

AGRODEP Technical Note TN-04

April 2013

The Gravity Model in International Trade

Version 2

Luca Salvatici

AGRODEP Technical Notes are designed to document state-of-the-art tools and methods. They are circulated in order to help AGRODEP members address technical issues in their use of models and data. The Technical Notes have been reviewed but have not been subject to a formal external peer review via IFPRI's Publications Review Committee; any opinions expressed are those of the author(s) and do not necessarily reflect the opinions of AGRODEP or of IFPRI.

Abstract Since Jan Tinbergen'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. 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 its simplest form, the analogy with Newton’s “Law of Universal Gravitation” implies that a mass of goods or labor or other factors of production at origin i , E_i , is attracted to a mass of demand for goods or labor at destination j , E_j , but the potential flow is reduced by distance between them, ϕ_{ij} . Strictly applying the analogy,

$$X_{ij} = E_i E_j / \phi_{ij}^2 \quad (1)$$

gives the predicted movement of goods or labor between i and j , X_{ij} .

The analogy between trade and the physical force of gravity, however, clashes with the observation that there is no set of parameters for which equation (1) will hold exactly for an arbitrary set of observations. Departing from strict analogy, traditional gravity allowed the coefficients of 1 applied to the mass variables and of 2 applied to bilateral distance to be generated by data to fit a statistically inferred relationship between data on flows and the mass variables and distance. Typically, the stochastic version of the gravity equation has the form

$$X_{ij} = a_0 E_i^{\alpha_1} E_j^{\alpha_2} \phi_{ij}^{\alpha_3} \varepsilon_{ij} \quad (2)$$

where α_0 , α_1 , α_2 , and α_3 are unknown parameters.

In the original version by Tinbergen (1962), the *model* is expressed in a *log-log* form, so that the parameters are elasticity of the trade flow with respect to the explanatory variables.¹ With respect to equation (2), *Adjacent countries* are assumed to have a more intense trade than what distance alone would predict; the adjacency is indicated by the dummy variable N_{ij} , that took the value 1 if the two countries share a common land border. Moreover, the equation is augmented with *political* factors: a dummy variable V_{ij} indicate that goods traded received a preferential treatment if they belonged to some unilateral or system of preferences. 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: we will come back to this issue in the following. As customary, a i.i.d. stochastic term ε_{ij} is included: $\ln X_{ij} = \underbrace{a_0}_{\text{constant}} + \underbrace{a_1 \ln E_i + a_2 \ln E_j}_{\text{economic attractors}} + \underbrace{a_3 \ln \phi_{ij} + a_4 N_{ij}}_{\text{distance}} + \underbrace{a_5 V_{ij}}_{\text{policy}} + \underbrace{\varepsilon_{ij}}_{\text{error term}}$ (3).

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.

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.

A first group of gravity models is derived under perfect competition. Anderson (1979) assumes a Constant Elasticity of Substitution (CES) import demand system where each country produces and sells goods on the international market that are differentiated from those produced in every other country. goods are purchased from multiple sources because they are evaluated differently by end users. An alternative derivation of a mathematically equivalent gravity model was proposed by Eaton and Kortum (2002), based on homogeneous goods on the demand side, iceberg trade costs, and Ricardian technology with heterogeneous productivity for each country and good due to random productivity draws. In the former case, like in any other ‘Armington’ structure (i.e., goods are differentiated by place of origin) there are only consumption gains from trade, whereas there are both consumption and production gains in the latter case (Arkolakis et al., 2012)

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.

¹ In Tinbergen’s version (1962), trade flows were measured both in terms of exports and imports of commodities and only non-zero trade flows were included in the analysis.

The existence of trade flows which have a bilateral value equal to zero is full of implications for the gravity equation because in Newton's equation the gravitational force can be very small, but never zero. Even if zeros may reflect mis-reporting and mis-measurement, particularly that of small and poor countries, observed zeros contain valuable information which should be exploited for efficient estimation. As a matter of fact, 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 fact that trade between several pairs of countries is literally zero may signal a selection problem (Chaney 2008; Helpman et al. 2008). In the following it will be shown how appropriate econometric techniques allow to extract more information from the data, particularly relating to the role of distance and other variables affecting the extensive margin of world trade.

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 misspecifying equation (3). 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 (2011), 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.

Gravity-type structures can be obtained imposing two crucial restrictions (Anderson and van Wincoop, 2004). The first requires the aggregator the aggregator of varieties to be identical across countries and CES.³ The CES form, as matter of fact, imposes homothetic (ensuring that relative demands are functions only of relative aggregate prices)⁴ as well as separable preferences (allowing the two stage budgeting needed to separate the allocation of expenditure across product classes from the allocation of expenditure within a product class). As it was already mentioned, product classes are defined by location since goods are differentiated by place of origin: a partition structure known as the "Armington assumption" (Armington, 1969).

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

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 Non-homotheticity has been first presented as an important assumption to explain trade in food products. More recently, Markusen (2010) emphasized the importance of explaining North-North and South-South trade "putting back" per-capita income in trade analysis.

Accordingly, the starting point of Anderson and van Wincoop (2003) is a CES utility function. If X_{ij} is consumption by region j consumers of goods from region i , consumers in region j maximize

$$\left(\sum_i \beta_i^{1/\sigma} X_{ij}^{(\sigma-1)/\sigma} \right)^{\sigma/(\sigma-1)} \quad (4)$$

subject to the budget constraint

$$\sum_i p_{ij} X_{ij} = E_j \quad (5),$$

where σ is the elasticity of substitution, β_i is a positive distribution parameter, E_j is the nominal income of region j residents, and p_{ij} is the price of region i goods for region j consumers.

The expenditure shares for region i goods by region j consumers satisfying maximization of (4) subject to (5) are:⁵

$$\frac{X_{ij}}{E_j} = \left(\frac{\beta_i p_i t_{ij}}{P_j} \right)^{1-\sigma} \quad (6),$$

where p_i is their factory gate price, and $t_{ij} > 1$ is the trade cost factor between origin i and destination j . 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:

$$P_j = \left(\sum (\beta_i p_i t_{ij})^{1-\sigma} \right)^{1/(1-\sigma)} \quad (7).⁶$$

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 gross domestic product (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 (4) tells us that the omission of the importing nation's price index P_j from the original gravity equation described in equation (3) 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 (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

⁵ The shares are invariant to income, since preferences are homothetic.

⁶ For intermediate goods, the same logic works replacing expenditure shares with cost shares.

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:

$$E_i = \sum_j X_{ij} = p_i^{1-\sigma} \sum_j (\beta_i t_{ij})^{1-\sigma} \frac{E_j}{P_j^{1-\sigma}} \quad (8).$$

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

$$p_i^{1-\sigma} = \frac{E_i}{\Omega_i} \quad (9),$$

with:

$$\Omega_i = \sum_j (\beta_i t_{ij})^{1-\sigma} \frac{E_j}{P_j^{1-\sigma}} \quad (10),$$

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), remoteness (Baier and Bergstrand 2009) or with the well known term of *multilateral resistance*. Using equation (10) in equation (6) yields a basic but correctly specified gravity equation:

$$\frac{X_{ij}}{E_j} = (\beta_i t_{ij})^{1-\sigma} \frac{E_i}{P_j^{1-\sigma} \Omega_i} \quad (11).$$

Hence, 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 the difference between (11) and (6) gets larger as the asymmetry among countries is more pronounced (De Benedictis and Taglioni, 2011).

As shown by Baldwin and Taglioni (2006), Anderson and van Wincoop (2003) assume that $\Omega_i = P_i^{1-\sigma}$ under three critical assumptions. 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, the two terms Ω_i and $P_i^{1-\sigma}$ could 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 term per year. If this point is acknowledged, it is simple to see that the gravity model in equation (11) 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.⁷ Often however, the need of correcting for omitted price indices clashes with problems of collinearity with the other variables, and it has been shown that a full-blown fixed-effects structure may capture the policy effect of interest (Matyas, 1997). 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 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).

The main weakness of the first method is that the existing price indexes may not accurately reflect the true border effects (Feenstra, 2003). 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 datasets, 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} . In conclusion, time invariant pair effects on top of time-varying importer and exporter fixed effects to address pair-specific invariant omitted variables can be used, if appropriate and if their introduction does not generate problems of collinearity with other explanatory variables (De Benedictis and Taglioni, 2011).

⁷ Obviously, in cross-sections, the Anderson van Wincoop specification is sufficient owing to the lack of time dimension.

2.3 Aggregation issues

Aggregation is embedded in the gravity model, since the main insight from the theory is that bilateral trade depends on relative trade barriers, and this requires a comparison between bilateral trade costs and the average trade resistance between a country and its trading partners: the latter, as we know, is summarized by the multilateral resistance terms. Moreover, the use of a value added concept such as the GDP raises the issue of its relationship with gross trade flows since such a relationship may not be constant across products.

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. Even sectoral data are not at the level of detail of reality featuring thousands of tariff lines and corresponding (potential) trade flows, and it should not be forgotten that the latter aggregate exports decisions of several different firms. On the other hand, the standard model raises a geographical aggregation issue. As a matter of fact, results depend on the measure of trade costs within a region or country since a country or region is itself an aggregate. In both cases, we face an aggregation bias resulting from estimating trade costs with aggregated data when trade costs (and the elasticities of trade with respect to these costs) vary at the disaggregated level either in terms of sectors or regions.

Even if aggregation (a feature of almost all gravity investigations) biases gravity estimates of the impact of trade costs on bilateral trade flows, the gravity equation can also be used in reverse to measure bilateral trade costs: in this respect they can be considered part of the solution to the problem of aggregating trade barriers. The idea is to solve a theoretical gravity equation for the trade costs term instead of trade flows and to express these costs as a function of the observable trade data (UNCTAD/WTO, 2012). This allows to estimate the tariff equivalent of non-tariff barriers and such an approach is often preferred to alternative approaches based either on price differences across border or on direct measures of certain trade costs (Cipollina and Salvatici, 2008).

Anderson and Yotov (2010) provide an extensive discussion of aggregation bias in gravity estimation, setting out forces pushing in either direction, and concluding that no theoretical presumption can be created. On the contrary, the only mention of the geographical aggregation issue we are aware of is provided by Engel (2002) who criticizes the use of elasticities of substitution estimated without considering the number of countries involved. Even if little is known about the theoretical sign and magnitude of aggregation bias, and some degree of aggregation is inevitable, the (possibly obvious) recommendation is to disaggregate as much as possible (Anderson and van Wincoop, 2004).

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 the demand side, and Chaney (2008), from the supply side, have both introduced heterogeneity in gravity models, allowing for the more general derivation of gravity with heterogeneous firms. One remarkable feature of this gravity equation is that the elasticity of trade flows with respect to variable trade costs depends not on the elasticity of substitution between firm varieties but rather on the shape parameter of the Pareto distribution for productivity.

In practice, the extension to disaggregated goods 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. More generally, models including a large number of sectors quickly become unmanageable due to the number of parameters involved. Even if the number of observations exceeds the number of parameters, gravity models with large numbers of fixed effects and interaction terms can be slow to estimate, and may even prove impossible to estimate with some numerical methods such as Poisson and Heckman. A more feasible alternative in such cases is to estimate the model separately for each sector in the dataset. The fact that each sector represents a separate estimation sample allows for multilateral resistance and the elasticity of substitution to vary accordingly. Indeed, it can often be useful from a research point of view to estimate separate sectoral models: knowledge of differences in the sensitivity of trade with respect to policy in particular sectors can be important in designing reform programmes, for example. This approach is therefore frequently used in the literature (De Benedictis and Salvatici, 2011).

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, 2011). 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 later on.

Starting with the issue of *typology*, in the large majority of studies the dependent variable is a measure of bilateral merchandise trade. Three choices of trade flows 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,⁸ it is common practice to justify the use of this procedure using the above arguments.

⁸ 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.

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.⁹ 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 (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).¹⁰ 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 Lugov-

⁹ 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.

¹⁰ 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 used (De Benedictis and Taglioni, 2011).

skyy 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. 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 wipe 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, 2010):

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

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).¹¹ Nowadays, all most common distance measures across virtually all country pairs in the world are freely available online or can be obtained from the applets of the most important geo-representations available on the web.¹² Econometric estimates of the constant elasticity of trade to distance range within an interval of –

¹¹ 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: $\phi_{ij} = a \cos(\sin(\text{lat}_i) \cdot \sin(\text{lat}_j) + \cos(\text{lat}_i) \cdot \cos(\text{lat}_j) \cdot \cos(\text{long}_j - \text{long}_i)) \cdot R$; where $R=6371$ is the radius of the earth, in km.

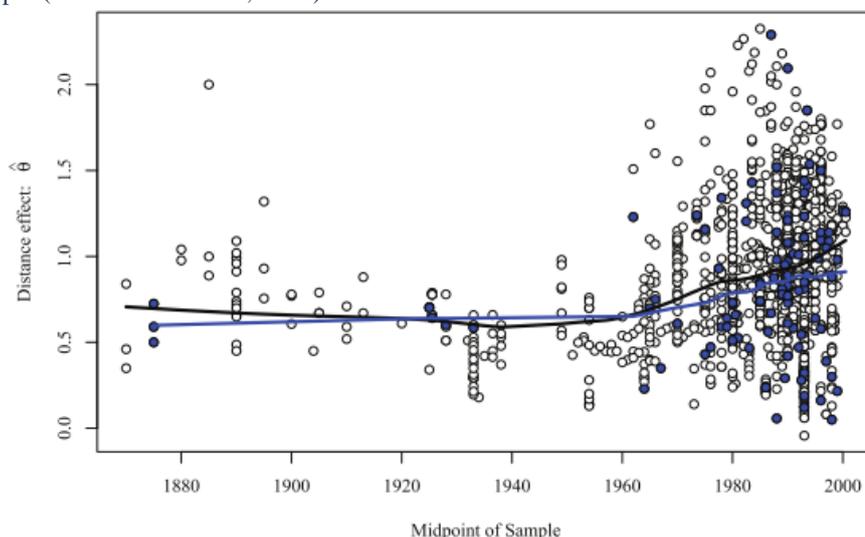
¹² CEPII generated a positive externality for all researchers by making freely available their measures of distance (see <http://www.cepii.fr/anglaisgraph/bdd/distances.htm>). Jon Haveman, Vernon Henderson and Andrew Rose were pioneers in this matter. Haveman’s collection of International Trade Data 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).

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.

As it was already mentioned, 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 as in the case of “iceberg costs”. 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).

If the distance variable is measured with error, we should expect the relevant coefficient to be biased. Disdier and Head (2008) examine 1467 distance effects estimated in 103 papers. Figure 1 distinguishes the highest R^2 estimate of each paper (shown with solid circle) and graphs all the estimates against time fitting time trends for both groups (i.e., highest R^2 and others)

Figure 1 - The variation of the distance effect graphed relative to the mid-period of the data sample (Disdier and Head, 2008)



Apparently, the estimated negative impact of distance on trade rose around the middle of the century and has remained persistently high since then. This result holds even after controlling for many important differences in samples and methods.

There is a general consensus that the estimated distance coefficient is higher than expected and the fact that it is highly persistent and also increasing over time is at odds with the evidence reported by Hummels and Lugovskyy (2006) of a decreasing pattern in freight costs.

Many have offered possible explanations for this “puzzling persistence of distance”. Felbermayr and Kohler (2006) argue that the distance puzzle may simply reflect a misspecification of the gravity equation that arises from inadequate treatment of the dual margin of world trade: as matter of fact, it may increase through time (intensive margin) as well as if a

trading bilateral relationship is newly established between countries that have not traded with each other in the past (extensive margin). While De Benedictis and Taglioni (2011) point out that 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.

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). Bernard et al. (2007 and 2009) shows that the extensive margin explains why trade falls off with distance. Lawless (2010) confirms that distance has a negative effect on both margins, but the magnitude of the effect is considerably larger and significant for the extensive margin.

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, 2010). 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 (2007) 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 at-

tractors 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.¹³

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.

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.

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 Bernoulli 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

¹³ See Anderson and van Wincoop (2004) for more discussion.

online¹⁴ and subsets are included in many datasets used and made available by experts in the field.¹⁵ 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, 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 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.¹⁶

¹⁴ 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

¹⁵ Andrew Rose’s homepage (<http://faculty.haas.berkeley.edu/arose/RecRes.htm>) is a good example of data sharing.

¹⁶ 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 cor-

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 (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 Persson (2001) and, showing that, in both cases, the relevant policy coefficient is substantially reduced.¹⁷ 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

related 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).

¹⁷ Propensity score is a matching technique that attempts to estimate the effect of a treatment, policy, or other intervention by accounting for the predicted probability of group membership – e.g., treatment vs. control group – based on observed predictors.

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 concerned, 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, 2011). An alternative economic specification explanation retains CES/Armington preferences and rationalizes zeros as due to fixed costs of export facing monopolistic competitive firms. Helpman et al. (2008) develop this idea.¹⁸

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 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 $y = \ln(1+M_{ij})$. Assuming that the problem is not of selection but truncation (censored data), this is the estimator to 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).¹⁹ 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. In other terms, there must be at least one variable which appears with a non-zero coefficient in the selection equation but does not appear in the equation of interest, the ‘exclusion restriction’. Such a restriction is crucially relevant, and if the variable included in the selection equation also affects

¹⁸ 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.

¹⁹ 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 outcome variable, it can lead to the researcher preferring simple OLS to the Heckman procedure (Puhani 2000). Helpman et al. (2008), for example, use as selection variables common religion or the regulation cost of firm's entry. 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 (3) 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 (ZIPML) or Zero-Inflated Negative Binomial Model (ZINBML). 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 ZIPML (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 ZIPML or a ZINBML 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. (2011), 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,²⁰ 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 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

²⁰ There is a difference between a good that is not produced and hence 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 at risk) to be exchanged because of non-economic reasons: trade embargos, religious prohibitions, etc..

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 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.

References

- Anderson, James E, (1979). "A Theoretical Foundation for the Gravity Equation," American Economic Review, American Economic Association, vol. 69(1), pages 106-116, March.
- Anderson JE (2011) The Gravity model. Annual Review of Economics, 3.
- Anderson JE, van Wincoop E (2003) Gravity with gravitas: A solution to the border puzzle. Amer. Econ. Rev. 63: 881-892.
- Anderson JE, van Wincoop E (2004) Trade costs. J of Econ Lit, 42:691-751.
- Anderson JE, Yotov YV (2010) The changing incidence of geography. Am Econ Rev, 100(5), 2157-2186.
- Anderson, JE (1979) A theoretical foundation for the gravity equation. Amer. Econ. Rev. 69: 106-116.
- Angrist JC, Pischke J-S (2008) Mostly harmless econometrics: an empiricist's companion. Princeton University Press, Princeton
- Arkolakis C, Costinot A, Andres Rodriguez-Clare A (2012) New Trade Models, Same Old Gains?, American Economic Review, American Economic Association, vol. 102(1), pages 94-130,
- Armington, PS (1969) A theory of demand for products distinguished by place of production, IMF Staff Papers, 16, pp. 159-178.
- Baier SL, Bergstrand JH (2001) The growth of world trade: tariffs, transport costs and income similarity. J. Int. Econ. 53: 1-27.
- Baier SL, Bergstrand JH (2004) Economic determinants of free trade agreements. J. Int. Econ. 64 (1): 29–63
- Baier SL, Bergstrand JH (2007) Do free trade agreements actually increase members' international trade? J. Int. Econ. 71: 72-95.
- Baier SL, Bergstrand JH (2009) Bonus vetus OLS: a simple method for approximating international trade-cost effects using the gravity equation. J. Int. Econ. 77: 77-85.
- Baldwin R, Taglioni D (2006) Gravity for dummies and dummies for gravity equations. NBER Working Paper N° 12516.
- Bergstrand JH (1985) The gravity equation in international trade: some microeconomic foundations and empirical evidence. Rev Econ Stat 67: 474-481.
- Bergstrand, JH (1989) The generalized gravity equation, monopolistic competition, and the factor-proportions theory in international trade. Rev. Econ. Statist. 71: 143-153.
- Bergstrand JH, Egger P. (2010), Gravity Equations and Economic Frictions in the World Economy: A Survey," University of Notre Dame Notre Dame Economics Working Paper Series, June.
- Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel S. Kortum, (2003) "Plants and Productivity in International Trade", American Economic Review, 93(4), 1268-1290
- Bernard A, Jensen JB, Redding S, Schott P (2007) Firms in international trade. J of Econ Persp, 21:105-130.
- Bernard, AB, Jensen, JB, Redding, SJ, Schott, PK, (2009) The Margins of US Trade, American Economic Review, American Economic Association, vol. 99(2), pages 487-93.
- Bernard, AB, Jensen, JB, Redding, SJ, Schott, PK, (2011) The Empirics of Firm Heterogeneity and International Trade, NBER Working Paper No. 17627.
- Blum BS, Goldfarb A (2006) Does the internet defy the law of gravity? J Int Econ 70(2):384-405.
- Brun J-F, Carrère C, Guillaumont P, de Melo J (2005) Has distance died? Evidence from a panel gravity model. World Bank Econ Rev 19:99-120.
- Bun M, Klaassen F (2002) The importance of dynamics in panel gravity models of trade. Tinbergen Institute Discussion Paper, No. 02–108/2.
- Burger MJ, van Oort FG, Linders GM (2009) On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation. Spat Econ Anal 4(2):167-190.
- Cameron AC, Trivedi PK (2005) Microeconometrics. Methods and Applications. Cambridge University Press, Cambridge.
- Cardamone P (2007) A survey of the assessments of the effectiveness of Preferential Trade Agreements using gravity models. International Economics 60(4):421-473.
- Chaney T. (2011) The Network Structure of International Trade, NBER Working Papers 16753, National Bureau of Economic Research.
- Chaney T (2008) Distorted gravity: the intensive and extensive margins of International trade. Am Econ Rev 98:1701-1721.
- Cipollina M, Salvatici L (2008) Measuring Protection: Mission Impossible?. Journal of Economic Surveys, Vol. 22, Issue 3, pp. 577-616
- Cipollina M, Salvatici L (2010a) The impact of European Union agricultural preferences. J of Econ Policy Review 13:87-106.

- Cipollina M, Salvatici L (2010b) Reciprocal trade agreements in gravity models: A Meta-Analysis. *Rev of International Econ* 18:63-80
- Cipollina M, Laborde D., Salvatici L (2011) Do preferential Trade Policies (Actually) Increase Exports? A Comparison between EU and US trade policies, SIE - Società Italiana degli Economisti, 52.ma Riunione Scientifica Annuale, Roma.
- Costantini J, Melitz MJ (2008) The dynamics of firm-level adjustment to trade liberalization. In: Helpman E, Marin D, Verdier T (eds) *The organization of firms in a global economy*. Harvard University Press, Boston, MA
- Daude C, Stein E (2007) Longitude matters: time zones and the location of foreign direct investment. *J Int Econ* 71:96-112.
- De Benedictis L, De Santis R, Vicarelli C (2005) Hub-and-Spoke or else? Free Trade Agreements in the enlarged EU. *European Journal of Comparative Econ* 2:245-260.
- De Benedictis L and Salvatici L eds. (2011), *The Trade Impact of European Union Preferential Policies: an analysis through gravity models*, Springer.
- De Benedictis L, Vicarelli C (2005) Trade potentials in gravity panel data models. *Topics in Econ Anal & Policy* 5(1):1-31
- De Benedictis L, Vicarelli C (2009) Dummies for gravity and gravity for policies: mission impossible? Mimeo.
- De Benedictis L, Taglioni D (2011) The Gravity Model in International Trade, in “The Trade Impact of European Union Preferential Policies: an analysis through gravity models”, Luca De Benedictis and Luca Salvatici (eds.), Springer.
- Disdier AC, Head K (2008) The puzzling persistence of the distance effect on bilateral trade. *R E Stat* 90(1):37-48.
- Eaton B, Kortum S (2002) Technology, geography and trade. *Econometrica* 70(5):1741-1779.
- Engel C. (2002), Comment on Anderson and van Wincoop in *Brookings Trade Forum 2001*, Susan Collins and Dani Rodrik, eds., Washington: The Brookings Institution.
- Feenstra RC (2003) *Advanced international trade: theory and evidence*. Princeton University Press, Princeton
- Felbermayr GJ, Kohler W (2006) Exploring the intensive and extensive margins of world trade. *Rev World Econ* 142(4):642-674.
- Fidrmuc J (2009) Gravity models in integrated panels. *Empir Econ* 37:435-446
- Francois J, Hoekman B, Manchin M (2006) Preference erosion and multilateral trade liberalization. *The World Bank Econ Rev* 20:197-216
- Frankel J (1997) *Regional trading blocs*. Institute for International Economics, Washington, DC
- Guiso L, Sapienza P, Zingales L (2009) Cultural biases in economic exchange? *The Quart J of Econ* 124(3):1095-1131.
- Harrigan J (2010) Airplanes and comparative advantage. *J Int Econ* 82(2):181-194.
- Head K, Mayer T (2000) Non-Europe: the magnitude and causes of market fragmentation in the EU. *Weltwirtschaftliches Archiv* 136 (2): 285– 314.
- Head K, Mayer T (2004) Market potential and the location of Japanese investment in the European Union. *R E Stat*.
- Heckman J (1979) Sample selection bias as a specification error. *Econometrica* 47(1):153-161.
- Helpman E, Krugman P (1985) *Market structure and foreign trade: increasing returns, imperfect competition and the international economy*. MIT Press, Cambridge
- Helpman E, Melitz M, Rubinstein Y (2008) Estimating trade flows: trading partners and trading volumes. *Quart J Econ* 123(2):441-487.
- Hummels D (1999) *Toward a geography of trade costs*. GTAP Working Paper 17, Global Trade Analysis Project, Purdue University.
- Hummels D (2007) Transportation Costs and International Trade in the Second Era of Globalization, *Journal of Economic Perspectives* 21 pp. 131-154.
- Hummels D, Lugovskyy V (2006) Are matched partner trade statistics a usable measure of transport costs? *Rev of Int Econ* 14(1):69-86.
- Imbens G, Wooldridge J (2009) Recent developments in the econometrics of program evaluation. *J of Econ Lit* 47(1): 5–86
- Jacks D, Meissner C, Novy D (2008) Trade costs. 1870-2000. *Am Econ Rev* 98(2):529-534.
- Lawless M (2010) Deconstructing gravity: trade costs and extensive and intensive margins. *Can J Econ* 43(4):1149-1172.
- Leamer EE, Levinsohn J (1995) International trade theory: the evidence. In: Grossman G, Rogoff K (eds) *Handbook of International Economics*, Volume 3. Elsevier, North-Holland
- Limao N and Venables T (2001) Infrastructure, geographical disadvantage and transport costs. *World Bank Economic Review* 15:451-479.

- Liu X. (2009) GATT/WTO promotes trade strongly: sample selection and model specification, *Review of International Economics* 17(3), 428-446.
- Markusen, J.R. (2010). Putting Per-Capita Income Back into Trade Theory, NBER Working Papers 15903, National Bureau of Economic Research, Inc
- Martínez-Zarzoso I, Nowak-Lehmann DF, Horsewood N (2009) Are regional trading agreements beneficial? Static and dynamic panel gravity models. *N Am J Econ Finance* 20(1):46-65
- Matyas L, (1997) Proper Econometric Specification of the Gravity Model, *The World Economy*, 20, pp. 363-369.
- Matyas L, Sevestre P (2007) *Econometrics of panel data*. Kluwer Academic Publishers
- Melitz MJ (2003) The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 1695-1725.
- Millimet DL, Tchernis R (2009) On the specification of propensity scores, with applications to the analysis of trade policies. *J Bus Econ Stat* 27(3)397-415.
- Persson T (2001) Currency Union and trade: how large is the treatment effect. *Economic Policy* 33:433-462.
- Portes R, Rey H (2005) The determinants of cross-border equity flows, *Journal of International Economics*, Elsevier, vol. 65(2), pages 269-296
- Puhani PA (2000) The Heckman Correction for Sample Selection and Its Critique. *J of Econ Survey* 14(1):53-69.
- Rose AK, van Wincoop E (2001). National Money as a Barrier to International Trade: The Real Case for Currency Union, *American Economic Review*, American Economic Association, vol. 91(2), pages 386-390.
- Santos-Silva JMC, Tenreyro S (2006) The log of gravity. *Rev Econ Statist* 88:641-658.
- Samuelson P. (1952), The Transfer Problem and the Transport Costs: The Terms of Trade When Impediments Are Absent, *Economic Journal*.
- Tinbergen J (1962), *Shaping the World Economy*, The Twentieth Century Fund, New York.
- UNCTAD (2003) Trade preferences for LDC: an early assessment of benefits and possible improvements. UNCTAD/ITCD/TSB/2003/8, United Nations, Geneva.
- WTO/UNCTAD (2012), *A Practical Guide to Trade Policy Analysis*, Geneva.