feature of gravity is due to its modularity: the distribution of goods or factors across space is determined by gravity forces conditional on the size of economic activities at each location.[2] Modularity readily allows for disaggregation at any scale and permits inference about trade costs not dependent on any particular model of production and market structure in full general equilibrium.

Modularity can be met by restricting the preferences (and/or technology) such that the cross effects in demand between classes of goods (either intermediate or final) flow only through aggregate price indexes. This demand property is satisfied when preferences or technology are homothetic and weakly separable with respect to a partition into classes whose members are defined by location, a partition structure called the Armington assumption. Notice that the assumption implies that goods are purchased from multiple sources because they are evaluated differently by end users, goods are differentiated by place of origin. It is usual

[2] Anderson and van Wincoop (2004) call this property *trade separability*.

to impose identical preferences across countries.[3] Separability also implies that each goods class has a natural quantity aggregate and a natural price aggregate, with substitution between goods classes occurring as if the quantity aggregates were goods in the standard treatment. Homotheticity ensures that relative demands are functions only of relative aggregate prices.

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. Expenditure shares in the CES case are given by:[4]

$$\frac{X_{ij}}{E_j} = \left(\frac{\beta_i p_i t_{ij}}{P_j} \right)^{1-\sigma} \qquad (2),$$

p_i is their factory gate price, $t_{ij} > 1$ is the trade cost factor between origin i and destination j, and E_j is nation-*j* expenditure. The 'distribution parameters' β_i for varieties shipped from i could be exogenous or, in applications to monopolistically competitive products, proportional to the number of firms from i offering distinct varieties (Bergstrand, 1989). The CES price index is given by:

$$P_j = \left(\sum \left(\beta_i p_i t_{ij} \right)^{1-\sigma} \right)^{1/(1-\sigma)} \qquad (3).^5$$

[3] There are, indeed, differences in demand across countries, such as a home bias in favor of locally produced goods. In practice it is very difficult to distinguish demand side home bias from the effect of trade costs, since the proxies used in the literature (common language, former colonial ties, or internal trade dummies, etc.) plausibly pick up both demand and cost differences.

[4] The shares are invariant to income, since preferences are homothetic.

[5] For intermediate goods, the same logic works replacing expenditure shares with cost shares.

Let us stress the point that the previous 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 *E_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, equation (2) tells us that the omission of the importing nation's price index P_j from the original gravity equation described in equation (1) leads to a misspecification. 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 (De Benedictis and Taglioni, 2011).

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 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.

Conversely, 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 GDPs 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 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, and set it equal to overall production. This can be written as follows:

$$E_i = \sum_j X_{ij} = p_i^{1-\sigma} \sum_j \left(\beta_i t_{ij}\right)^{1-\sigma} \frac{E_j}{P_j^{1-\sigma}} \qquad (4).$$

Solving for $p_i^{1-\sigma}$ yields:

$$p_i^{1-\sigma} = \frac{E_i}{\Omega_i} \qquad (5),$$

with:

$$\Omega_i = \sum_j \left(\beta_i t_{ij}\right)^{1-\sigma} \frac{E_j}{P_j^{1-\sigma}} \qquad (6),$$

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 equation (6) in equation (2) yields a basic but correctly specified gravity equation:

$$\frac{X_{ij}}{E_j} = \left(\beta_i t_{ij}\right)^{1-\sigma} \frac{E_i}{P_j^{1-\sigma}\Omega_i} \qquad (7).$$

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 (De Benedictis and Taglioni, 2011).

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 two terms Ω_i and $P_i^{1-\sigma}$ with the well known term of *multi-*

lateral resistance; which can be empirically controlled for by a time-invariant country-fixed effect.

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 equation (7) is missing a time-varying dimension. An easy and practical solution to match the theory with the data is to introduce time-varying importer and exporter fixed effects.[6] Often however, the need of correcting for omitted variables biases clashes with problems of collinearity with the other variables. 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 (De Benedictis and Taglioni, 2011).

2.2 MULTILATERAL RESISTANCE TERM

The previous model showed that because there are many origins and many destinations in any application, a theory of the bilateral flows must account for the relative attractiveness of origin-destination pairs. Each sale has multiple possible destinations and each purchase has multiple possible origins: any bilateral sale interacts with all others and involves all other bilateral frictions. After this contribution, the omission of a Multilateral Resistance term is considered a serious source

[6] Obviously, in cross-sections, the Anderson van Wincoop specification is sufficient owing to the lack of time dimension.

of bias and an important issue every researcher should deal with in estimating a gravity equation.

In literature three methods are suggested to account for price effects in the gravity equation: (1) the use of published data on price indexes (Bergstrand, 1985, 1989; Baier and Bergstrand, 2001; Head and Mayer, 2000); (2) direct estimation à la Anderson and van Wincoop (2003); (3) or the use of country fixed effects (Hummels, 1999; Rose and van Wincoop, 2001; Eaton and Kortum, 2002; Feenstra, 2002; Redding and Venables, 2000).

The main weakness of the first method is that the existing price indexes may not accurately reflect the true border effects (Feenstra, 2002). Accordingly, Anderson and van Wincoop (2003) estimate the structural equation with nonlinear least squares after solving for the multilateral resistance indices as a function of the observables bilateral distances and a dummy variable for international border.

However, the computationally easier method for accounting for multilateral price terms in cross section – that will also generate unbiased coefficient estimates – is to estimate the gravity equation using country-specific fixed effects. Moreover, since detailed data on consumption shares are not available, the only way to take account of the unobserved shares is to include commodity fixed effects. The advantage of using fixed effect specifications lies in the fact that they represent by far the simplest solution: 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}.

On the other hand, caution 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 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. 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 (De Benedictis and Taglioni, 2011).

2.3 THE ZEROS PROBLEM

Thus far, all treatment of flows has been of a generic good which most of the literature has implemented as an aggregate: the value of aggregate trade in goods for example. However, there is aggregation bias due to sectorally varying trade costs and sectorally varying elasticities of trade with respect costs (see Anderson and van Wincoop, 2004, for analysis and Anderson and Yotov, 2009 and 2010 for evidence on downward bias). The second aggregation problem is specification bias because GDP is a value added concept with a variable relationship to gross trade flows. Disaggregation and use of the appropriate sectoral output and expenditure variables fixes both problems.

Introducing disaggregated goods or firm heterogeneity in models of international trade 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 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.

The gravity specification with disaggregated goods or firm-heterogeneity differs from previous specifications in two broad ways. 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. When per-unit trade costs move, both the intensive and the extensive margin of trade are affected and the degree of competition in the market, plays an important role in the dynamics. On the other hand, 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.

The extension to disaggregated goods is straightforward, but it leads to two types of shortcomings: (i) the elevated percentage of “zero trade flows”; (ii) the impossibility, for some variables, to get information at the level of details at which tariff lines are specified.

3 A piecewise analysis of the gravity equation

3.1 Dependent variable

The gravity equation has also been used extensively for understanding the determinants of observed bilateral foreign direct investment and migration flows, although to an extent less than for trade flows. As with trade flows, the model always fits well. But, in contrast to the recent development of a theory-based gravity model of trade, there has been little progress in building a theoretical foundation (Anderson, 2010). In the following, the discussion will focus on goods movements.

According to De Benedictis and Taglioni (2011), there are three main 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 later on.

Starting with the issue of *typology*, in the large majority of studies the dependent variable is usually a measure of bilateral merchandise trade. 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,[7] 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 important piece of information. Second, if carried out wrongly, averaging leads to mistakes (De Benedictis and Taglioni, 2011).

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.[8] 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

[7] 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.

[8] Since the gravity equation is mostly estimated in logs, the practice of averaging trade flows often results in using the log of the sum of the flows instead of the sum of the logs.

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 (De Benedictis and Taglioni, 2011).

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).[9] 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 (De Benedictis and Taglioni, 2011).-

3.2 Covariates

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.

[9] 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

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. 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 (De Benedictis and Taglioni, 2011).

Frictions that impede international trade flows are usually called “trade costs.” Trade costs can be decomposed into two main sources: “natural” trade costs and “unnatural” (or policy-based) trade costs (Bergstrand and Egger,):

- Natural trade costs refer to those costs incurred largely – though not exclusively – by geography. Distance between a pair of countries is an example of a natural trade cost.
- Policy-based trade costs refer to those additional costs impeding trade if physical distances (or other natural costs) were absent. These costs are largely “man-made” or “artificial,” and are mainly attributable to policy decisions of governments.

Anderson and van Wincoop (2004), in a comprehensive discussion of trade costs, estimate that the average cost of delivering a good from the point of manufacture to the destination (including international tariff and non-tariff policy barriers) is about a 170 percent add-on to the cost of producing the good. They decompose this into 74 percent international trade costs (21 percent natural and

note that China, besides providing SITC classifications also provides data series for processing trade used (De Benedictis and Taglioni, 2011).

44 percent international-border-related: (74=1.21*1.44-1).) and 55 percent associated with domestic retail and wholesale distribution costs (1.7=1.55*1.74-1).

3.2.1 Trade frictions

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). Others used great-circle or orthodromic formulas (De Benedictis and Taglioni, 2011).[10] Nowadays, all most common distance measures across virtually all country pairs in the world are freely available online[11] or can be obtained from the applets of the most important geo-representations available on the web. 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). 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.

[10] 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: $\phi_{ij} = a\cos\ (\sin(lat_i)\cdot\sin(lat_j)+\cos(lat_i)\cdot\cos(lat_2)\cdot\cos(long_j - long_i))\cdot R$; where R= 6371 is the radius of the earth, in km.

[11] 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 (De Benedictis and Taglioni, 2011).

The most popular assumption is to restrict the trade costs so that the distribution of goods uses resources in the same proportion as the production of those same goods. Samuelson (1952) invented iceberg melting trade costs in which the trade costs were proportional to the volume shipped, as the amount melted from the iceberg is proportional to its volume. Mathematically, the generalized iceberg trade cost is linear in the volume shipped. However, 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 (De Benedictis and Taglioni, 2011).

The iceberg metaphor still applies when allowing for a fixed cost, as if a chunk of the iceberg breaks off as it parts from the mother glacier. Fixed costs are realistic and potentially play an important role in explaining why many potential bilateral flows are equal to zero. 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 (De Benedictis and Taglioni, 2011).

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 rereported by Hummels and Lugovskyy (2006) of a decreasing pattern in freight costs. Many have offered possible explanations. Felbermayr and Kohler (2006) argue that the distance puzzle may simply reflect a mis- specification of the gravity equation that arises from inadequate treatment of the dual margin of world trade; while De Benectis and Taglioni (2011) 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, the bias would tend to be positive and the magnitude would depend on the signal-to-noise ratio.

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. The most common measure of transport costs is referred to commonly as the "cif-fob factor" (Bergstrand and Egger,). Trade flows from one country to another are often measured "free on board" (fob), which refers to the value of a shipment of goods delivered to and put "on board" an overseas vessel for potential shipment. The same trade flows are often also measured reflecting "cost-insurance-freight" (cif), which refers to the value of the same shipment at the destination port (or airport), including the cost of insurance and freight charges. The ratio of these two values minus unity provides an *ad valorem* "rate" for the add-on associated with international transport. Baier and Bergstrand (2001) report that average cif-fob factors for 16 Organization for Economic Co-operation and Development (OECD) countries in 1958 and 1988 were 8.2 percent and 4.3 percent, respectively. Moreover, they show that the decline in such costs explain about 8 percent of the increase in world trade from the late 1950s to the late 1980s, after accounting for expanding GDPs and falling tariffs.

While cif-fob factors are the most common method for estimating the costs associated with transit of a good from country i to country j, this measure is not without flaws. Hummels (2007) raises the concern that this measure may underestimate the true transport costs. He finds that the average level (and variances) of cif-fob factors in disaggregated data is much higher than that in aggregate data.

Time also is a natural trade cost. It takes longer on average for the same good to move between countries than within countries. Hummels (2001) found that every additional day in ocean travel for a shipment to arrive reduces the probability of outsourcing manufactures by 1 percent. In the same vein, 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.

Jacks, Meissner and Novy (2008) work in the opposite direction, deriving distance measures from a Anderson-van Wincoop type gravity equation, 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 (De Benedictis and Taglioni, 2011).

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). Many studies, the large part of them in a cross-sectional setting, augment the gravity equation with variables that could ease trade costs. Sharing a common 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.[12]

Many of those issues are of interest per se and are worth to be explored. However, the researcher should be aware that most of these variable have in general very low time variability. For this reason, one 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.

[12] See Anderson and van Wincoop (2004) for more discussion.

Finally, De Benedictis and Taglioni (2011) point out 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. 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. 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.

3.2.2 Trade policy

Artificial trade costs can be decomposed into the exhaustive categories of “tariffs” – taxes on goods crossing international borders – and “nontariff barriers” on international trade. While measures of tariff rates are available, nontariff barriers (or measures, NTBs) are even more difficult to quantify. One method of measurement of the importance of NTBs is to calculate the share of industries in a country that are subject to NTBs in that country; this is typically referred to as the “NTB coverage ratio.”

One of the oldest and most prominent uses of the gravity equation has been to estimate the impacts of economic integration agreements (EIAs) – notably, free trade agreements (FTAs), customs unions, and other forms of preferential trade agreements (PTAs) – on trade. 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 with positive realization of a Bernoully process. In all these cases, 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[13] and subsets are included in many datasets used and made available by experts in the field.[14] 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 coun-

[13] 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

try-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 (De Benedictis and Taglioni, 2011).

An alternative exists, and 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 (2010a). It opens an interesting research agenda and also offers some methodological challenges and some puzzling results. For instance, the estimated effects of Regional Trade Agreements (RTAs) vary widely, from study to study and sometimes even within the same study. Cipollina and Salvatici (2010b) by means of meta-analysis techniques, we statistically summarized 1827 estimates collected from a set of 85 studies. After filtering out publication impact and other biases, the MA confirms a robust, positive RTAs effect, equivalent to an increase in trade of around 40%. The estimates tend to get larger for more recent years, which could be a consequence of the evolution from “shallow” to “deep” trade agreements. From the methodological point of view, there appears to be evidence of a significant downward bias due to omitted variables problems, while data measurement and specification problems are less likely to

[14] Andrew Rose’s homepage (http://faculty.haas.berkeley.edu/arose/RecRes.htm) is a great example of data sharing.

produce (statistically speaking) “good results,” and estimates tend to be biased in the opposite direction.

A couple of issues are worth discussing. 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 variables focus on disaggregated data on trade. This strategy expands 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, 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.

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.[15] The same can happen if trade policy is measured with error (as certainly is in the dummy strategy case)

[15] 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 (De Benedictis and Taglioni, 2011).

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 (De Benedictis and Taglioni, 2011).

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; Martinez-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 treatment. Propensity score matching estimators have been used by Pearson (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 (De Benedictis and Taglioni, 2011).

4 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 review to a close.

4.1 The zeros problem and the choice of the estimator

One well recognized problem in empirical trade is that trade datasets often contain zeros: the trade matrix is sparse. The prevalence of zeros rises with disaggregation, so that in finely grained data a large majority of bilateral flows appear to be inactive.

The traditional *log-log* form of the gravity equation calls for particular caution in dealing with zeros. 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. The data presented to the analyst may record a zero that is a true zero or it may reflect shipments that fall below a threshold above zero. In addition there may be missing observations that may or may not reflect true zeros.

The zeros present two distinct issues for the analyst: appropriate specification of the economic model and appropriate specification of the error term on which to base econometric inference. As far as the former is comcerned, one way to rationalize zeros is to modify the demand specification so as to allow 'choke prices' above which all demand is choked off (Anderson, 2010). An alternative economic specification explanation retains CES/Armington preferences and rationalizes zeros as due to fixed costs of export facing monopolistic competitive firms. Helpman, Melitz and Rubinstein (HMR, 2008) develop this idea.[16]

As far as the latter is concerned, 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 zeros. and estimate the log-linear form by OLS. Even without mentioning the fact that the omission of zero flows could strongly reduce the sample and then lead to a considerable loss of information, limiting of the analysis to observations where bilateral trade flows are positive is a significant source of bias since the selected sample is not random. Zeros may be the result of rounding errors. If these rounded-down observations were partially compensated by rounded-up ones, the overall effect of these errors would be relatively minor. However, the rounding down is more likely to occur for small or distant countries and, therefore, the probability of rounding down will depend on the value of the covariates, leading to the

[16] The key mechanism is a Pareto productivity distribution of potential trading firms: the Pareto distribution is capable of capturing the empirical observation that the largest and most productive firms export the most and to the most destinations.

inconsistency of the estimators. The zeros can also be missing observations which are wrongly recorded as zero. This problem is more likely to occur when small countries are considered and, again, measurement error will depend on the covariates, leading to inconsistency.

A second solution is to replace the zeros with a very small positive trade flow, i.e. replace them in the data-series by x_{ij}+1. As a matter of fact, many gravity works perform Tobit estimates by constructing a new dependent variable y = ln(1+Mij). Assuming that the problem is not of selection but truncation (censored data), this is the estimatorto be used according to the econometric theory. However, this procedure relies on rather restrictive assumptions that are not likely to hold since the censoring at zero is not a "simple" consequence of the fact that trade cannot be negative. Zero flows, as a matter of fact, do not reflect unobservable trade values but they are the result of economic decision making based on the potential profitability of engaging in bilateral trade at all. 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).[17] 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. For instance, Helpman et al. (2008) use common religion, a specification that many find dubious. The exclu-

[17] 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.

sion restriction is crucially relevant, 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. Alternatively, Francois suggested the use of a 'network index', namely the number of common partners in trade between two countries. Such an index could be a viable selection variable, since Chaney (2011) showed that once a firm has acquired some foreign contacts, it can meet the contacts of those contacts. The possibility to use existing contacts to find new ones gives an advantage to firms with many contacts: in other terms, the more contacts a firm has, the more likely it is to acquire additional contacts. As a consequence, the entry of individual exporters into a given country is influenced by changes in aggregate trade flows between third countries. In conclusion, the question of the most appropriate selection variable is still open and more research on the topic is needed.

Given the inability of log-linear models to efficiently account for zeros, 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 equation (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).

De Benedictis and Taglioni (2011) rightly warn that the choice is not straightforward and 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.

When the number of zeros 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. However, the 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. In a two-steps procedure, as a matter of fact, an increased probability of registering a positive trade flows in the first stage means that a larger set of products is traded (extensive margin), while a positive coefficient in the second stage refers to a larger volume of trade (intensive margin). 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.

To conclude, the literature offers several strategies to deal with the zeros problem and results are quite relevant: Liu (2009) using a large bilateral panel dataset including zero trade flows and state-of-the-art econometric methods, finds that the GATT/WTO has been very effective in promoting world trade at both the intensive and extensive margins. In all cases, though, one ought to answer a simple (but far from trivial) question: where are *all* those zeros coming from? Cipollina et al. (), for instance, distinguish between two different kinds of zero-valued trade flows: products that are never traded and products that are not traded, but could be (potentially, at least) traded. Hence, a distinction is made between flows with exactly zero probability of positive trade, flows with a non-zero trade probability who still happen to be zero, and positive flows. Since preferential policies cannot possibly influence the first group, in their analysis they only keep exporters that have at least one export flow at the world level for the product, assuming that excluded commodities are not produced, and exclude products that are not imported at all in the foreign markets. This avoids the inclusion of irrelevant information that may bias the estimate, [18] and greatly reduces the dimension of the dataset.

4.2 Dynamics

Dynamics is largely a missing piece in the gravity model story. However there are at least two good reasons to take dynamics into consideration (De Benedictis and Taglioni, 2011). The first one is a direct consequence of deriving the gravity equation from a micro-founded trade model with heterogeneous firms. 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

[18] There is a difference between a good that is not produced and hece is not exported, and a good that is produced but it is still not exported. In the same vein, it should be taken into account that not all products have the potential (or are

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 (De Benedictis and Taglioni, 2011).

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 (De Benedictis and Taglioni, 2011).

5 Conclusions

This review has shown how the 50-year long progress in the research agenda on gravity equation has allowed over the years to bring new, more efficient so-

at risk) to be exchanged because of non-economic reasons: trade embargos, religious prohibitions, etc..

lutions 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. 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 (De Benedictis and Taglioni, 2011).