ARMINGTON ELASTICITIES AND THE DISTANCE PUZZLE

Elizaveta Archanskaia (Sciences Po (OFCE))
Guillaume Daudin (Lille-I (EQUIPPE) / Sciences Po (OFCE))

Introduction

The estimated effect of distance in gravity equations has remained stable, or even slightly increased, in the past 50 years\(^1\). This has been called a puzzle because of the common opinion that technological developments in transports (e.g. the airplane and the container) had contributed to the flattening of the world, leading to the death of distance\(^2\).

Recent work on the distance puzzle has gone in two directions. Following Santos Silva and Tenreyro, one strand of the literature has argued that the log-linear estimation strategy resulted in biased distance coefficients, and has sought to provide consistent estimates of the distance elasticity of trade since the 1950s using non linear estimators\(^3\). Indeed, as a consequence of Jensen’s inequality\(^4\), the log-linear estimation provides inconsistent estimates if the trade equation in levels is subject to heteroskedasticity linked to explanatory variables such as trading partners’ characteristics. Furthermore, the log-linear estimation strategy suffers from sample selection bias because it can neither take into account the abundance of zero flows in trade statistics, nor the reduction in the proportion of zeros overtime\(^5\). The growth of trade has been both intensive, in the sense that the volume of trade of established

---

\(^1\) Disdier and Head (2008) find that distance impedes trade by 37% more since 1990 than it did from 1870 to 1969 in a metaanalysis based on 1467 estimations of the distance coefficient in gravity equations spanning the period 1870-2001. Berthelon and Freund (2008) find that the distance coefficient has increased by 10% in absolute value over the period 1985-2005. See also Coe, Subramanian, and Tamirisa* (2007); Combes, Mayer, and Thisse (2006); Brun, Carrere, Guillaumont, and Melo (2005) and Buch, Kleinert, and Toubal (2004).


\(^3\) Silva and Tenreyro (2006); Coe, Subramanian, and Tamirisa* (2007); Bosquet and Boulhol (2009).

\(^4\) \(E(\ln x) \neq \ln E(x)\). See Silva and Tenreyro (2006) and Bosquet and Boulhol (2009) for a discussion.

\(^5\) Felbermayer and Kohler (2006) argue that the initial estimates of the distance coefficient were biased downwards because of a greater presence of zeros, and that this bias was subsequently reduced thanks to the extensive margin of trade.
trade relations had increased, and extensive, in the sense that new trade relations had been established\textsuperscript{6}. Taking zeros into account might change the evolution of the distance coefficient if new trade relations have in priority been established between distant partners. The use of a non-linear estimator leads to a rejection of the increase in the distance elasticity of trade overtime. However, once the potential biases of the log linear estimation strategy have been corrected, the distance coefficient is still found to be \textit{non decreasing} overtime\textsuperscript{7}.

The second strand of the literature has looked into theoretical and empirical explanations of a \textit{non decreasing} distance elasticity of trade overtime, arguing that it is not certain there is any puzzle\textsuperscript{8}. First, because the evolution of the level of trade costs relative to the price of traded goods is still debated\textsuperscript{9}. Even if the general level of trade costs has an income effect on openness to trade of each country, this is accounted for in the estimation of the theoretical gravity model by country-specific fixed effects. More fundamentally, the value of the distance coefficient in theoretically derived gravity models of trade is not directly related to the level of trade costs\textsuperscript{10}. It is the product of two elasticities: the elasticity of trade flows relative to trade costs and the elasticity of trade costs relative to distance. Whether transport costs have declined or increased, the crucial question for the evolution of the coefficient is whether non-distance dependent transport costs (e.g. loading costs at ports) have declined relatively to distance-dependent transport costs (e.g. fuel costs)\textsuperscript{11}.

\textsuperscript{6}Helpman, Melitz, and Rubinstein (2008), Baldwin and Harrigan (2007).
\textsuperscript{7}Using the PPML estimator, Bosquet and Boulhol (2009) find that the distance elasticity of trade has been stable within the .6-.75 range in 1948-2006 while the gap between the PPML elasticity and the elasticity estimated with the log-linear specification had increased overtime. Coe, Subramanian, and Tamirisa\textsuperscript{9}(2007) find that the distance elasticity had decreased from 0.5 to 0.3 in 1975-2000 using NLS. In their estimation, the sum of exports and imports is used as the dependent variable. Note also that Brun, Carrere, Guillaumont, and Melo (2005) and Márquez-Ramos, Martínez-Zarzoso, and Suárez-Burguet (2007) find decreasing distance elasticities once the sample is restricted to developed countries using a log-linear estimation strategy.
\textsuperscript{8}See Bosquet and Boulhol (2009); Duranton (2008); Buch, Kleinert, and Toubal (2004).
\textsuperscript{10}The three specifications we consider in this paper are Eaton and Kortum (2002), Anderson and Wincoop (2003) and Chaney (2008).
\textsuperscript{11}See Feyrer (2009).
Second, a composition effect might explain a non decreasing distance coefficient. It might be that the composition of traded goods has changed toward goods which are either less transportable or which consumption is more sensible to trade costs. Theoretically, Duranton and Storper show how falling transport costs can induce firms in trading goods with higher transaction costs, leading to an increasing distance sensitivity of trade.\textsuperscript{12} Empirically, Berthelon and Freund test the impact of the changing composition of world trade on the distance coefficient between 1985 and 2005 but find that it has had a negligible effect\textsuperscript{13}.

The discussion on the distance puzzle in this second strand of the literature has thus far been mainly concerned with the evolution of the elasticity of trade costs relative to distance. This is forgetting half of the picture. One should look as well at the elasticity of trade flows relative to trade costs. For example, since the 1960s, an increasing number of countries have demonstrated their ability to climb up the quality ladder and produce a set of goods bearing the same brands as the goods produced in developed countries. This may have resulted in more uniformity in country-specific varieties of each good. This could have increased the elasticity of substitution between varieties and hence the elasticity of trade flows relative to trade costs, giving an additional explanation for the distance puzzle\textsuperscript{14}.

This paper’s original contribution is to assess the role of the evolution of the Armington elasticity of substitution between goods produced in different countries in the distance puzzle. First, it evaluates the role of the composition effect on the distance coefficient using data for 1962-2006 rather than 1985-2005, thus covering a longer time period than Berthelon and

\textsuperscript{12} See Duranton (2008).
\textsuperscript{13} Berthelon and Freund (2008) find that 39% of products and 54% of trade as become more sensitive to distance over this period against 2.8% of trade becoming less sensitive to distance. Hummels (2001) suggests another explanation: the increasing importance of certain components of trade costs, such as trading on time, for most traded goods.
\textsuperscript{14} The elasticity of substitution between varieties of different origin could play a role in explaining the distance puzzle according to the Anderson and Wincoop (2003) derivation of the gravity model. If Eaton and Kortum (2002) were the relevant framework, it could be argued that countries’ technical capabilities have become less dispersed overtime. See section 2 for a discussion.
Freund. Second, it uses a Poisson estimate to correct for potential biases of the log-linear estimation. Third, this paper innovates relatively to existing literature in that it assesses the ability of the three most common specifications of the theoretically grounded gravity model of trade\textsuperscript{15} to explain the distance puzzle. Fourth, it measures elasticities of substitution between goods from different countries directly, using unit values as a proxy for prices.

The first part studies the composition effect. Since the 1960s, the composition of world trade has evolved towards an ever-growing proportion of manufactured goods’ flows in total trade. Manufactured goods are more differentiated than primary goods. Since within-product elasticities of substitution depend on the degree of varieties’ differentiation within a given product sector, manufactured goods should have a lower within-product elasticity of substitution than homogeneous goods\textsuperscript{16}. This entails that trade in these goods should be relatively less sensitive to trade barriers.

We use the Poisson estimator in a theoretically grounded gravity setting, and work with a COMTRADE world bilateral trade dataset spanning the years 1962-2006 and reporting trade at the 4-digit SITC Rev. 1 disaggregation level in order to quantify the impact of the changing composition of trade on the aggregate distance coefficient, assuming that the Armington elasticity of substitution for each good at the 4-digit level has stayed constant. In line with the growing importance of differentiated goods in world trade, we find that the composition effect would entail a decreasing elasticity of trade to distance of approximately 20\% of the initial value. The composition effect thus only deepens the puzzle.

In a second part, we check whether firm level product differentiation\textsuperscript{17} or Ricardian models of comparative advantage\textsuperscript{18} help explain the distance puzzle. Specifically, we discuss

\textsuperscript{15} These are Eaton and Kortum (2002), Anderson and Wincoop (2003) and Chaney (2008).
\textsuperscript{16} Rauch (1999).
\textsuperscript{17} This builds on the work of Melitz (2003) and Chaney (2008).
theoretical and empirical reasons which could have led to a change in the parameter corresponding to the elasticity of trade flows relative to trade costs in the Eaton and Kortum framework, and in the Chaney framework: the dispersion of each country’s technical capabilities and firm productivities, respectively.

In a third part, we discuss theoretical and empirical grounds which indicate that the evolution of sector-specific Armington elasticities of substitution since the 1960s might be relevant to explain the evolution of the distance coefficient. From the theoretical standpoint, a number of models suggest we must move away from the CES assumption to understand the evolution of world trade. From the empirical standpoint, Armington elasticities might have increased as an increasing number of countries started producing a set of goods bearing identical brands. This paper collects estimates of Armington elasticities in 1950-2000 and suggests its own estimates based on the use of unit values. It then assesses whether the evolution of Armington elasticities explains part of the distance puzzle.

\[18\] This builds on the work of Davis (1995) and Eaton and Kortum (2002).
1. **Across sectors’ composition effect**

Berthelon and Freund (2008) have studied the impact of a potential change in the composition of trade on the distance coefficient from the following angle: if traded goods have variable transportability, it is possible that the decrease in trade costs brought about an increase in the share of traded goods which have a high sensitivity to trade costs, resulting in an overall increased sensitivity of trade to trade costs. They find that this composition effect is actually nil.

The impact of the composition effect on the distance coefficient can also be studied from the angle of the degree of product differentiation. Product differentiation varies across sectors. Manufactured products should have a lesser sensitivity to trade barriers than homogeneous commodities because the degree of feasible product differentiation is higher for manufactured goods\textsuperscript{19}. This means that the elasticity of substitution for manufactured goods is relatively low. Indeed, Broda and Weinstein’ computations of substitution elasticities for American imports in 1972-2001 show that trade in homogeneous goods is approximately three times more sensitive to a change in relative prices than trade in differentiated goods\textsuperscript{20}. Thus, if manufactured goods’ share in overall trade increases, the elasticity of substitution should decrease at the aggregate level, leading to a decrease of the distance coefficient.

If both the composition and the substitution effects were found to deepen the distance puzzle, a *non decreasing* distance coefficient would have to be explained by an increasing sensitivity of trade to trade costs and/or a change in the structure of trade costs, e.g. in the

---

\textsuperscript{19} Rauch (1999)  
\textsuperscript{20} Broda and David (2006).
relative shares of fixed and variable costs. In this section, we investigate the role of the composition effect on a longer time period than Berthelon and Freund (2008) while correcting for potential biases of the log linear estimation strategy used by these authors.

1.1. Data and methodology

Our baseline database is a world bilateral trade dataset spanning the years 1962-2006 and reporting trade at the 4-digit SITC Rev.1 disaggregation level comprising 228 trading partners extracted from COMTRADE. The SITC Rev.1 classification is chosen because it provides the longest coverage of bilateral trade flows at the same disaggregation level. As trade data is of better quality for imports, the estimation is conducted only on data reported as imports to COMTRADE.

We verify the robustness of our results by using an alternative dataset spanning the years 1962-2000 and reporting trade at the 4-digit SITC Rev.2 disaggregation level for 184 trading partners. This dataset is provided by Feenstra, Lipsey, Deng, Ma and Mo.

We follow the Anderson and van Wincoop derivation of the gravity trade model. These authors use the Armington assumption of goods differentiated by country of origin, with each country producing only one good. The supply of each good is assumed fixed. Trade costs are assumed symmetric for each trading country pair. Approximating identical homothetic preferences of consumers in all countries by a CES utility function, Anderson and van Wincoop derive the following gravity equation for aggregate trade flows:

\[ X_{ij} = \frac{Y_i Y_j}{Y} \left( \frac{t_{ij}}{\Pi_i \Pi_j} \right)^{1-\sigma} \]

21 Felbermayer and Kohler (2006) have suggested a different solution for the distance puzzle: they argue that early estimates of the distance coefficient were biased downwards. A reduced number of country pairs traded, and OLS yielded inconsistent estimates of the distance coefficient.


24 Balistreri and Markusen (2009) show this is equivalent to assuming perfect competition at the subnational firm, e.g. an infinite elasticity of substitution between domestic firms’ varieties.
Here $X_{ij}$ is the value of exports from country $i$ to country $j$, $Y_{ij}$ is nominal income of each trading partner, $Y$ is world income, $t_{ij}$ are bilateral trade costs, $\cdot_i$ and $P_j$ are respectively inward and outward multilateral trade resistance terms, and $\cdot$ is the Armington elasticity of substitution.\(^{25}\)

Anderson and van Wincoop show that for a given specification of bilateral trade costs, the log linear estimation in which exporter and importer fixed effects are used as proxies for multilateral trade resistance terms provides consistent estimates of this gravity equation. However, Santos Silva and Tenreyro have shown that inference based on estimates from the log linear estimation is problematic if the trade equation in levels is subject to heteroskedasticity.\(^{26}\) These authors show that the Poisson pseudo maximum likelihood estimator is both a consistent and efficient estimator of the gravity equation under the hypothesis that the conditional variance of trade flows is proportional to their conditional expected value.\(^{27}\) This estimation strategy has the added advantage of integrating observations signalling no trade in the sample.

Following Santos Silva and Tenreyro, we use the PPML estimator in order to quantify the impact of the changing composition of trade on the aggregate distance coefficient, assuming that the Armington elasticity of substitution for each good at the 4-digit level has stayed constant. The estimated equation in cross section is:

$$ M_{ij} = \exp(\alpha_0 + \alpha_1 \ln d_{ij} + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + f_{exp} + f_{imp}) e_{ij} $$

where $M_{ij}$ is the value of imports from country $i$ to country $j$, $d_{ij}$ is bilateral distance, $f_{exp}$ and $f_{imp}$ are importer and exporter fixed effects, $Z_1$ comprises bilateral trade costs’ controls taken from the CEPII database,\(^{28}\) e.g. adjacency, common language, same country, and

\(^{25}\) See Anderson and Wincoop (2003).
\(^{26}\) Silva and Tenreyro (2006).
\(^{27}\) Bosquet and Boulhol (2009) also provide support for the PPML estimator.
\(^{28}\) This data is available online at www.cepii.fr.
colonial linkages, $Z_2$ comprises additional trade costs’ controls constructed on the basis of documentation provided by the WTO\(^{29}\), e.g. common membership of GATT/WTO and common membership of an FTA, while $Z_3$ is a proxy for a common level of development of the two trading partners\(^{30}\), and $\varepsilon_\ell$ is a multiplicative error term.

We proceed as follows for computing the composition effect. The distance coefficient is first estimated for each year of the sample for aggregate trade. This allows checking whether Bosquet and Boulhol’s finding that the distance elasticity has been stable within the 0.6-0.7 range when controlling for common membership of an FTA holds for our dataset\(^{31}\) Second, we estimate the distance coefficient for aggregate trade each year of the sample while fixing each product’s share at its share in a given base year\(^{32}\). This gives the distance elasticity of trade if the composition of trade had stayed unchanged at base year weights in all years. We then compute the moving average of the two point estimates and evaluate the gap between the two estimates over the sample\(^{33}\)

1.2. Results: the composition effect deepens the distance puzzle

The evolution of the distance coefficient is shown in Figure 1. Our baseline regression ("without any FTA") only includes distance, colonial links, contiguity and common languages. It shows an increase in the absolute value of the elasticity of trade to distance from the 0.55-0.6 range to the 0.67-0.72 range. The absolute value of the elasticity increases from 1962 to 1985, decreases from 1985 to 1993 and increases afterward. The turning point of 1985-1986 might be linked to the sharp decrease of oil prices at that time.\(^{34}\)

______________________________

\(^{29}\) Data for membership of GATT/WTO was taken from the WTO website. Data for common membership of an FTA was constructed by the authors from Crawford and Fiorentino (2005).

\(^{30}\) This variable is constructed on the basis of country groupings in high, upper middle, lower middle, and low income per capita made available on the World Bank website. The year 2000 classification is used for all years.

\(^{31}\) Bosquet and Boulhol (2009).

\(^{32}\) The following base years have been used: 1962, 1984, 2000, 2006.

\(^{33}\) We computed 3-year, 5-year, and 10-year moving averages.

\(^{34}\) The confidence intervals are not represented in the figure: the 95% ones are equal to +/- 0.06 to 0.1.
The 1985-2006 part of the story is not robust to the introduction of dummies for all FTAs and comparable levels of development. Once they are introduced (“all FTA”), the absolute value of the elasticity of trade to distance decreases sharply from 1985 to 1986. It then stabilizes at a low level, comparable to the one in the early 1960s. One might conclude that the distance puzzle is hence solved. Of course, this might be misleading, as it might be the case that FTAs are endogenous to the intensity of trade. Indeed, in that specification, common membership to the EEC is associated to lower trade.

A specification introducing only NAFTA, the EEC, the ASEAN and the GATT as dummies (“FTA (-)”) gives more plausible values to the EEC coefficients, but does not modify dramatically the picture changed by the introduction of all FTAs. 1985-1986 is still a breaking point; the elasticity coefficient stabilizes at an absolute level slightly higher than in the early 1960s. Both variations suggest that increased regionalization of trade plays an important role in the apparent higher role for distance.

35 The full list of included FTAs and of the years these appear in the estimation is provided in the Appendix.
This evolution might be linked to changes in the list of countries reporting their trade. The first way to check that is to restrict the analysis to the panel of countries importing every year from 1962 to 2006 and the panel of countries exporting every year from 1962 to 2006. Figure 2 shows this makes a difference for the 1990s and 2000s. Following the fall of the USSR, many countries were created during this period. They apparently had a tendency to trade relatively less with distant partners. Removing them from the panel reduces the absolute value of the elasticity of trade to distance to less than in the 1960s.
The results are robust to using Feenstra data rather than COMTRADE data (see Figure 3).
Finally, we study the impact of the composition effect. Figure 4 shows that the evolution of the elasticity of trade to distance based on 1962 weights differs from the evolution based on unweighted data in two respects. First, the breaking point of 1985-86 disappears. Second, even when some FTAs are introduced, the distance puzzle does not disappear: the absolute value of the elasticity of trade to distance is 25% larger in the mid 2000s than in the early 1960s. This is in line with the growing importance of differentiated goods in world trade. The composition effect thus only deepens the distance puzzle.

Figure 4

This result leads us to investigate two additional explanations of a common evolution effect in a majority of traded goods which would have led to a non-decreasing distance coefficient. On the supply side (section 2), we investigate the possibility of a change in firms’ productivity heterogeneity in most sectors as a consequence of trade integration. We also
evaluate the ability of the Ricardian framework to help explain the distance puzzle\textsuperscript{36}. On the demand side (\textit{section 3}), we investigate the implications of a potentially endogenously increasing elasticity of substitution in the number of varieties as a consequence of trade integration as suggested by Krugman (1979) and Neary (2009)\textsuperscript{37}. In the Armington framework, this would entail increasing sector-specific Armington elasticities of substitution in the number of country-specific varieties traded. We check the validity of this hypothesis on historical data on Armington substitution elasticities and on our own sample.

2. \textbf{Interpreting the distance coefficient}

In their specification of the gravity model of trade, Anderson and van Wincoop do not model explicitly the supply side of the economy. Each country is specialized in the production of one good, and supply is perfectly inelastic\textsuperscript{38}. Implicitly, this replicates the supply framework used in CGE modeling, e.g. constant returns to scale in production and perfect competition between firms at the sub-national level.\textsuperscript{39} The mechanism which drives the model is a preference mechanism in consumption. As consumers’ preferences are approximated by a CES utility function, they value variety for itself because an increase in the number of varieties consumed mechanically reduces the aggregate price index.\textsuperscript{40} Furthermore, following Armington, it is assumed that consumers perceive products of different national origin as intrinsically imperfect substitutes.\textsuperscript{41} As noted by Eaton and Kortum\textsuperscript{42}, this means that higher production and/or trade costs leave the set of traded goods unaffected, but decrease

\textsuperscript{36} Davis (1995) and Eaton and Kortum (2002).
\textsuperscript{37} An appendix to Melitz (2003) considers the implication of an endogenously increasing substitution elasticity in the framework of heterogeneous firms’ models.
\textsuperscript{38} Anderson and van Wincoop (2003).
\textsuperscript{39} See Balistreri and Markusen (2009).
\textsuperscript{40} Krugman (1980).
\textsuperscript{41} Armington (1969).
\textsuperscript{42} Eaton and Kortum (2002).
expenditure on each imported good. Consequently, in the context of trade integration, the adjustment is restricted to the *intensive* margin, e.g. the volume of trade.

The crucial parameter which measures the elasticity of trade to trade costs is the elasticity of substitution between these country-specific varieties, the “Armington elasticity of substitution”. The more substitutable country-specific varieties are, the higher the sensitivity of trade to trade costs. Consequently, the Armington substitution elasticity has a direct impact on the distance coefficient. We investigate the implications of the hypothesis of a changing Armington elasticity in the following section. In this section, we pick up the criticism of the Armington assumption formulated by Davis. This author notes that this assumption is *ad hoc*. Indeed, absent modeling of the supply side, it is not clear why a given variety would have to be produced in only one specific country and why consumers would perceive these national varieties as imperfect substitutes.

This criticism leads us to investigate two theoretical derivations of the gravity model which differ from Anderson and van Wincoop in that they explicitly model the structure of production in the world economy while maintaining the CES utility function in modeling consumer behavior. These are the monopolistic competition framework with firm heterogeneity and the Ricardian framework. Both models introduce the mechanism of trade flows’ adjustment through the *extensive* margin in the context of trade integration. In these models, the parameter which corresponds to the elasticity of trade flows to distance is no longer the Armington elasticity of substitution, but firms’ productivity dispersion and technical efficiencies’ variability across goods respectively. In this section, we ask whether

43 See Davis (1995).
44 Firm heterogeneity models are subject to this criticism for firm-level varieties: in these models, consumers perceive all firm varieties as imperfect substitutes, and it is not clear why each variety is specific to a given firm-country combination. Note that in Anderson and van Wincoop (2003), the assumption of nationally differentiated varieties is less problematic in that the equation is estimated for aggregate trade flows: it is indeed plausible that the *composite* national good be country-specific.
models in which the adjustment takes place at both the extensive and the intensive margin help explain the distance puzzle.

2.1. Firm heterogeneity in monopolistic competition models and the distance puzzle

2.1.1. In theory

Building on Melitz (2003), Chaney has derived a theory-based gravity model of trade in the framework of monopolistic competition which allows for domestic firms’ heterogeneity in production efficiency.\(^{45}\) Melitz (2003) relaxed the assumption of representative firms which characterized imperfect competition models of international trade developed in the late 1970s and the 1980s\(^{46}\) in order to understand the impact of trade integration on resources’ reallocation across firms.

Within a given sector, Melitz (2003) assumes that goods are horizontally differentiated varieties at the firm level. The Melitz model, followed by Chaney, modifies the traditional Armington framework in that it assumes a finite constant elasticity of substitution between firm-level varieties and no country-specific product differentiation whereas the Armington assumption is based on the hypothesis of perfect competition between domestic firms and imperfect substitution between goods produced by different countries\(^{47}\).

In the context of trade integration, the Melitz model predicts that the elasticity of substitution between firm-level varieties determines the relative weight of the intensive and the extensive margins in the total increase in the volume of trade following a decrease in variable trade costs, i.e. the increase in the volume of already exporting firms relatively to volumes exported by firms acquiring exporter status. The higher is the elasticity of

\(^{45}\) Chaney (2008).
\(^{47}\) Armington (1969).
substitution, the higher is the share of the intensive margin in the total increase in the volume of trade.

Chaney’s main result is to show that in the context of a change in variable trade costs, the total change in the volume of trade would be *insensitive* to the level of the elasticity of substitution between varieties because this elasticity of substitution has an inversely proportional effect on the extensive and the intensive margins of trade. It follows that the elasticity of trade to distance is invariant to differing values of the firm-level elasticity of substitution. Rather, it is the inverse of firms’ productivity distribution log standard deviation, a parameter exogenously given in the Chaney model, which has a direct impact on the distance coefficient. If firm level product differentiation were the relevant framework, Chaney shows that it is the evolution of the degree of firms’ productivity heterogeneity which would have a role in explaining the value of the distance coefficient: the lower the degree of firm heterogeneity in domestic markets, the higher the distance coefficient.48

It should be borne in mind that *by definition* the elasticity of substitution in Chaney’s model is *not* the Armington elasticity of substitution49.

Before focusing on the parameter measuring the inverse of the degree of firms’ productivity heterogeneity in the Chaney model, we briefly go over the main features of Melitz (2003). A firm’s marginal cost depends on its productivity draw. The firm charges a fixed mark-up over its marginal cost because consumers are characterized by a constant elasticity of substitution between firm-level varieties. Consumers have an intrinsic love for variety which leads them to consume all of the existing varieties. The CES assumption results in a price distribution of varieties produced in any given sector, and this price distribution

______________________________

48 Chaney (2008).
49 See Balistreri and Markusen (2009) for a discussion of nested firm-level and national-level CES utility.
replicates the productivity distribution of viable firms. The more productive a firm is, the lower the price it charges, and the higher its market share. Entry in a given market hinges on the firm’s ability to cover a fixed entry cost. The zero profit cut-off condition combined with free entry determines the cut-off productivity level of viable firms as a function of the level of fixed costs and of the firm-level elasticity of substitution. Trade integration results in an increasing number of domestic firms which acquire export status. This shifts up the zero-profit condition which leads to an upward shift in the equilibrium cut-off productivity level of viable firms. The least productive firms exit the market while market shares and profits are redistributed towards more efficient firms.

We now turn to two specificities of Chaney (2008). The author assumes a Pareto distribution of firms’ productivity. The Pareto distribution parameter captures the inverse of firms’ productivity heterogeneity and is increasing in the fraction of small size low-productivity firms. Furthermore, Chaney does not impose the free entry condition. This implies that the cut-off productivity level for the viability of domestic firms does not necessarily increase in the context of trade integration, contrary to the original Melitz framework: the parameter measuring the inverse of firms’ productivity dispersion is exogenously given.

However, if we want to examine the impact of a potential change in this parameter on the elasticity of trade flows to distance, we have to look for a mechanism which renders endogenous the inverse of firms’ productivity dispersion. This is why we turn back to the original Melitz (2003) framework in which firms’ productivity heterogeneity is assumed endogenous. It follows that in the context of trade integration, low productivity firms exit the market.\textsuperscript{50} This leads to a decrease in the fraction of small size low-productivity firms, e.g. a an

\textsuperscript{50} Melitz (2003).
increased dispersion of firms’ productivities. The Pareto distribution parameter which measures the inverse of this dispersion decreases.\textsuperscript{51}

Thus, if the framework suggested by Chaney (2008) were reinterpreted in terms of the original Melitz (2003) model, the parameter which captures the inverse of firms’ productivity dispersion would be expected to decrease in the context of trade integration. This would lead to a \textit{decreasing} distance coefficient, deepening the distance puzzle.\textsuperscript{52, 53}

\subsection*{2.1.2. Empirical evidence}

In heterogeneous firms’ models, each firm charges a fixed mark-up over its marginal cost. This marginal cost is determined by the firm’s productivity draw. In turn, the firm’s market share is a decreasing function of its price relative to the aggregate price index. This implies that the observed size distribution of firms is an adequate proxy of the underlying productivity distribution of viable firms in a given sector. Empirical evidence as to the evolution of the observed size distribution of firms overtime is rather meagre.

Cabral and Mata (2003)\textsuperscript{54} compute firms’ size distribution in Portugal in 1983 and 1991 and find this distribution to have remained roughly unchanged.

For Canada, Trefler (2004) finds that trade integration with the United States had not induced any scale effects, but rather a strong firm turnover effect. This selection effect has not

\textsuperscript{51} See infra, section II. Indirect evidence of the Pareto distribution parameter decrease – assuming firms were initially Pareto distributed in productivity heterogeneity - is provided by Trefler (2004). He finds that the driving force behind the Canadian 17\% long-run labour productivity gain following the implementation of the FTA with the US was a plant selection effect which induced a rise in average firm efficiency. Trefler (2004)

\textsuperscript{52} Combes, Duranton, Gobillon, Puga, and Roux (2009) show that in the absence of a selection effect but allowing for agglomeration effects, firms’ productivity distribution following trade integration would be right-shifted and dilated at the right-hand side: the log standard deviation of productivity dispersion would also increase, leading to a decrease in the parameter capturing the inverse of firms’ productivity dispersion.

\textsuperscript{53} For multiproduct firms, Nocke and Yeaple (2009) show that globalization reduces the degree of skewness in firms’ size distribution which, in the heterogeneous firms’ framework, is an adequate proxy of firms’ productivity distribution. Note that these authors make the hypothesis that domestic firm size is inversely linked to the firm’s marginal cost. This is because in the context of multiproduct firms, the firm is all the more efficient as it produces goods closely situated to its core competence. Thus, the bigger the scope of the firm, the higher its marginal costs. We should be careful in comparing MPF implications of globalization for firms’ size distribution with single product firms for which the scope effect is absent by definition, and for which the higher is their productivity draw, the lower is their marginal costs, and the bigger is their size.

\textsuperscript{54} Cabral and Mata (2003).
resulted in market share shifts towards high productivity firms, but rather in plant turnover which resulted in rising average technical efficiency.

Combes, Duranton, Gobillon, Puga and Roux find that trade integration results in a right-shift and a right-hand side dilatation of firms’ productivity distribution for France, combined with a minor left-truncation of the distribution.\textsuperscript{55}

We conclude that there is insufficient empirical evidence in order to formulate a clear statement as to the evolution of the inverse of firms’ productivity dispersion overtime.

In a nutshell, the gravity model derived from a monopolistic competition model allowing for firm heterogeneity does not help elucidate the distance puzzle.

2.2. Ricardian models’ ability to explain the distance puzzle

In the Eaton and Kortum (2002) framework, goods are assumed homogeneous in any given sector, and each country chooses to consume the least expensive good in each sector. The absence of national product differentiation entails that the Armington elasticity of substitution does not exist in this framework. The CES utility function is defined at the inter-sectoral level only. The distance coefficient does not depend on this inter-sectoral elasticity of substitution as it is invariant across countries. Instead, the distance coefficient is directly impacted by the parameter which corresponds to the inverse of the degree of heterogeneity in relative country efficiencies in production across the whole range of goods. The lower the degree of heterogeneity in relative production efficiency, and the higher is the sensitivity of trade flows to distance.

\textsuperscript{55} Combes, Duranton, Gobillon, Puga, and Roux (2009) conclude that agglomeration effects dominate selection effects.
This section has shown that the heterogeneous firms’ framework points to potentially increasing firm heterogeneity dispersion in the context of trade integration, e.g. a decreasing distance coefficient. As for Ricardian trade models, the prediction is less clear. Indeed, Davis shows that if countries have similar endowments and production technologies such that, within each sector, varieties’ inputs’ vectors are linearly dependent, but each trading partner has a technical comparative advantage in the production of one specific variety, then trade would foster national specialization in a given specific variety. This means that even if we make the hypothesis that production functions and factor endowments of countries have grown more similar since the 1960s, it would not entail decreasing national specialization if each country retained a technical advantage in the production of one specific variety. Indeed, as noted by Head and Ries, Ricardian type models can be reinterpreted as part of a broader class of national product differentiation models, e.g. models in which a change in market size does not modify countries’ variety characteristics or firms’ production locations. To sum up, further investigation of the evolution of the dispersion in countries’ relative efficiency in production across goods is needed to make a prediction as to the evolution of the distance coefficient in the Ricardian type framework.

We conclude that on the supply side, it is the Ricardian framework which would be potentially compatible with a non-decreasing distance coefficient. We now turn to the analysis of the demand side explanation of a non decreasing distance coefficient, e.g. the hypothesis of

56 See Davis (1995).
58 The distance coefficient would be increasing if the dispersion in countries’ relative efficiencies in production across goods was shown to have decreased. It would be unchanged if increased national specialization exactly offset increasing similarity of production capabilities and factor endowments.
endogenously increasing sector-specific Armington elasticities of substitution in the number of traded national varieties.

3. **III. Armington elasticities in the distance puzzle**

In this section, we study the ability of the most influential theoretically derived gravity model of trade to elucidate the distance puzzle. The model by Anderson and van Wincoop assumes that each country produces a specific variety of the good of interest. The magnitude of the substitution effect included in the distance coefficient \( \alpha_1 \) depends on the elasticity of trade costs to distance \( \beta \) and the potential for substituting away from goods in distant localities to goods in nearer localities \( 1-\beta \).\(^{59}\) This potential is constant in each market for all varieties coming from different countries, whatever their market share.

This hypothesis replicates the one presented by Armington in the 1960s that the goods produced in each country be considered a different, country-specific, variety.\(^{60}\) The elasticity of substitution between these country-specific varieties is called an “Armington elasticity of substitution”. The one usually estimated is between domestic products and an aggregate import good: the “upper tier” Armington elasticity of substitution.\(^{61}\) The one we are looking at in the context of the gravity equation is between all imported varieties of a good, e.g. the “lower-tier” Armington elasticity of substitution.\(^{62}\)

In the Anderson and van Wincoop specification of the gravity equation, the *level* of the lower-tier Armington elasticity of substitution \( \beta \) has a direct impact on the distance

---

\(^{59}\) The distance coefficient is equal to \( \alpha_1 = \beta(1-\beta) \) following Anderson and van Wincoop (2003). See also Anderson and Wincoop (2004).

\(^{60}\) Armington (1969).


coefficient. In this section, we put forward the hypothesis that this lower-tier Armington elasticity of substitution could have changed since the 1960s.

3.1. Armington elasticities and trade integration: increased substitutability or increased product differentiation?

3.1.1. Armington elasticities could be increasing in the number or the degree of similarity of available country-specific varieties

The Anderson and van Wincoop gravity model is derived from assumptions on consumer preferences and firm characteristics of the monopolistic competition models of international trade developed in the late 1970s and the 1980s. These assumed a world comprised of identical countries and firms, in which consumers are characterized by a constant elasticity of substitution (CES) across goods, and across time periods. Consumers value variety for itself because an increase in the number of varieties mechanically reduces the aggregate price index. Given consumers’ preference for variety, they choose to spread expenditure on all of the available horizontally differentiated varieties in a given representative sector.

In the Anderson and van Wincoop model, the CES monopolistic competition framework is applied to country-specific varieties of a given good. By definition, this specification does not allow for a mechanism conducive to a change in the lower-tier Armington elasticity of substitution. However, if we want to examine the impact of a potential change in this parameter on the elasticity of trade flows to distance, we have to look for a mechanism which renders endogenous the lower-tier Armington elasticity of substitution.

64 Bilbiie, Ghironi, and Melitz (2008).
65 If there were no trade costs, expenditure would be evenly spread on all of the equally priced varieties available in the world economy. See Krugman (1980).
In consumer theory, two commonly admitted mechanisms could be at work. First, an increasing similarity of country-specific varieties would entail their increased substitutability. Second, it could be argued that the degree of substitutability is itself increasing in the number of available horizontally differentiated varieties.\footnote{This is the argument put forward by Melitz in an appendix to Melitz (2003): the author suggests that the elasticity of substitution could be endogenously increasing if firms are unable to maintain the same degree of product differentiation once a larger number of varieties have been made available.} It could be argued that trade integration has encouraged both mechanisms since the 1960s in that a growing number of countries have started producing a set of goods similar to that of developed countries\footnote{Schott (2004).}, both increasing the number of available varieties and, potentially, their degree of similarity. This could have resulted in an increased lower-tier Armington elasticity of substitution.\footnote{In this paper, we deliberately stay within the CES framework. Additional mechanisms of increasing price responsiveness of demand in the number of available varieties can be suggested if one moves away from this specification of consumer preferences.}

Empirically, an additional explanation of increasing \textit{estimated} Armington elasticities as a consequence of trade integration is provided by Welsch (2006)\footnote{Welsch (2006).}. If the estimation method of Armington elasticities does not allow separating the effect of the perceived difference in product characteristics between imported varieties from the effect of unobserved resistance to trade, then estimated Armington elasticities will increase if product characteristics of imported varieties are perceived to become more similar (\textit{higher degree of substitutability}). But they will also increase if unobserved trade barriers’ impact is reduced. Thus, trade integration could potentially result in \textit{de facto} higher substitutability between imported varieties as unobserved relative trade resistances diminish.

3.1.2. \textit{Armington elasticities could be decreasing in the degree of product differentiation}.

In the Armington framework, firm-level varieties are assumed to be perfectly homogeneous at the sub-national level. This impedes considering firm-level strategies of

\footnotesize{\textsuperscript{66} This is the argument put forward by Melitz in an appendix to Melitz (2003): the author suggests that the elasticity of substitution could be endogenously increasing if firms are unable to maintain the same degree of product differentiation once a larger number of varieties have been made available. \textsuperscript{67} Schott (2004). \textsuperscript{68} In this paper, we deliberately stay within the CES framework. Additional mechanisms of increasing price responsiveness of demand in the number of available varieties can be suggested if one moves away from this specification of consumer preferences. \textsuperscript{69} Welsch (2006).}
increased product differentiation as firms in any given sector could seek to amplify the perceived difference of their product characteristics in order to counter the effect of an increasing number of varieties in the sector following trade integration.

To consider the possibility of decreasing lower-tier Armington substitution elasticity, we have to look for mechanisms which would drive country-level product differentiation. As suggested by Davis, the Ricardian trade framework is well suited for this purpose.\textsuperscript{70} If we put forward the hypothesis that since the 1960s, countries have grown more similar in their factor endowments and in their technical capabilities while retaining technical comparative advantages for producing one specific variety within any given sector, then we could observe increased country-level product specialization. For example, this could have resulted in an enhanced quality-specialization at the country level. Enhanced country specialization could potentially come to dominate the direct effect of trade integration, i.e. the increasing number of available varieties. Lower-tier Armington elasticities of substitution could then be found to decrease overtime at the intra-sectoral level.

Empirically, this hypothesis is borne out by Broda and Weinstein (2006) finding that lower-tier Armington elasticities of substitution for American imports in 1972-1988 and 1990-2001 decreased particularly strongly for goods classified as differentiated according to Rauch (1999). It is also consistent with Schott (2004)\textsuperscript{71} finding of increasing unit value dispersion in American imports between products of different national origin.

An additional explanation for potentially decreasing estimated Armington elasticities is the \textit{within-sector} composition effect\textsuperscript{72}. If the estimation is conducted on data which is not highly disaggregated, Armington elasticities would be retrieved for sectors which comprise heterogeneous products, and sectoral product composition may change overtime. As trade

\textsuperscript{70} Davis (1995).
\textsuperscript{71} Schott (2004)
\textsuperscript{72} Welsch (2006).
integration intensifies the degree of competition, it could bring about increased specialization at the country level in certain product groupings within a given sector. The within-sector composition effect would result in decreasing estimated Armington elasticities overtime as sectors become increasingly differentiated across countries. If not in the short term, in the medium-long term decreasing intra-sectoral Armington elasticities would be observed. In accordance with this conjecture, Welsch (2006) has found that lower-tier Armington elasticities peaked in the 1970s, then progressively decreased in the 1990s for the French market.

We conclude that it is necessary to conduct a series of empirical tests to check whether intra-sectoral Armington elasticities’ evolution overtime can help elucidate the distance puzzle.

3.2. Have Armington elasticities increased since the 1950s?

3.2.1. Armington elasticities’ estimates in 1950-2000: historical overview

Traditional estimation methodologies and their drawbacks

[TO BE COMPLETED]

At the aggregate level: time-series estimations, cross-section estimations

[TO BE COMPLETED]

At the sectoral level: time-series estimations, cross-section estimations

[TO BE COMPLETED]

3.2.2. Armington elasticities’ estimates overtime: evidence of a decrease?

Increased differentiation could be due to two distinct phenomena. According to Welsch (2006), as Armington elasticities are retrieved for heterogeneous product groups, the composition of these groups may change overtime becoming increasingly different among countries. An alternative explanation would be linked to quality-specialization: the composition of the product groups would be similar across countries, but each country would specialize in a specific quality-segment, and the range of existing qualities would expand overtime.
Broda and Weinstein (2006) study the evolution of substitution elasticities between 1972-1988 and 1990-2001 for American imports. They find that these have decreased for all types of goods, and at all levels of product disaggregation (10-digit, 5-digit, 3-digit). This finding is consistent with an increased dispersion of imports’ prices observed by Broda and Romalis (2009). It is also consistent with Welsch (2006)’s hypothesis of increased country-level product differentiation spurred by an intensifying degree of international competition. It is in direct contradiction with the hypothesis of increased goods’ substitutability and would actually deepen the distance puzzle.

Berthelon and Freund (2008) use Broda and Weinstein elasticity estimates to test a hypothesis grounded in Buch, Kleinert, and Toubal (2004)’ finding that the distance coefficient is a function of the evolution of the relative share of trade between countries situated close by and countries situated far apart. If the share of close neighbors’ trade increases quicker than the share of trade of far apart countries, the distance coefficient of aggregate trade would be increasing. Berthelon and Freund (2008) put forward the hypothesis that trade among close neighbours would increase quicker than trade between countries far apart if the Armington elasticity of substitution between goods had increased.

This hypothesis is not rejected by the data. Berthelon and Freund find that changes in substitution elasticities estimated by Broda and Weinstein (2006) in 1972-1988 and 1990-2001 are positively correlated to changes in the absolute value of the distance coefficient. The magnitude of this effect is however small: a one unit increase in the elasticity of substitution leads to a 0.004 increase in the absolute value of the distance coefficient. Moreover, Berthelon and Freund (2008) note that changes in substitution elasticities used in their work cannot

74 See Hummels (2001) for a formal presentation of a multi-sector model in which producers minimize trade costs by locating production of complementary goods among close neighbors. This means that substitutability works through two channels: the direct effect of substituting away from higher priced varieties captured by the Armington substitution elasticity, and the indirect effect of producers’ cost minimization which induces locating production within short distances, and, in turn, increases the weight of close neighbors’ trade in total trade. Hummels (2001)
explain the average increase in the distance coefficient insofar as these substitution elasticities are decreasing on average for all types of goods. Thus, even though the hypothesis of increased substitutability is not rejected, Berthelon and Freund (2008) conclude that increasing distance coefficients are due to the decreasing share of non-distance dependent costs relatively to distance-dependent costs in trade, and not to increasing goods’ substitutability.

Thus, previous empirical evidence is inconclusive as to the evolution of lower-tier Armington elasticities of substitution and their ability to help explain the distance puzzle. This is why we provide new estimates of these elasticities in 1962-2006.

3.2.3. By product estimations of Armington substitution elasticites in 1962-2006 on the COMTRADE database

Following Feenstra, we note that the Armington and van Wincoop gravity model is derived from the following specification of the value of bilateral trade flows:\footnote{Feenstra (2004), chapter 5, equation 5.26.}

\[
X_{ij} = N_i Y_j \left( \frac{P_{ij}}{P_j} \right)^{1-\sigma}
\]

Here \(X_{ij}\) is the value of exports from country \(i\) to country \(j\), \(Y_j\) is total expenditure (income) of country \(j\), \(N_i\) is the number of varieties exported by country \(i\), \(p_{ij}\) is the cif price of \(i\)’s exports to country \(j\), \(P_j\) is \(j\)’s aggregate price index, and \(\sigma\) is the Armington elasticity of substitution.

Loglinearizing this expression, and including fixed exporter and importer effects, gives the following equation to be estimated for each product (k) in order to directly retrieve by product Armington elasticities of substitution for each year in 1962-2006:

\[
\ln(x_{ij,t}^k) = (1-\sigma_t^k) \ln(p_{ij,t}^k) + f_{exp}^k \ln(1) + f_{imp}^k \ln(1)
\]
We conduct this estimation for products present in the dataset in all years, with unit values used as proxies for prices, and with the sample restricted to importer reported data in order to guarantee that the reported price is the cif price.

We then compute the weighted mean price elasticity of trade. Figure 5 shows that, despite the uncertainty around the estimate of the coefficient, the absolute value of the elasticity has decreased from the early 1980s to the mid-2000s based on COMTRADE data at the 4-digit level. The evolution of Armington elasticities thus deepens the distance coefficient puzzle.

![Figure 5](attachment:figure5.jpg)

[additional graphs to be included]

3.3. **The role of the Armington elasticities of substitution in the distance puzzle**

[TO BE COMPLETED]

4. **Conclusion**

[TO BE COMPLETED]
5. **References**


Hummels, David. 1999. “Have International Transportation Costs Declined?”. Purdue University.
6. **Appendix: list of included FTAs**

*[in squared brackets: years for which this FTA appears in the database]*

EC (European Communities), then EU (European Union): [1962-2006]

EFTA (European Free Trade Association): [1962-2006]


COMECON (Union of Mutual Economic Assistance): [1964-1990]


EC-OCT (EC FTA with Overseas Countries and Territories): [1971-2006]

CARICOM (Caribbean Community and Common Market): [1973-2006]

EEA (European Economic Area: EC-EFTA FTA): [1973-2006]


EFTA-Spain FTA: [1980-1985]


SPARTECA (South Pacific Regional Trade and Economic Cooperation Agreement): [1981-2006]

CER (Australia-New Zealand FTA): [1983-2006]

US-Israël FTA: [1985-2006]

US-Canada FTA: [1988-2006]


EC-Andorra FTA: [1991-2006]

EFTA-CEE FTA: [1992-2003]

EU-CEE FTA: [1992-2003]

AFTA (Association of South East Asian Nations FTA): [1992-2006]
CEFTA (Central European FTA): [1993-2006]
CIS (Commonwealth of Independent States): [1995-2006]
SAPTA (South Asian Preferential Trade Arrangement): [1995-2006]
WAEMU (West African Economic and Monetary Union): [1996-2006]
PAFTA (Pan Arab FTA): [1998-2006]
SACU (Sub Saharan South African Customs Union): [2000-2006]
EAC (East African Community): [2000-2006]
MERCOSUR (Southern Common Market): [2006]

Additional variables which appear in the database in all years:
GATT/WTO membership

dev1: high income group (GNI per capita > 9205 US dollars)
dev2: upper middle income group (GNI per capita between 2975 and 9205 US dollars)
dev3: lower middle income group (GNI per capita between 746 and 2975 US dollars)
dev4: low income group (GNI per capita <=745 US dollars).

Income groupings were taken from the World Bank classification for the year 2000.