IFN Working Paper No. 957, 2013

Heterogeneous Firms, Globalization and the Distance Puzzle

Mario Larch, Pehr-Johan Norbäck, Steffen Sirries and Dieter Urban

Research Institute of Industrial Economics
P.O. Box 55665
SE-102 15 Stockholm, Sweden
info@ifn.se
www.ifn.se

Heterogeneous Firms, Globalization and the Distance Puzzle*

Mario Larch† Pehr-Johan Norbäck‡ Steffen Sirries§

Dieter M. Urban¶

March 1, 2013

Abstract

Despite the strong pace of globalization, the distance effect on trade is persistent or even growing over time (Disdier and Head, 2008). To solve this *distance puzzle*, we use the recently developed gravity equation estimator from Helpman, Melitz and Rubinstein (2008), HMR henceforth. Using three different data sets, we find that the distance coefficient increases over time when OLS is used, while the non-linear estimation of HMR leads to a decline in the distance coefficient over time. The distance puzzle thus arises from a growing bias of OLS estimates. The latter is explained by globalization more significantly reducing the downward bias from omitting zero trade flows than it reduces the upward bias from omitting the number of heterogeneous exporting firms. Furthermore, we show that including zero-trade flows cannot solve the distance puzzle when using HMR. The HMR estimates are strongly correlated with the time pattern in freight costs reported by Hummels (2007).

JEL-Classification: F13, F14, F23

Keywords: Distance puzzle, gravity estimation, zero trade flows, firm heterogeneity

*We thank Eddie Bekkers, Benedikt Heid, Sebastian Krautheim, Simon Loretz and seminar participants at the ETSG, the annual meeting of the German Economic Association and at Bayreuth for helpful comments. Pehr-Johan Norbäck gratefully acknowledges financial support from the Jan Wallander and Tom Hedelius Foundation.

†University of Bayreuth, ifo Institute for Economic Research, CESifo and GEP at the University of Nottingham, Universitätsstraße 30, 95447 Bayreuth, Germany. E-mail: mario.larch@uni-bayreuth.de.

‡Research Institute of Industrial Economics (IFN), P.O. Box 55665, SE-102 15 Stockholm, Sweden. E-mail: pehr-johan.norback@ifn.se. Home page: www.ifn.se/pjn.

§University of Bayreuth, 95447 Bayreuth, Germany. E-mail: steffen.sirries@uni-bayreuth.de.

¶Dieter sadly passed away on 7th March, 2011. We miss his friendship, kindness and brilliance.

1 Introduction

> *"From the telegraph to the Internet, every new communications technology has promised to shrink the distance between people, to increase access to information, and to bring us ever closer to the dream of a perfectly efficient, frictionless global market." (Thomas Friedman, 2005, p. 204)*

The many facets of globalization like the increased trade in final goods, intermediate inputs and services, or the increased international mobility of capital and labor, are perceived to bring countries closer together, shrinking the impediments of distance. However, gravity estimations regressing bilateral trade on distance, inter alia, tell us the opposite. Disdier and Head (2008) undertake a meta analysis of the magnitude of the distance coefficient based on 103 empirical studies and find that (i) the mean effect of the distance coefficient is about $|-0.9|$ across studies, and (ii) the negative impact of distance on trade rose around the middle of the century and has remained persistently high ever since.[1]

A stable or *rising* distance coefficient over time is puzzling because the distance coefficient has the structural interpretation of the elasticity of bilateral trade with respect to distance (e.g. Anderson and van Wincoop, 2003). Transport technology is known to be biased in favor of long distances (see Hummels, 2007), which should lead to a decrease of the distance effect. Hence, the elasticity of bilateral trade with respect to distance should fall with increasing globalization.

In this paper, we use the recently developed gravity equation estimator from Helpman, Melitz and Rubinstein (2008), henceforth HMR, which controls for sample selection and exporter heterogeneity, to solve this *distance puzzle*.

We apply the HMR estimator on world trade for three different data sets, two aggregate data sets and one at the industry-level over different time periods.[2] We find that the HMR estimates of the distance coefficient (in absolute value) are decreasing on average over time as expected. These estimated coefficients are also strongly correlated with the time pattern in freight costs reported by Hummels (2007) and Brun, Carrère, Guillaumont and de Melo (2005), which in turn depend on fluctuations of oil prices.

Having shown that the HMR estimator does produce decreasing distance coefficients over time, we then compare the outcome with OLS. We first confirm the finding of HMR that OLS produces larger distance coefficients (in absolute value). More importantly, we show that these distance coefficients increase over time. Hence, the distance puzzle arises

[1] This paper also provides a good collection of references for the "distance puzzle". Hence, we here dispense with a discussion of all relevant papers and with providing all references.

[2] Berthelon and Freund (2008) document the distance puzzle on bilateral industry data rather than on bilateral country data.

due to the fact that *the bias of OLS increases over time.*

To explain the increasing OLS bias, we formally derive how the bias of the OLS distance coefficient evolves over time if the true data generating process is the HMR model and the elasticity of trade with respect to distance[3] decreases during globalization through, for instance, improved transport and communication technologies.

We note that if the HMR model is the true model, OLS estimates suffer from two sources of bias. First, there is a *sample selection bias* because bilateral trade is measured as a logarithm and zero values of bilateral trade turn into missing values. As small or distant countries are more likely to have small trade flows, measurement errors in export flows will more likely lead to zero trade flows for those countries. This leads to a positive correlation of the error term with distance, causing a downward bias in the distance coefficient, i.e. the value of the distance coefficient is too small in absolute terms. Hence, accounting for zero trade flows does not explain the distance puzzle.

Second, there is an *omitted variable bias* from ignoring that firms are heterogeneous in productivity. If an index of the size and the number of exporting firms in an industry is not included as a control in the gravity estimation, then it appears in the regression error, causing a negative correlation between error and distance, because there are less exporters to more distant destinations. Hence, the distance coefficient is upward biased through omitting a control on firm productivity, i.e. the value of the distance coefficient is too large in absolute terms.

As these two biases work in opposite directions, the overall change of the bias from OLS estimates is ambiguous. We first reproduce previous findings that the OLS estimates are upward biased, i.e. the distance coefficient is too large in absolute value. Hence, the downward bias from sample selection due to omitting zero trade flows is outweighed by the upward bias due to ignoring that firms are heterogeneous.

We then show how the two biases evolve over time in the course of globalization measured as a fall in the elasticity of trade with respect to distance. We first show that the downward bias through sample selection must decrease over time. Intuitively, as trade costs decrease, ever less country pairs have zero trade flows and eventually all countries trade with each other. But then the sample selection bias disappears, i.e. the distance coefficient rises. We then show that the upward bias from omitting the number of exporting firms also becomes smaller over time when the elasticity of trade with respect to distance falls. Intuitively, at a lower trade elasticity, most firms will export, reducing

[3]We follow this interpretation of the distance coefficient throughout the paper. Assuming decreasing distance costs would lead to a *flatter world* without relative differences of trade volumes across trading partners w.r.t to distance. However, Buch, Kleinert and Toubal (2004) argue that the distance puzzle is not that puzzling when the effect of distance is interpreted in absolute terms. Under the assumption of linear dependency of trade costs with respect to distance, they show that a potential decline in the impact of distance would be caught by the constant term in the gravity equation. But still we should – but do not – observe a decline in the relative impact of distance on bilateral trade, which is exactly measured by the elasticity we look at.

the upward bias, i.e. the distance coefficient decreases. Since both biases decrease in the course of globalization (measured as a fall of distance with respect to trade) globalization has an ambiguous effect on the bias of OLS in general.

Our estimates show that the bias of OLS increases over time. Hence, it must be the case that the downward bias from sample selection decreases faster than the upward bias from not controlling for the number and size of exporting firms. Thus, the HMR model implies that the distance puzzle arises from firm heterogeneity having become relatively more important over time. This is nicely in line with empirical evidence provided by Poschke (2011), that, as countries develop, the distribution of firm sizes becomes more dispersed.

For future work, we also suggest a linearization of the HMR estimator, which is comparable to the non-parametric approach of Helpman, Melitz and Rubinstein (2008). This approach is easy to implement with standard econometric programs because it is estimable via OLS. We show that such a simplified estimator performs just as well as the original nonlinear least squares version.

We also show that a Heckman estimator deviates from the HMR estimates and produces bigger distance coefficients and an increasing difference to the OLS estimates over time. The Heckman correction results lead to the conclusion that taking into account zero trade flows cannot solve the distance puzzle, as expected from our theoretical results.

Alternative attempts to solve the distance puzzle stem from Felbermayr and Kohler (2006), using Tobit estimates to take zero trade flows into account.[4] Other studies explain why the substitution elasticity may have been rising over time (Glaeser and Kohlhase (2004), Krautheim (2011), Lawless and Whelan (2007), Berthelon and Freund (2008)), possibly overcompensating the fall in trade costs, which both determine the distance coefficient in theory. Duranton and Storper (2008) provide an alternative model to rationalize rising overall trade costs besides falling transport costs. They assume vertically linked industries in which the quality of inputs is not contractible and where providing a given level of quality to suppliers becomes more costly with distance. Their main finding is that lower transport costs imply that higher quality inputs are traded in equilibrium,

[4]There is ample evidence from microdata for particular countries that the extensive margin matters. Bernard, Jensen and Schott (2006) use firm-level data to distinguish the entry and exit of firms into and out of exporting (extensive margin) from the export volumes of exporting firms (intensive margin). They find that a reduction in trade costs may increase industry productivity through changes on the extensive margin. Hummels and Klenow (2005) use disaggregated product-level data to distinguish between the variety dimension (extensive margin) and the quality as well as the quantity dimension (intensive margin). One of their main results is that adverse terms-of-trade effects occur more frequently if growth takes place mainly at the extensive margin. Similarly, Baldwin and Harrigan (2011) use product-level data on bilateral U.S. exports demonstrating that a large part of potential export flows are zero, and showing that the incidence of these zero export flows is strongly correlated with distance and importing country size. Hillberry and Hummels (2008) analyze trade at the five-digit zip codes and decompose the extensive and intensive margins of shipments. Their main finding is that distance reduces aggregate trade values primarily by reducing the number of commodities shipped and the number of establishments shipping. However, the extensive margin is important over very short distances.

and the effect of this higher quality is that there is an increase in trade costs. Yotov (2012) proposes to measure the effects of distance on international trade relative to the effects of distance within national borders as a simple and useful solution of the distance puzzle. He finds a drop in the impact of distance on trade of roughly 50% from the mid-sixties to 2005. Finally, using bilateral country data for the year 1986, HMR find that their estimated distance coefficient represents a drop of roughly one third as compared to OLS. However, HMR do not examine the evolution of the distance coefficient over time. Hence, none of the mentioned papers discusses the role of the omitted variable problem of firm heterogeneity in creating an increasing bias over time, which is the contribution of our paper.

The remainder of the paper is organized as follows. Section 2 derives the gravity equation controlling for zero trade flows and firm-level heterogeneity following HMR in subsection 2.1, whereas we calculate the biases of OLS estimates in subsection 2.2. Section 3 presents our estimation equation in subsection 3.1, describes the data in subsection 3.2, and gives the results in subsection 3.3. The last section concludes the paper.

2 Theory

2.1 Deriving the gravity equation from Helpman, Melitz and Rubinstein (2008)

The HMR model is a multi-country monopolistic competition model with heterogeneous firms and identical consumers with CES "love-of-variety" utility functions à la Dixit and Stiglitz (1977). Since we have bilateral industry trade data, we will add the assumption of multiple sectors in the world economy, each characterized by monopolistic competition. There are N_{ih} firms in a sector h of country i, each producing a differentiated variety l. For ease of notation, we will drop the subscript h for sector whenever obvious. With a substitution elasticity between any two varieties $\varepsilon > 1$, the demand $x_{ij}(l)$ for a variety l, consumed in country i and produced in country j is

$$x_{ij}(l) = \frac{p_{ij}(l)^{-\varepsilon}}{P_i^{1-\varepsilon}} \mu_i Y_i, \tag{2.1}$$

where $p_{ij}(l)$ is the associated price, $P_i = \left[\int_{\mathcal{B}_i} p_{ij}(l)^{1-\varepsilon} dl\right]^{\frac{1}{1-\varepsilon}}$ is the price index on the set of $\mathcal{B}_i$ symmetric domestic and imported differentiated goods consumed in country i, Y_i is the income in country i and μ_i is the (constant) share of income spent on a sector under consideration by consumers of country i.

A firm l in country j produces one unit of output at the cost $c_j a$, where c_j is the minimum cost of a bundle of inputs which is country- and sector-specific, and where $a(l)$

is a firm-specific input coefficient implying that firm l's productivity is given by $1/a(l)$. As in Melitz (2003), firms can be identified by their productivity, allowing us to exchange the index l for a. Shipping goods across borders involves iceberg trade costs, which implies that $\tau_{ij} > 1$ units of output need to be shipped from country j to country $i \neq j$ in order for one unit to arrive. Delivering to home country customers involves no trade costs, i.e. $\tau_{ii} = 1$. Exporting across borders is also associated with country- and sector-specific fixed export costs $f_{ij}c_j$, i.e. $f_{ii} = 0$ and $f_{ij} > 0$ for $j \neq i$.

The operating profit from producing a variety a in country j and selling it to country i is then $\pi_{ij}(a) = [p_{ij}(a) - \tau_{ij}ac_j]\, x_{ij}(a) - f_{ij}c_j$. From (2.1), this implies that a firm with productivity $1/a$ producing in country j for exports to country i will charge the price:

$$p_{ij}(a) = \frac{1}{\alpha}\tau_{ij}ac_j, \tag{2.2}$$

where $1/\alpha = \varepsilon/(\varepsilon - 1)$ is the standard mark-up. It also follows that domestic consumers are priced at $p_{jj}(a) = 1/\alpha ac_j$.

The assumption of the absence of fixed costs in home sales operations and fixed set-up costs incurred in exporting operations implies that only a fraction of country $j's$ N_j firms will export to country i. To characterize exporters, define the reduced-form operating profit for country j exporters as $\pi_{ij}(a) = \pi_{ij}(p_{ij}(a))$, or:

$$\pi_{ij}(a) = (1/a)^{\varepsilon-1}\,(1-\alpha)\left(\frac{\tau_{ij}c_j}{\alpha P_i}\right)^{1-\varepsilon}\mu_i Y_i - f_{ij}c_j. \tag{2.3}$$

Let firm-productivity $1/a$ be characterized from the cumulative distribution $G(a)$ with density $g(a)$ over the finite support $a \in [a_L, a_H]$, where a_L is the firm-specific input coefficient of the most productive firm and a_H that of the least productive firm, respectively. We need the lower bound because we want to generate zero trade flows and the upper bound because we will assume a Pareto distribution.[5]

The cut-off productivity for being an exporter to country i based in country j, $1/a_{ij}$, is then determined from the zero-profit condition, $\pi_{ij}(a_{ij}) = 0$, or:

$$a_{ij} = \left[(1-\alpha)\,\tfrac{\mu_i Y_i}{f_{ij}c_j}\right]^{\frac{1}{\varepsilon-1}}\frac{\alpha P_i}{\tau_{ij}c_j}. \tag{2.4}$$

In Figure 1(iii), we show the operating profits for a country j firm as a function of the firm-specific input-coefficient a in exporting $\pi_{ij}(a)$ and home sales $\pi_{jj}(a)$. The operating profits decrease in a and, hence, increase in productivity $1/a$, as shown in Figure 1(ii). Thus, firms in country j can only recover export fixed costs and export to country i if their productivity exceeds the cut-off productivity. Firms with a productivity lower than

[5]For an infinitely productive firm ($a = 0$), the operating profits would always be large enough for finite positive fixed costs ($f_{ij}c_j$) to ensure exports to every country.

the cut-off productivity will only sell on the domestic market.

Trade between countries i and j can now be characterized as follows: >From (2.1) and (2.2), the export revenue for a country j firm is $p_{ij}(l)x_{ij}(l) = (1/a)^{\varepsilon-1}\left(\frac{\tau_{ij}c_j}{\alpha P_i}\right)^{1-\varepsilon}\mu_i Y_i$. As shown in Figure 1(i), the number of firms endowed with a productivity $1/a$ is $N_j g(a)$. Hence, the aggregate imports of county i from country j, M_{ij}, are:

$$M_{ij} = \int_{a_L}^{a_{ij}} (1/a)^{\varepsilon-1}\left(\frac{\tau_{ij}c_j}{\alpha P_i}\right)^{1-\varepsilon}\mu_i Y_i N_j g(a)da. \tag{2.5}$$

Define V_{ij} as a term indicating the share of exporting firms ($g(a)$) weighted by a measure of firm size ($a^{1-\varepsilon}$):

$$V_{ij} = \begin{cases} \int_{a_L}^{a_{ij}} a^{1-\varepsilon} g(a)da & \text{for} \quad 1/a_L \geq 1/a_{ij}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

Combining (2.5) and (2.6), imports from country j to country i can be written as:

$$M_{ij} = \left(\frac{c_j\tau_{ij}}{\alpha P_i}\right)^{1-\varepsilon}\mu_i Y_i N_j V_{ij}. \tag{2.7}$$

To obtain an estimation equation from (2.7), HMR proceed in several steps. First, they specify trade costs as follows:

$$\tau_{ij}^{\varepsilon-1} = D_{ij}^{\gamma}, \tag{2.8}$$

where D_{ij} is the distance between source and destination country and γ is the elasticity of bilateral trade with respect to distance, which may vary across industries.

Second, they assume a truncated Pareto distribution $G(a) = (a^k - a_L^k)/(a_H^k - a_L^k)$, where $k > (\varepsilon - 1)$ is the shape parameter and show that V_{ij} in (2.6) takes the form $V_{ij} = \psi W_{ij}$, where:

$$W_{ij} = \max\left\{\left(\frac{a_{ij}}{a_L}\right)^{k-\varepsilon+1} - 1, 0\right\}, \tag{2.9}$$

and $\psi = (ka_L^{k-\varepsilon+1})/((k-\varepsilon+1)(a_H^k - a_L^k))$ is a constant. Note that trade is observed whenever $W_{ij} > 0$, that is, if $a_{ij} > a_L$, implying that the cut-off productivity is smaller than the productivity of the most efficient firm in the industry, $1/a_{ij} < 1/a_L$. From Figure 1(i), note that the number of exporting firms $\int_{a_L}^{a_{ij}} N_j g(a)da$ is increasing in a_{ij}. Importantly, from (2.9), it then follows that the number of country j exporters is increasing in W_{ij}, since W_{ij} is increasing in a_{ij}.[6] Note also that a reduction of γ thus increases the number of exporting firms since it lowers the threshold productivity level $1/a_{ij}$.

Third, they add a multiplicative error $\exp(u_{ij})$ on the right-hand side of equation

[6] As illustrated in Figure 1, note that positive trade $M_{ij} > 0$ requires $a_{ij} > a_L$. But then $W_{ij} = \max\left\{\left(\frac{a_{ij}}{a_L}\right)^{k-\varepsilon+1} - 1, 0\right\} = \left(\frac{a_{ij}}{a_L}\right)^{k-\varepsilon+1} - 1$. Thus, we have $\omega_{ij} \equiv \ln W_{ij} = \ln\left[\left(\frac{a_{ij}}{a_L}\right)^{k-\varepsilon+1} - 1\right]$.

(2.7). u_{ij} is assumed to be log normally distributed such that $u_{ij} \sim N(0, \sigma_u^2)$.

Inserting (2.8) and (2.9) into (2.7) and adding $\exp(u_{ij})$, taking logs and denoting logged variables with lower-case letters, HMR obtain the gravity estimation equation in logs:

$$m_{ij} = \beta_0 + \lambda_j + \chi_i - \gamma d_{ij} + \omega_{ij} + u_{ij}, \tag{2.10}$$

where $\beta_0 = (\varepsilon - 1)\ln(\alpha) + \ln(\psi)$, importer country-fixed effects χ_i contain $\chi_i = (\varepsilon - 1)p_i + \ln \mu_i + y_i$, and exporter country-fixed effects λ_j contain $\lambda_j = (1 - \varepsilon)\ln c_j + n_j$ and the number of exporters is captured by

$$\omega_{ij} = \ln \left[\left(\frac{a_{ij}}{a_L} \right)^{k-\varepsilon+1} - 1 \right]. \tag{2.11}$$

Note that the term ω_{ij} is the only new one in the gravity equation as compared to Anderson and van Wincoop (2003).

The estimation of (2.10) is hampered by two problems. First, it is only estimated on data with positive trade flows, since the dependent variable, the log of trade volume m_{ij}, is not defined for zero import values, $M_{ij} = 0$. Second, there is an omitted variable problem through ω_{ij}, which captures the degree of firm heterogeneity in country j, information which is typically not available for gravity estimations on a world trade data set.[7]

HMR note that both problems are related to the extensive margin of trade. Rearranging (2.4), they then define an auxiliary variable Z_{ij}:

$$\begin{aligned} \left(\frac{a_{ij}}{a_L} \right)^{\varepsilon-1} &= \frac{(1-\alpha)\left(\frac{\tau_{ij} c_j a_L}{\alpha P_i} \right)^{1-\varepsilon} \mu_i Y_i}{c_j f_{ij}} \\ &\equiv Z_{ij}. \end{aligned} \tag{2.12}$$

As illustrated in Figures 1(i) and 1(iii), changes in Z_{ij} will indicate both, i.e. changes in the number of exporting firms through the cut-off a_{ij} as well as zero trade links. Noting that changes in Z_{ij} are exclusively driven by changes in a_{ij}, Figure 1(iii) shows that there will be no exports if $a_{ij} < a_L$, which is equivalent to the condition $Z_{ij} < 1$. In contrast, the trade flows will be non-zero when $a_{ij} > a_L$, which implies $Z_{ij} > 1$.

Now, we can express the omitted variable ω_{ij} by inserting equation (2.12) into (2.11)

$$\omega_{ij} = \ln \left[exp\left[\delta z_{ij} \right] - 1 \right], \tag{2.13}$$

where $\delta = (k - \varepsilon + 1)/(\varepsilon - 1)$ and $z_{ij} = \ln Z_{ij}$. The estimation strategy of HMR is to obtain an estimate of the expected value of the omitted variable ω_{ij} by estimating an

[7] Flam and Nordström (2008) have recently included a proxy variable for ω_{ij}, which is available for Swedish exports. However, they did not estimate the distance coefficient over time, which is the focus of this paper.

expected value of the auxiliary variable z_{ij}.

To obtain another estimation equation, another error term is introduced by decomposing the fixed trade costs f_{ij} as follows:

$$f_{ij} = \exp\left[\left(\phi_{EX,j} + \phi_{IM,i} + \kappa\phi_{ij} - \nu_{ij}\right)\right], \tag{2.14}$$

where $\nu_{ij} \sim N\left(0, \sigma_\nu^2\right)$ and $\phi_{EX,j}$ is a measure of the fixed export cost common across all export destinations, $\phi_{IM,i}$ is a fixed trade barrier imposed by the importing country on all exporters, and ϕ_{ij} is an observed measure of any additional country-pair specific fixed trade costs.[8]

Taking logs of Z_{ij} in (2.12) and using (2.8) and (2.14), HMR obtain an equation for a latent variable $z_{ij} \equiv \ln Z_{ij}$:

$$\begin{aligned} z_{ij} &= \gamma_0 + \xi_j + \zeta_i - \gamma d_{ij} - \kappa\phi_{ij} + \eta_{ij} \\ &= E\left[z_{ij} \left| d_{ij}, \xi_j, \zeta_i, \phi_{ij}\right.\right] + \eta_{ij}, \end{aligned} \tag{2.15}$$

where $\eta_{ij} = u_{ij} + \nu_{ij}$, $\zeta_i = (\varepsilon - 1)\, p_i + y_i + \ln \mu_i - \phi_{IM,i}$ is an importer fixed effect, $\xi_j = -\varepsilon \ln c_j - \phi_{EX,j}$ is an exporter fixed effect and σ_η^2 is the variance of η_{ij}. While the latent variable z_{ij} cannot be observed, one can observe if trade takes place. Thus, an indicator variable $T_{ij} = \mathcal{I}_{[z_{ij}>0]}$ can be defined from which the selection equation for the probability of strictly positive exports is obtained:

$$\begin{aligned} \Pr\left(T_{ij} = 1 \left| d_{ij}, \xi_j^*, \zeta_i^*, \phi_{ij}\right.\right) &= \Pr\left(z_{ij}^* > 0 \left| d_{ij}, \xi_j^*, \zeta_i^*, \phi_{ij}\right.\right) \\ &= \Pr\left(\gamma_0^* + \xi_j^* + \zeta_i^* - \gamma^* d_{ij} - \kappa^* \phi_{ij} > -\eta_{ij}^* \left| d_{ij}, \xi_j^*, \zeta_i^*, \phi_{ij}\right.\right) \\ &= \Phi\left(\gamma_0^* + \xi_j^* + \zeta_i^* - \gamma^* d_{ij} - \kappa^* \phi_{ij}\right) \\ &= E\left[z_{ij}^* \left| d_{ij}, \xi_j^*, \zeta_i^*, \phi_{ij}\right.\right], \end{aligned} \tag{2.16}$$

where $\Phi(.)$ is the cumulative distribution function of the unit normal distribution and every starred coefficient represents the original coefficient divided by σ_η.[9]

One can now in a first stage estimate (2.16) by a probit estimation. Inverting the predicted probability from (2.16) yields an estimate of the underlying latent variable $\hat{z}_{ij}^*$.

Defining $\delta = \sigma_\eta \frac{k-\varepsilon+1}{\varepsilon-1} > 0$, HMR use $\hat{\bar{\omega}}_{ij}^* \equiv \ln\left\{\exp\left[\delta\left(\hat{z}_{ij}^* + \hat{\eta}_{ij}^*\right)\right] - 1\right\}$ as an estimate for $E[\omega_{ij}|., z_{ij}^* > 0]$,[10] where $\hat{\eta}_{ij}^* = \phi(\hat{z}_{ij}^*)/\Phi(\hat{z}_{ij}^*)$ is the inverse Mills ratio from the

[8] ϕ_{ij} takes the role of an instrument in the empirical implementation and is assumed to be statistically independent of ν_{ij} as well as u_{ij}.

[9] As in every discrete choice model, the scale can be arbitrarily chosen, i.e. the model must be properly normalized. We normalize by dividing through σ_η, following HMR. This leads the error term $\eta_{ij}^* = \eta_{ij}/\sigma_\eta$ to be distributed unit normal.

[10] Santos Silva and Tenreyro (2006, 2008) note that this is not a consistent estimate because of Jensen's inequality. However, Santos Silva and Trenreyro (2008) also note that it is a reasonably accurate approximation in many practical situations. The similarity of our results from the linear approximation of

first-stage probit estimation, which itself is well-known to be a consistent estimate of $E[u_{ij}|., z^*_{ij} > 0]$.[11] Inserting these terms into (2.10), HMR show that estimation of the gravity model requires estimating the following specification:

$$m_{ij} = \beta_0 + \lambda_j + \chi_i - \gamma d_{ij} + \ln\left\{\exp\left[\delta\left(\hat{z}^*_{ij} + \hat{\eta}^*_{ij}\right)\right] - 1\right\} + \beta_{u\eta}\hat{\eta}^*_{ij} + e_{ij}, \tag{2.17}$$

where $\beta_{u\eta} \equiv corr(u_{ij}, \eta_{ij})(\sigma_u/\sigma_\eta) = corr(u_{ij}, u_{ij}+\nu_{ij})(\sigma_u/\sigma_\eta) > 0$. The term $\ln\{\exp[\delta(\hat{z}^*_{ij}+\hat{\eta}^*_{ij})] - 1\}$ corrects for the omitted variable ω_{ij} in the presence of sample selection[12] and $\beta_{u\eta}\hat{\eta}^*_{ij}$ is the well-known correction of the error term u_{ij} in the presence of sample selection. As a result, e_{ij} is an i.i.d. error term satisfying $E[e_{ij}|., T_{ij} = 1] = 0$. Therefore, one can estimate (2.17) using NLS and obtain an estimate of the distance coefficient γ, having the structural interpretation of the elasticity of bilateral trade with respect to distance for all country-pairs in the population, i.e. for positive and zero trade flows.

2.2 HMR and the distance puzzle

In this section, we use the HMR model to examine the distance puzzle. We will assume the HMR model to be the data generating process and examine to what extent the OLS estimation of (2.10) is biased and in what direction this bias goes. Then, we examine how the bias of OLS is affected by globalization.

2.2.1 The Bias of OLS

Let us start by examining the properties of an OLS estimate of the distance coefficient, $\hat{\gamma}^{OLS}$, from estimating gravity equation (2.10) without a sample selection correction and when not controlling for the omitted variable bias due to firm heterogeneity by ω_{ij}. To gain some intuition on these two biases and their direction, we first look at them in turn before considering them simultaneously. We begin by first discussing the sample selection bias and then continue with the omitted variable bias.

Selection bias We first show that the selection bias leads to an underestimation of the elasticity of bilateral trade with respect to distance. Consider Figure 2, which contains distance d_{ij} on the horizontal axis and imports m_{ij} on the vertical axis. We depict by circles imports to country i from countries $j = 1, 2, 3, 4, 5$, holding the control variables constant over countries j for the purpose of graphical illustration. From the selection equation for the probability of strictly positive exports (2.16), we note that $\frac{\partial \Pr(T_{ij}=1|\cdot)}{\partial d_{ij,}} = -\gamma^* \phi(\cdot) < 0$, where $\phi(\cdot)$ is the normal density function. Thus, missing observations

HMR below support this claim (see 3.3.1).

[11] This term is also known as Heckman's lambda (see Heckman, 1979).

[12] In the absence of a sample selection bias but in the presence of the omitted variable bias, the correction term would simplify to $\ln\left\{\exp\left(\delta\hat{z}^*_{ij}\right) - 1\right\}$, since $\text{plim}\hat{\eta}^*_{ij} = E[u_{ij}|., z^*_{ij} > 0] = 0$ in this case.

are more likely the larger is the distance. In addition, the smaller is the error term u_{ij}, the more likely is trade to be predicted to be zero. For this reason, we draw potential imports between countries i and $j = 4$ and $j = 5$ such that the distance is large and the error terms u_{i4} and u_{i5} negative, causing these two observations to drop out of the sample, which we indicate by hollow circles. Since the negative u_{i4} and u_{i5} are not only contained in the selection equation (2.16) but also in the gravity equation (2.10), the imports that drop out do not only occur at a large distance but also at unusually low values of imports.[13] The non-missing imports at large distances, indicated by filled circles, are those with positive values of u_{ij}, i.e. $E[u_{ij}|d_{ij}, T_{ij} = 1] > 0$ if the distance d_{ij} is large. Since the unconditional expected value of u_{ij} is zero by construction of the OLS estimator,[14] i.e. $E[u_{ij}|T_{ij} = 1] = 0$, the conditional expected value of u_{ij} is negative, $E[u_{ij}|d_{ij}, T_{ij} = 1] < 0$, if the distance d_{ij} is small. But then the error term in the outcome equation u_{ij} and distance d_{ij} are positively correlated. The conditional expected value of the error term feeds back into the estimated regression line, since

$$E[m_{ij}|d_{ij}, T_{ij} = 1] = E[m_{ij}|d_{ij}] + E[u_{ij}|d_{ij}, T_{ij} = 1], \tag{2.18}$$

for all observations with $T_{ij} = 1$,[15] where $E[m_{ij}|d_{ij}, T_{ij} = 1]$ is fitted by an OLS regression on the remaining three strictly positive import data from $j = 1, 2, 3$ and $E[m_{ij}|d_{ij}]$ is fitted by an OLS regression on the entire population including $j = 4$ and $j = 5$. This in turn, is asymptotically equivalent to an OLS regression with a sample selection correction denoted by *Heckman* in Figure 2. Hence, the positive slope of $E[u_{ij}|d_{ij}, T_{ij} = 1]$ in d_{ij} results in a flatter declining slope of an OLS regression without sample selection, $E[m_{ij}|d_{ij}, T_{ij} = 1]$, than one with sample selection, $E[m_{ij}|d_{ij}]$. This implies that the distance coefficient estimated by OLS, ignoring zero trade flows, will be too small in absolute values. Accounting for zero trade flows will aggravate the distance puzzle, as the estimates of the distance coefficient based on *Heckman* will be larger in absolute values. *To sum up: the selection bias due to omitting zero trade flows does not explain the distance puzzle.*

Omitted variable bias Next we show that the omitted variable bias leads to an overestimation of the elasticity of bilateral trade with respect to distance. We first need to understand how the omitted variable ω_{ij} is correlated with distance d_{ij}. This can easily be seen by inserting (2.15) into (2.13) and taking the expected value conditional on

[13] Note that we have drawn negative values of m_{ij}. Naturally, negative values of m_{ij} can never exist, but are generated by the gravity equation (2.10), since shocks are, by assumption, normally distributed on a range from $-\infty$ to $+\infty$. However, whenever m_{ij} is negative, it is not observed.

[14] The estimated regression constant will always ensure that the unconditional expected value of the error term is zero in an OLS regression, whereas the conditional expected value of the error term is only zero for a correctly specified model, i.e. a model without endogeneity problems.

[15] See, e.g., equation (16.34) in Cameron and Trivedi (2005, p. 549).

distance d_{ij} and the other control variables $z_0 \equiv \gamma_0 + \zeta_i + \xi_j - \kappa\phi_{ij}$ to obtain

$$E[\omega_{ij}\,|d_{ij}, z_0] = \int f_{\omega_{ij}}(\omega_{ij}) \ln\left[exp\left[\delta\left(z_0 - \gamma d_{ij} + \eta_{ij}\right)\right] - 1\right] d\omega_{ij} \equiv \Omega\left(z_0, d_{ij}\right), \qquad (2.19)$$

where $f_{\omega_{ij}}(\omega_{ij})$ is the marginal distribution function of ω_{ij}, and we take ω_{ij} to be conditionally independent of d_{ij} and z_0, i.e. we investigate the omitted variable bias after having properly controlled for the selection bias (such as by the Heckman correction factor) or, equivalently, considering the case where no trade flows are missing.[16] This has the purpose of comparing the Heckman estimator, which controls the selection bias but suffers from the omitted variable bias with the HMR estimator which controls for both biases. Controlling conceptually for the selection bias while analyzing the omitted variable bias implies that $e^{\delta(z_o - \gamma d_{ij} + \eta_{ij})} > 1$ to ensure that there are no missing observations causing selection bias. Moreover, $\Omega\left(z_o, d_{ij}\right)$ is the non-linear conditional expectation function, the shape of which is easy to analyze. Taking the derivative of (2.19) with respect to distance d_{ij}, we obtain

$$\frac{\partial E[\omega_{ij}\,|d_{ij}, z_0]}{\partial d_{ij}} = -\gamma\delta \int f_{\omega_{ij}}(\omega_{ij}) \frac{e^{\delta(z_0 - \gamma d_{ij} + \omega_{ij})}}{e^{\delta(z_0 - \gamma d_{ij} + \eta_{ij})} - 1} d\omega_{ij} < 0. \qquad (2.20)$$

Hence, there is a negative correlation between ω_{ij} and d_{ij}, because the share of exporting firms becomes smaller the larger is distance. If ω_{ij} is known (and other controls are kept constant), a linear OLS regression of import values m_{ij} on distance d_{ij} controlling for ω_{ij} is like a regression of $(m_{ij} - \omega_{ij})$ on distance d_{ij}. This follows from the fact that the regression coefficient of d_{ij} explains the remaining variation of the corresponding variable that is not at the same time common variation with another control variable (Frisch-Waugh theorem) and ω_{ij} enters the regression equation (2.10) with coefficient one. An OLS estimator fitting the regression line $E\left[m_{ij} - \omega_{ij} | d_{ij}\right]$ then has the same slope in d_{ij} as one fitting $E\left[m_{ij} | d_{ij}, \omega_{ij}\right]$ or, indeed, one using a consistent correction factor that controls for ω_{ij}, i.e. the *HMR* estimator (while at the same time controlling for the sample selection effect).

To obtain $(m_{ij} - \omega_{ij})$ in Figure 2, which is indicated by crosses, we can read off the difference between m_{ij} and $m_{ij} - \omega_{ij}$, an example of which is given for ω_{i1}. As can be seen, the crosses indicating $(m_{ij} - \omega_{ij})$ are systematically located below the circles indicating m_{ij} at low distances and above at large distances.[17] Hence, a fit of the crosses by the

[16] To see how this equation is obtained, note that by definition of a conditional expected value $E[\omega_{ij}\,|d_{ij}, z_0] = \int \hat{f}\left(\omega_{ij}\,|d_{ij}, z_0\right) \omega_{ij} d\omega_{ij}$, where $\hat{f}\left(\omega_{ij}\,|d_{ij}, z_0\right)$ is the conditional distribution of ω_{ij}. According to Greene (2012), (B-51), this can be written as $E[\omega_{ij}\,|d_{ij}, z_0] = \int f\left(\omega_{ij}\,|d_{ij}, z_0\right) \omega_{ij} d\omega_{ij}$ with $f\left(\omega_{ij}\,|d_{ij}, z_0\right)$ being the conditional distribution of ω_{ij}. If we then assume that ω_{ij} is conditionally independent of d_{ij} and z_0, we obtain from (B-60) $f\left(\omega_{ij}\,|d_{ij}, z_0\right) = f_{\omega_{ij}}(\omega_{ij})$, where $f_{\omega_{ij}}(\omega_{ij})$ is the marginal probability density (see B-45). Inserting this relation above, we obtain: $E[\omega_{ij}\,|d_{ij}, z_0] = \int f_{\omega_{ij}}(\omega_{ij}) \omega_{ij} d\omega_{ij}$. Inserting (2.15) and (2.13) into this relation yields (2.19).

[17] Note that the *HMR* regression line fits all crosses for $j = 1, 2, 3, 4, 5$, because it does not only correct

solid *HMR* line $E[m_{ij} - \omega_{ij}|d_{ij}]$ rather than the circles by the Heckman line $E[m_{ij}|d_{ij}]$ is flatter, implying an upward bias of the distance coefficient. Hence, the OLS estimator omitting ω_{ij} overestimates the elasticity of bilateral trade with respect to distance. This implies that the distance coefficient estimated by OLS, ignoring the omitted variable bias, will be too large in absolute values. Accounting for the omitted variable bias it is therefore possible to solve the distance puzzle, as the estimates of the distance coefficient based on *HMR* will be smaller in absolute values. *To sum up: the omitted variable bias due to neglecting the heterogeneity of firms can potentially explain the distance puzzle.*

Interacting the two biases Distance has three influences on imports. There is a direct one on positive trade flows through the intensive margin (and which is the only one present in homogeneous firm models), and an indirect one on positive trade flows through the extensive margin at the firm-level, i.e. the share of exporting firms ω_{ij}. A third margin is given by the extensive margin of positive overall trade flows. If omitting proper controls for the extensive margin ω_{ij} and for the selection of countries into positive trade flows, the distance coefficient captures all three margins.[18]

Overall, it is then indeterminate whether the OLS line $E[m_{ij}|d_{ij}, T_{ij} = 1]$ is flatter or steeper than the HMR line $E[m_{ij} - \omega_{ij}|d_{ij}]$. In anticipation of our empirical results, we have drawn it such that the OLS line is steeper than the HMR line, which implies that the omitted variable bias dominates the sample selection bias in levels. We depict this as *Bias OLS* in Figure 2.

Let us now consider both biases simultaneously, formally taking into account the interaction of the two biases. For this purpose, we need to draw on an approximation of (2.13),

$$\omega_{ij} \approx \delta z_{ij}, \tag{2.21}$$

where $\delta = \partial \omega_{ij} / \partial z_{ij}$ evaluated at the mean of z_{ij}.

We then have the following proposition:

Proposition 1. *When assuming that the HMR model is the data generating process, the OLS estimate of γ in (2.10) may then be (asymptotically) up- or downward biased, depending on whether the omitted variable bias from the share of exporting firms or the sample selection bias due to the omission of zero trade flows dominates, respectively.*

We derive the simultaneous bias term in the Appendix, which is given by the following

for the omitted variable bias, but also for sample selection simultaneously. If only the omitted variable bias was controlled for but not the sample selection bias, such a regression line would only fit the crosses corresponding to $j = 1, 2, 3$.

[18] Chaney (2008) and Krautheim (2011) theoretically derive the effects on the elasticity of bilateral trade flows with respect to distance disentangling the intensive and extensive margin.

simple expression:

$$Bias(\hat{\gamma}^{OLS}) = \gamma\delta - \Xi\left[\delta + \beta_{u\eta}\right]\bar{\eta}^{*}_{ij} \gtreqless 0, \tag{2.22}$$

where $\Xi = \sum_i \sum_j d_{ij} / \sum_i \sum_j (d_{ij})^2$.

Thus, as shown in Figure 2, the term $\gamma\delta > 0$ in (2.22) represents an upward bias in OLS (and Heckman) from not controlling for the number of exporting firms, and the last two terms measure a downward bias from sample selection in OLS, when omitting zero trade flows, as $\beta_{u\eta}$, $\bar{\eta}^{*}_{ij}$ and Ξ are positive.

2.2.2 Globalization

How would the bias of OLS evolve over time when globalization reduces the responsiveness of bilateral trade flows with respect to distance, due to new and better communication and transport technologies? Make the following assumption:

Assumption Increased globalization implies that $\frac{\partial\gamma}{\partial t} < 0$.

We then have the following proposition:

Proposition 2. *When assuming that the HMR model is the data generating process, both the downward bias from sample selection due to zero trade flows and the upward bias from omitting the number of exporting firms decrease in the pace of globalization.*

The change in the bias of the distance coefficient $\frac{\partial \text{Bias}(\hat{\gamma}^{OLS})}{\partial t}$ can once more be understood intuitively, looking at the two biases separately. Beginning with the change of the sample selection bias over time, we first notice that the bias depends on how the slope of $E[u_{ij}|d_{ij}]$ changes when γ changes over time. To understand this, we need to first look at how the selection process is influenced by a reduction in γ. An observation is missing whenever $z^{*}_{ij} < 0$ according to (2.16). Obviously, a reduction in γ decreases z^{*}_{ij} $\left(\partial z^{*}_{ij}/\partial\gamma = -d_{ij} < 0\right)$, where some missing trade links turn positive. Eventually, all missing trade links have turned into positive ones at sufficiently low γ. Hence, the true line fitting the data after globalization becomes flatter.

Turning to the change of the omitted variable bias over time, we once more need to understand how the slope of the conditional expectation function $\mathrm{E}[\omega_{ij}|d_{ij}]$ changes with a reduction of γ. For this purpose, it is sufficient to look at how ω_{ij} changes for each observation when γ falls. From (2.15) and (2.13), we immediately obtain

$$\frac{\partial\omega_{ij}}{\partial\gamma} = -d_{ij}\delta\frac{e^{\delta z_{ij}}}{e^{\delta z_{ij}} - 1} < 0, \tag{2.23}$$

for all ω_{ij} that are non-missing. Hence, the share of exporting firms of a country j exporting to country i is increasing for each country pair when γ falls. More importantly,

this share increases less for increasingly distant trading partners:

$$\frac{\partial^2 \omega_{ij}}{\partial \gamma \partial d_{ij}} = -\delta \frac{e^{\delta z_{ij}}}{e^{\delta z_{ij}} - 1} - d_{ij} \gamma \delta^2 \frac{e^{\delta z_{ij}}}{\left(e^{\delta z_{ij}} - 1\right)^2} < 0, \tag{2.24}$$

for all ω_{ij} that are nonmissing.

Since $\mathrm{E}_a[\omega_{ij}|d_{ij}]$ is flatter after globalization than $\mathrm{E}_b[\omega_{ij}|d_{ij}]$ is before globalization, the upward bias in the distance coefficient from omitting the variable ω_{ij} also becomes smaller.

Considering changes in both biases simultaneously, we cannot tell whether the difference in slopes between the HMR-line and the OLS line will increase or decrease over time, because the downward bias from sample selection decreases and the upward bias from the omitted variable ω_{ij} also decreases. Since we cannot tell how the bias of OLS will behave under globalization, the OLS estimate of the distance coefficient may also increase or decrease over time.

The HMR estimator and the distance puzzle Let us now show how the HMR estimator can be used to explain the distance puzzle. That is, let us now show how the HMR estimator can be used to explain the finding in the literature that the OLS estimates of the elasticity of bilateral trade with respect to distance do not fall over time.

Suppose that the omitted variable bias dominates in levels at the beginning of the data period such that there is an overall upward bias in the distance coefficient (see the estimates of Helpman, Melitz and Rubinstein (2008)), i.e. the OLS estimated schedule is steeper than the true line (HMR) just as in Figure 2. A decrease in the upward bias through the omitted variable ω_{ij} makes the OLS estimate flatter and a decrease in the downward bias through less sample selection makes the OLS schedule steeper. Now, if the downward bias from sample selection due to the omission of zero trade flows decreases faster than the upward bias from omitting the share of exporting firms, then, overall, the estimated OLS schedule will become steeper.

Note also that the sample selection bias alone cannot solve the distance puzzle if the HMR model is the data generating process, as was suggested by Felbermayr and Kohler (2006) without being specific about underlying data generating process. As the sample selection bias leads to a downward bias, the importance of distance will be underestimated. Hence, the level cannot be correctly captured accounting for sample selection alone. However, to capture the change of the bias in the distance coefficient, we need a larger decrease in the zero-trade flows bias as compared to the omitted variable bias due to firm heterogeneity. A first glance at the data and anecdotal evidence cope with these facts. Whereas there has been a dramatic decrease in zero-trade flows over the last two decades, firm sizes and productivities are still heavily dispersed (Poschke (2011)) and the share of exporting firms remains small.

3 Econometric analysis

3.1 Base-line estimation equation and alternative estimators

Our baseline estimation equation is the HMR gravity equation (2.17). Since our main interest rests on the coefficient of the distance variable γ and how it evolves over time, we will estimate this equation separately by year and industry. We use the following augmented specification:

$$m_{ij} = \beta_0 - \gamma d_{ij} + \alpha \mathbf{X}_{ij} + \lambda_j + \chi_i + \ln\left\{\exp\left[\delta\left(\hat{z}^*_{ij} + \hat{\eta}^*_{ij}\right) - 1\right]\right\} + \beta_{u\eta}\hat{\eta}^*_{ij} + e_{ij}, \quad (3.1)$$

where we explain the additional variables below. Once more, note that $\ln\left\{\exp\left[\delta\left(\hat{z}^*_{ij} + \hat{\eta}^*_{ij}\right) - 1\right]\right\}$ captures the omitted variable bias due to firm-level heterogeneity in the presence of sample selection, whereas $\hat{\eta}^*_{ij}$ captures the sample selection bias of the error term from estimating (3.1) for non-zero trade. To estimate these correction terms, we add a first-stage equation in order to estimate (2.16), where:

$$z^*_{ij} = \varphi^*_0 - \gamma^* d_{ij} + \vartheta^* \mathbf{X}_{ij} + \varphi^*_1 COMM\_REL_{ij} + \varphi^*_2 COMM\_LANG_{ij} + \xi^*_j + \zeta^*_i + \eta_{ij}. \quad (3.2)$$

3.1.1 Other estimators

We have shown that the distance puzzle can be studied by systematically comparing the estimates from HMR with corresponding estimates obtained with OLS. The OLS estimator estimates equation (3.1), omitting the correction terms for firm-level heterogeneity and sample selection, i.e. excluding $\ln\left\{\exp\left[\delta\left(\hat{z}^*_{ij} + \hat{\eta}^*_{ij}\right) - 1\right]\right\}$ and $\hat{\eta}^*_{ij}$. By comparing the HMR and OLS estimators, we can evaluate how the bias of OLS evolves over time as predicted by Propositions 1 and 2. We will also compare our estimates with HMR with a number of other estimators.

Heckman The usual Heckman estimator estimates equation (3.1) omitting the correction terms for firm-level heterogeneity but including that for sample selection, i.e. excluding $\ln\left\{\exp\left[\delta\left(\hat{z}^*_{ij} + \hat{\eta}^*_{ij}\right) - 1\right]\right\}$ but including $\hat{\eta}^*_{ij}$.

Linear approximation of HMR As δ enters the estimation equation non-linearly, we first estimate equation (3.1) via non-linear least squares, as proposed by HMR. However, as discussed in Santos Silva and Tenreyro (2008), this correction term is biased if their theoretical model is the data generating process. However, for a wide range of $\hat{z}^*_{ij} + \hat{\eta}^*_{ij}$, the term $\ln\left\{\exp\left[\delta\left(\hat{z}^*_{ij} + \hat{\eta}^*_{ij}\right) - 1\right]\right\}$ may be well approximated by $\overline{\delta}\left(\hat{z}^*_{ij} + \hat{\eta}^*_{ij}\right)$ for some appropriate parameter $\overline{\delta}$, which can be estimated by OLS (see our discussion in section

2.1). Hence, we also estimate the model via OLS and include $\varpi_{ij} = \overline{\delta}\left(\hat{z}^*_{ij} + \hat{\eta}^*_{ij}\right)$ instead of $\ln\left\{\exp\left[\delta\left(\hat{z}^*_{ij} + \hat{\eta}^*_{ij}\right) - 1\right]\right\}$.[19]

3.2 Data

The first of three data sets which we employ is borrowed from the original HMR paper (Helpman, Melitz and Rubinstein, 2008). Despite that HMR provide their main results for the year 1986, they also offer results for 1980s, adding year fixed effects to a panel. A comprehensive description of these data can be found in *Appendix I* in the HMR paper; the data are available at http://scholar.harvard.edu/helpman . The second data set is the standard CEPII gravity data set (available at www.cepii.fr), originally generated by Keith Head, Thierry Mayer and John Ries. A full description can be found in the appendix of Head, Mayer and Ries (2010). The CEPII data enables us to explore the distance coefficients a longer period as the original HMR data set. Although the CEPII data set already starts in the 1940s, – due to the number of observations – we use data from 1980 to 2006 which is the latest available year. Thirdly, we use an industry-level data set where imports are taken from Nicita and Olarreaga (2001), who have compiled an industry data set corresponding to the 3-digit ISIC, revision 2, level that contains 28 manufacturing industries for up to 100 countries during 1976-2004. Because there is a large number of missing values in the early years and we are lacking a control variable in the last year, we have restricted the sample to 1978-2003. This data set is available for downloading from the World Bank (www.worldbank.org\trade). In turn, this data set draws its bilateral industry import data from COMTRADE of the UN which is based on the Standard International Trade Classification (SITC) and then transformed into ISIC. Production data are taken from UNIDO (International Yearbook of Industrial Statistics).

3.2.1 Dependent variable

The dependent variable m_{ij} in (3.1) is the natural logarithm of bilateral imports of country i from country j at a given year t; for the industry-level data additionally in a given industry l, measured in million US$ converted by the Penn World Tables 6.0 purchasing power parity exchange rate (PPP) and deflated by the U.S. consumer price index.

3.2.2 Explanatory variables

The original HMR data set and the CEPII data set contain geographical information. The industry-level trade data set is merged into a balanced geography data set covering 170 countries. Thus, all three data sets contain geographical variables common to

[19] HMR use a polynomial of degree 3 in the score variable in one of their robustness checks. We will point out that even a linear approximation works well in practice.

gravity estimations. These geography variables appear in (3.1) and (3.2) and the different data sets as follows: common to all data sets, d_{ij} is the log of the distance between countries i and j. λ_j and χ_i are a full set of exporter and importer dummy variables, respectively, which control for, among others, the multilateral resistance terms pointed out by Anderson and van Wincoop (2003). $\mathbf{X}_{ij}$ contains a dummy variable indicating a common border between i and j in all data sets as well as a indicator whether there is a common trade agreement between exporter i and importer j. Dummy variables for a common legal system, a common colonial history, a currency union and bilateral membership within GATT/WTO are only available and included for the HMR and the CEPII data sets. Common island and landlock status indicators are included in the HMR and the industry-level data sets. All these variables are captured by $\mathbf{X}_{ij}$ in (3.1) and (3.2).

3.2.3 Exclusion restriction variables

To overcome the weak identification just through functional form, HMR propose at least three exclusion restriction variables for their procedure.

HMR prefer a specification where, in the first stage probit, a proxy variable of bilateral fixed export costs is employed. This variable–measuring the bilateral number of procedures needed to start exporting–might not influence the intensive margin but the probability of a positive trade flow. Since this variable does not cover a rich country sample they offer alternative exclusion restrictions. Beside the coverage issues of this variable, we suspect that the fixed exporting costs might change a lot over time. Therefore, using this variable which is, at best, available for periods after year 2000 would not fit our multi-period trade data sets starting in the seventies.

Alternatively HMR use the bilateral measures *common religion* and *common language* and do not find a qualitative difference in their results across any employed exclusion restrictions. The common religion variable measures to what extent the importer and the exporter share a common religion in the population according to data from the Christian Research Association for the year 2003. In particular, the measure takes the sum over the set of all existing religions summing up a population's share of the importer country confessing a religion multiplied with the same share of the exporter country. This measure is bounded between 0 and 1, with large numbers indicating a large degree of overlap in the religious structure of importer and exporter country. The second excluded variable indicates whether the importer and the exporter share a common language. Below we stick to this choice of exclusion restrictions and use the same control variables as in (3.1) (including the importer and exporter fixed effects) in addition to both excluded variables to estimate the probability of exporting in the first stage. We do so for all three data sets.

3.3 Results

To explore the distance puzzle, we thus estimate (3.1) for all three data sets by year and additionally by industry for the industry-level data set. With ten years from the original HMR data set, 27 years from the CEPII data set and data for 28 industries over 26 years from the industry-level data set and with four specifications respectively, this amounts to estimating 765 first-stage regressions and 3060 second-stage regressions. For expositional reasons, we show our results graphically.

3.3.1 HMR versus OLS

Figure 3 depicts distance coefficients estimated with OLS and the non-linear method from HMR for the original HMR data set. For each year, the distance coefficient is calculated, which is then plotted over the available time period from 1980-1989. To indicate the time pattern for each estimator, we have added a quadratic trend with an associated 95 percent confidence interval. Several interesting features are present in Figure 3.

Note that the trend of the distance coefficient, when estimated by OLS, $\hat{\gamma}^{OLS}$, is slightly *increasing* over time. This confirms the puzzling result in previous studies that the negative impact of distance on trade seems to increase rather then decrease over time, which would be expected from the globalization process. Turning to the HMR distance coefficient, $\hat{\gamma}^{HMR}$, we note that $\hat{\gamma}^{HMR}$ is indeed decreasing over time. Examining the bias of OLS, $\hat{\gamma}^{OLS} - \hat{\gamma}^{HMR}$, we note that this is positive. From Proposition 1, this is consistent with the upward bias from omitting the number of exporters dominating the selection bias from omitting zero trade flows. In addition, the bias grows over time. From theory, this suggests that globalization and reduced trade costs seem to decrease the downward bias from selection more than they reduce the upward bias from the number of exporters, see Proposition 2. Hence, the omitted variable bias seems to dominate the selection bias, and becomes relatively more important than the selection bias over time.

In Figure 4, we compare OLS with the linear approximation of HMR. We note that the results are qualitatively the same as in Figure 3: the HMR distance coefficient is decreasing over time, whereas the OLS coefficient increases with the associated bias of OLS increasing. Comparing Figures 3 and 4 we note that the linear approximation of HMR gives very similar results to the non-linear version of HMR. That the linear approximation of the HMR works satisfactorily is useful information for a future application of the linear approximation of the HMR methodology, given the cumbersome estimation of the non-linear version of HMR.

This main empirical finding holds for all three data sets as can be seen from Figures 5-8. Figures 6 show for the CEPII data qualitatively the same results as Figures 3 and 4 do for the original HMR data set. Again, we find this for the non-linear method of HMR and the linear approximation we propose. When we estimate (3.1) by year and industry

and then average the estimated distance by year, we find a very similar pattern shown in Figures 7-8.[20]

3.3.2 Heckman versus OLS

Next, we make a comparison by results obtained with the usual Heckman procedure. Since Heckman does not correct for the omitted variable bias, but the sample selection, we expect it's estimated distance coefficients to be larger in absolute values than those from OLS. This is exactly what our results in Figure 9 for the original HMR data depict: the estimated distance coefficients are bigger than those estimated from OLS in every single year in our data.[21] This empirical finding is very much in line with our theoretical result that accounting for zero trade flows cannot solve the distance puzzle when HMR is the data generating process. The results for the CEPII data (Figure 10) and the averaged distance coefficients from the industry-level estimates (Figure 11) again support this theoretical result: we find no evidence for a reduction of estimated distance coefficients when accounting for sample selection from ignoring zero trade flows compared to OLS estimates. Figure 11 also shows bigger distance coefficients in every single year and an increasing trend for the Heckman estimates. The importance of zero trade flows seems to be less for the CEPII data set given that the Heckman estimates are very similar to the OLS results. This is reasonable since Head, Mayer and Ries (2010) fill up many zero trade flows which actually have not been zero while generating the CEPII data set (see appendix of Head, Mayer and Ries, 2010).

To sum up our results up until here, we do not find a qualitative difference between the three data sets. Some quantitative differences are quite reasonable since for example the results for the industry data are averaged over industries with equal weights.

3.3.3 Industries

Figures 12a-12d show changes over time in the level of distance coefficient for each of the 28 industries from HMR and OLS. Most industries show a similar pattern, where the distance coefficient with OLS is increasing over time and the HMR distance coefficient is decreasing over time, producing an increasing bias of the OLS estimates.[22] In particular, these patterns are present in industries that are characterized by intra-industry trade

[20] Note here that, although the linear approximation works best for values of δ around 1 (see footnote 24), it still performs well for different values of correction factors.

[21] Note that we do not depict confidence intervals in Figures 9 and 10 since they are overlapping, which does not contradict our theoretical expectations.

[22] Actually, the bias can be identified visually from Figures 12a-12d. Therefore we added again quadratic fits and 95% confidence intervals over time to our estimates. We mostly observe an increase in the difference between the quadratic fit of the OLS estimates and the quadratic fit of the HMR estimates over time, at least for the second half of our data period. Note that this difference is always significant since an overlap of the confidence intervals between the two fits is not indicated at any figure and never converges to the end of our data period, except for "petrolium refineries".

(e.g. "Footwear" or "Manufacture of machinery"), whereas the patterns seem weaker in industries where the pattern of trade is to a larger extent explained by comparative advantage (e.g. "Tobacco manufactures" or "Petrolium refineries"). This is also what should be expected since trade in the HMR model generates intra-industry trade.

Descriptive evidence of these results shows Table 1 where the ISIC classification is linked to the industry classification with respect to product differentiation according to Rauch (1999) and the information of whether OLS bias increases or not. Rauch classifies industries at the SITC 4-digit level as differentiated or not. However, we first subsume these SITC 4-digit classification into our ISIC classification which actually aggregates the SITC 4-digit industries at a higher level, i.e. the ISIC codes consist of more than one SITC 4-digit code. We then calculate the share of differentiated SITC 4-digit industries according to Rauch (1999) within our 28 ISIC industries (*Share of differentiated industries*).

In Table 1 we do find a correlation between the dummy *Increase in bias* and *Share of differentiated industries* of 0.34.[23] The mean *Share of differentiated industries* within the 23 industries where we do find an increasing bias is 0.75 which is much higher than 0.40 within the 5 industries where we do not find an increase in the bias. If we draw an arbitrary cut-off for differentiated versus homogeneous industries at a *Share of differentiated industries* of 0.5 we would see that 17 out of 19 cases are differentiated according to the Rauch classification. Since the size of the SITC 4-digit industries is not accounted for when subsuming them into the ISIC classification we now concentrate on ISIC codes where we calculated a clear-cut *Share of differentiated industries* of either 0 or 1. Within these 15 observations we find 12 matches, either between no increase in the bias and a clear-cut *Share of differentiated industries* of 0 or between increase in the bias an a clear-cut *Share of differentiated industries* of 1.

3.3.4 Globalization and transport costs

Finally, we provide evidence that the HMR data generating process fits the data well and that (3.1) might consistently estimate the distance coefficient. Figures 13-18 show the results of relating the estimated distance coefficient $\hat{\gamma}^{HMR}$ to actual trade costs. Firstly, Figures 15 and 17 show that the estimated distance coefficients are strongly positively correlated with shipping costs in data recently published by Hummels (2007). Figure 13 does not support this finding, which we suspect to happen because of the low number of observations here. Secondly, Figures 14, 16 and 18 shows that the $\hat{\gamma}^{HMR}$ is also positively correlated with oil prices, which should be an important determinant of transport costs.

[23] However, left with 28 industries/observations, the regression results lack in their precision, but can serve as additional descriptives: Point estimates of regressions (probit, logit or linear probability) of the dummy which indicates a bias increase on *Share of differentiated industries* give results in our favor (positive) and are significant at the 10% level.

Additionally, we note that the OLS estimate of the distance coefficient is negatively correlated with these data on transport costs. Once more, this non-intuitive correlation can be explained because OLS neither controls for the omitted variable of the number of exporters nor for the omission of zero trade flows.

4 Conclusions

Globalization has advanced rapidly during the last two decades. In contrast, the influence of distance in empirical estimates of bilateral trade flows has remained high and has not declined. In this paper, we use the model by Helpman, Melitz and Rubinstein (2008), emphasizing zero trade flows and firm heterogeneity, to resolve this "distance puzzle".

Using different trade data sets, the non-linear estimation of HMR leads to declining distance coefficients over time. These coefficients also reflect the variation in "true trade costs" as the estimated HMR distance coefficients are also strongly correlated with the variation in freight costs and oil prices. When estimating the effect of distance on trade with OLS, we do not only find a larger distance coefficient but also that it increases over time. Thus, the distance puzzle arises from a growing bias of OLS estimates.

We show how the growing bias of OLS estimates can be explained from the two sources of bias generated from applying OLS to a gravity estimation when the HMR model is the data generating process. The upward bias of the OLS estimates implies that the omitted variable bias (from the number of heterogenous exporting firms) must dominate the sample selection bias (due to the omission of zero trade flows). When relating globalization to a fall of the true distance coefficient, both the downward bias from sample selection from omitting zero trade flows and the upward bias from omitting the number and size of exporting firms will decrease with increasing globalization (in absolute value). Since we find that the bias of OLS increases over time, the distance puzzle must arise because globalization had a weaker impact on the omitted variable bias from the number of heterogenous exporters.

On a final note, the gravity equation is perhaps the most widely used tool in empirical work using aggregate international trade data. While firm-level data is becoming more frequent, applying gravity equations on aggregate trade data will also remain common in the future when various policy issues are investigated. In this paper, we have shown how taking sample selection and exporter firm heterogeneity into account is crucial for understanding the effect of distance on international trade when aggregate trade data is used. Then, we showed the usefulness of a linear approximation of the HMR estimator. As this estimator is much simpler to apply than the non-linear estimator of HMR, we suggest that the linear approximation could be fruitfully used in many other research questions.

References

Anderson, J.E. and E. van Wincoop (2003), "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review* **93**(1), pp. 170-192.

Anderson, J.E. and E. van Wincoop (2004), "Trade Costs," *Journal of Economic Literature* **42**(3), pp. 691-751.

Baldwin, R. and J. Harrigan (2011), "Zeros, Quality and Space: Trade Theory and Trade Evidence," *American Economic Journal: Microeconomics* **3**(2), pp. 60–88

Bernard, A.B., J.B. Jensen, and P.K. Schott (2006), "Trade Costs, Firms and Productivity," *Journal of Monetary Economics* **53**(1), pp. 917–937.

Berthelon, M. and C. Freund (2008), "On the Conservation of Distance in International Trade," *Journal of International Economics* **75**(2), pp. 310-320.

Blum, B. and A. Goldfarb (2006), "Does the Internet Defy the Law of Gravity?", *Journal of International Economics* **70**(2), pp. 384–405.

Brun, J.-F., C. Carrère, P. Guillaumont and J. de Melo (2005), "Has Distance Died? Evidence from a Panel Gravity Model," *World Bank Economic Review* **19**(1), pp. 99-120.

Buch, C. M., Kleinert, J. and Toubal, F. (2004), "The Distance Puzzle: On the Interpretation of the Distance Coefficient in Gravity Equations," *Economics Letters* **83**(3), pp. 293-298

Cameron, A.C. and P.K. Trivedi (2005), *Microeconometrics - Methods and Applications*, Cambridge University Press, Cambridge, United Kingdom.

Chaney, T. (2008), "Distorted Gravity: The Intensive and Extensive Margins of International Trade," *American Economic Review* **98**(4), pp. 1707–1721.

Disdier, A.-C. and K. Head (2008), "The Puzzling Persistence of the Distance Effect on Bilateral Trade," *Review of Economics and Statistics* **90**(1), pp. 37-48.

Dixit, A.K. and J.E. Stiglitz (1977), "Monopolistic Competition and Optimum Product Diversity," *American Economic Review* **67**(3), pp. 297-308.

Dreher, Axel, Noel Gaston and Pim Martens (2008), *Measuring Globalization – Gauging its Consequences*, Springer, New York, United States.

Duranton, G. and M. Storper (2008), "Rising Trade Costs? Agglomeration and Trade with Endogenous Transaction Costs," *Canadian Journal of Economics* **41**(1), pp. 292-319.

Felbermayr, G.J. and W. Kohler (2006), "Exploring the Intensive and Extensive Margins of World Trade," *Review of World Economics* **142**(4), pp. 642-674.

Flam, H. and H. Nordström (2010), "Gravity Estimation of the Intensive and Extensive Margin: An Alternative Procedure and Alternative Data," unpublished manuscript, available at: http://www-2.iies.su.se/∼flamh/workingpapers.html.

Glaeser, E.L. and J.E. Kohlhase (2004), "Cities, Regions and the Decline of Transport Costs," *Papers in Regional Science* **83**(1), pp. 197–228.

Greene, W. (2012), *Econometric Analysis*, 7$^{\text{th}}$ edition, Pearson, Upper Saddle River, United States.

Grossman, G.M. (1998), "Comment" (pp. 29–31), in J. A. Frankel (Ed.), *The Regionalization of the World Economy*, Chicago, University of Chicago Press, National Bureau of Economic Research project report, 1998.

Head, K., T. Mayer and J. Ries (2010), "The Erosion of Colonial Trade Linkages after Independence," *Journal of International Economics* **81**(1), pp. 1-114.

Heckman, J.J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica* **47**(1), pp. 153-161.

Helpman, E., M. Melitz, Y. Rubinstein (2008), "Estimating Trade Flows: Trading Partners and Trading Volumes," *Quarterly Journal of Economics* **123**(2), pp. 441-487.

Hillberry, R. and D. Hummels (2008), "Trade Responses to Geographic Frictions: A Decomposition Using Micro-data," *European Economic Review* **52**(3), pp. 527-550.

Hummels, D. (2007), "Transportation Costs and International Trade in the Second Era of Globalization," *Journal of Economic Perspectives* **21**(3), pp. 131-154.

Hummels, D. and P.J. Klenow (2005), "The Variety and Quality of a Nation's Exports," *American Economic Review* **95**(3), pp. 704–723.

Krautheim, S. (2011), "Heterogeneous Firms, Exporter Networks and the Effect of Distance on International Trade,"*Journal of Internatinal Economics* **87**(1), pp. 27–35.

Krugman, P. (1980), "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review* **70**(5), pp. 950-959.

Lawless, M. and K. Whelan (2007), "A Note on Trade Costs and Distance," *Research Technical Paper of the Central Bank and Financial Services Authority of Ireland.*

Nicita, A. and M. Olarreaga (2001), "Trade and Production: 1976-1999," *World Bank Policy Research Working Paper No. 2701*, available at: http://www.worldbank.org\trade

Poschke, M. (2011), "The Firm Size Distribution across Countries and Skill-Biased Change in Entrepreneurial Technology," *CIREQ Working Paper*, 08-2011.

Rauch, J.E. (1999), "Networks Versus Markets in International Trade," *Journal of International Economics* **48**(1), pp. 7-35.

Sampford, M.R. (1953), "Some Inequalities on Mill's Ratio and Related Functions," *The Annals of Mathematical Statistics* **24**(1), pp. 130-132.

Santos Silva, J.M.C. and S. Tenreyro (2006), "The Log of Gravity," *Review of Economics and Statistics* **88**(4), pp. 641-658.

Santos Silva, J.M.C. and S. Tenreyro (2008), "Trading Partners and Trading Volumes: Implementing the Helpman-Melitz-Rubinstein Model Empirically," unpublished manuscript, available at: http://privatewww.essex.ac.uk/∼jmcss/research.html?.

World Trade Report (2008), *Trade in a Globalizing World*, World Trade Organization.

Yotov, Y.(2012), "A Simple Solution to the Distance Puzzle in International Trade," *Economic Letters* **117**(3), pp. 794-798.

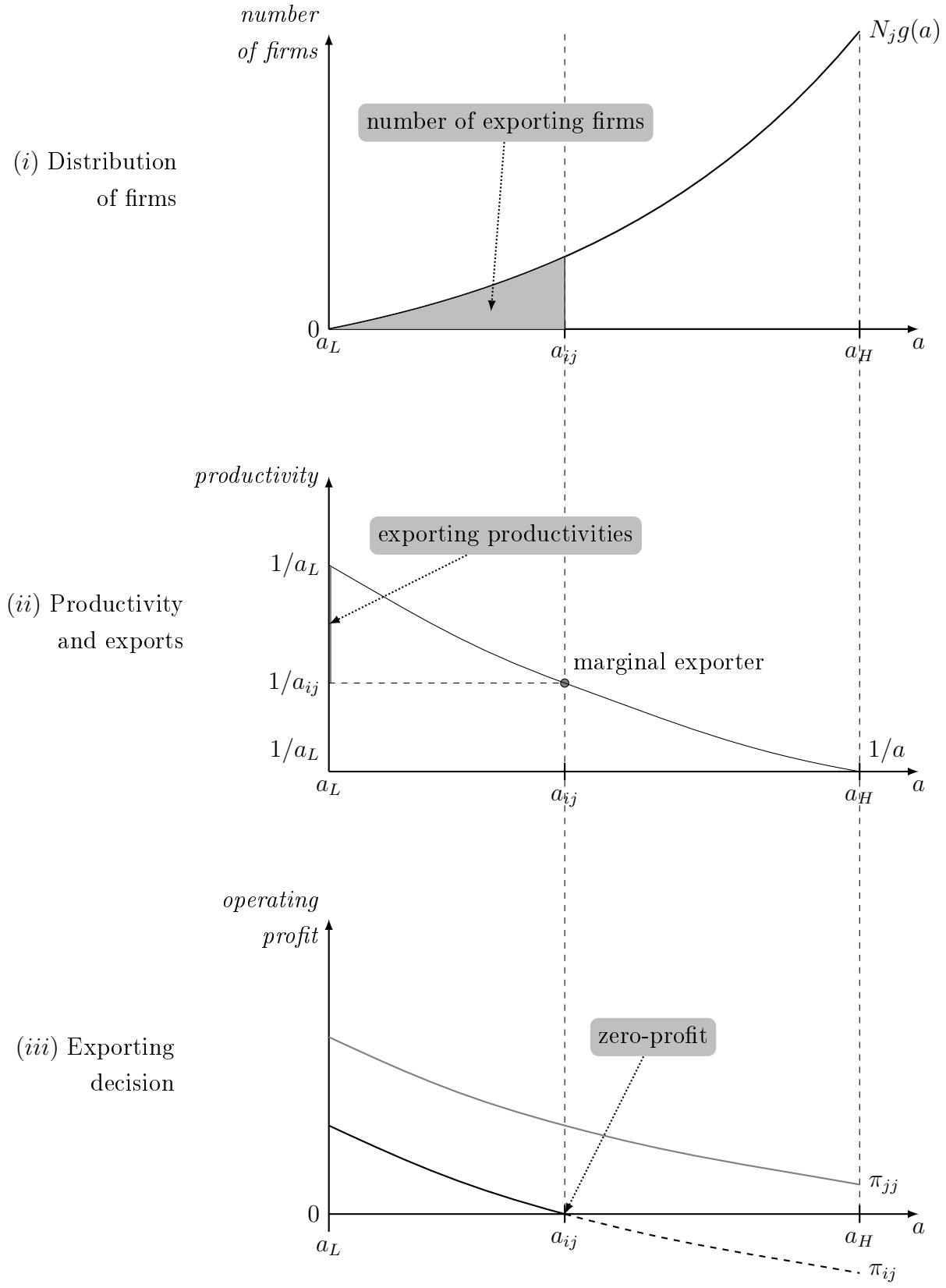

Figure 1: The HMR Model, where (i) illustrates the number of exporting firms of a Pareto distribution $N_j g(a)$ with productivity higher than marginal exporter's productivity $1/a_{ij}$, which is illustrated in (ii), while (iii) shows the zero-profit condition for export sales π_{ij} and the home sales π_{jj} function over the firm-specific input coefficient.

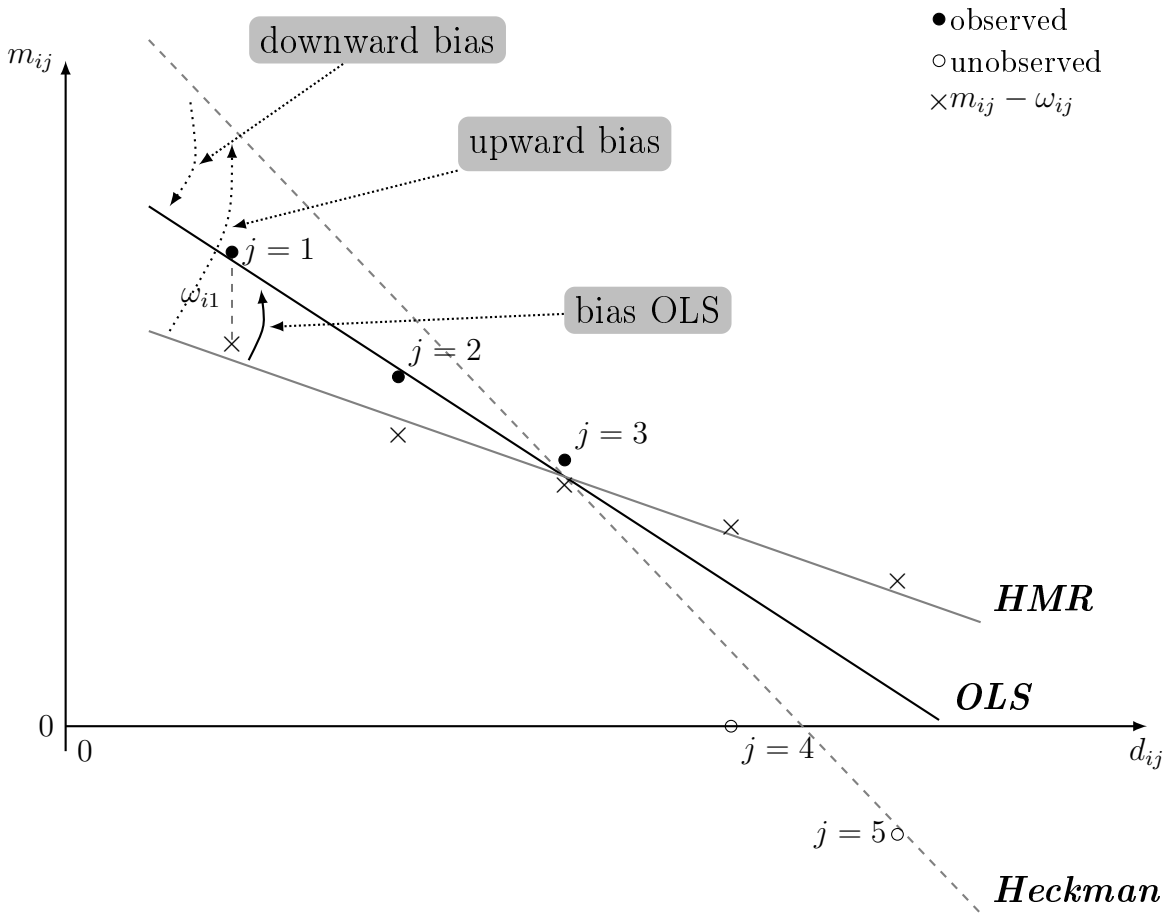

Figure 2: Illustrating the bias of OLS, where d_{ij} denotes distance between i and j and m_{ij} imports from j to i. ***HMR*** is given by $E[m_{ij} - \omega_{ij}|d_{ij}]$ with distance coefficient γ. ***Heckman*** is given by $E[m_{ij}|d_{ij}]$ and ***OLS*** by $E[m_{ij}|d_{ij}, T_{ij} = 1]$. The *bias OLS* corresponds to $Bias(\hat{\gamma}^{OLS}) = \gamma\delta - \Xi[\delta + \beta_{u\eta}]\,\bar{\eta}^*_{ij}$, where $(\gamma\delta)$ is denoted by *upward bias* and $(-\Xi[\delta + \beta_{u\eta}]\,\bar{\eta}^*_{ij})$ by *downward bias* in the figure. ω_{ij} controls for the omitted variable due to firm heterogeneity.

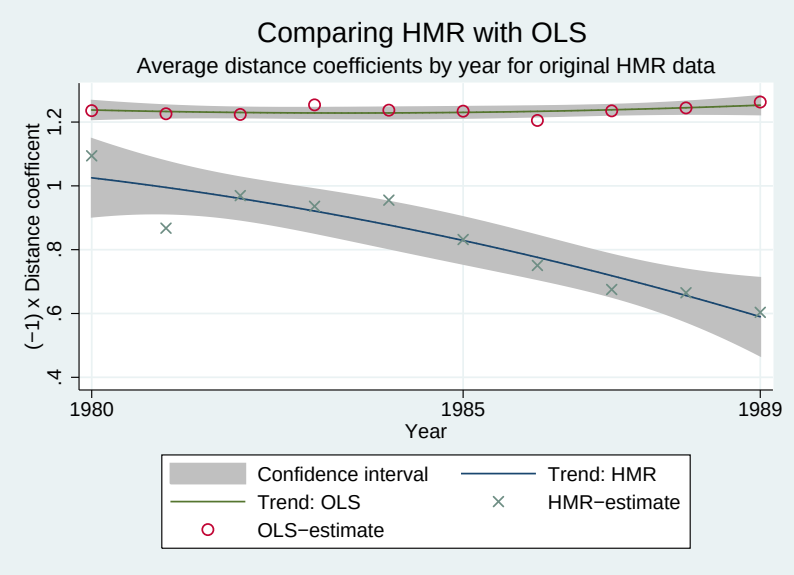

Figure 3: Comparing estimates of HMR with OLS for original HMR data.

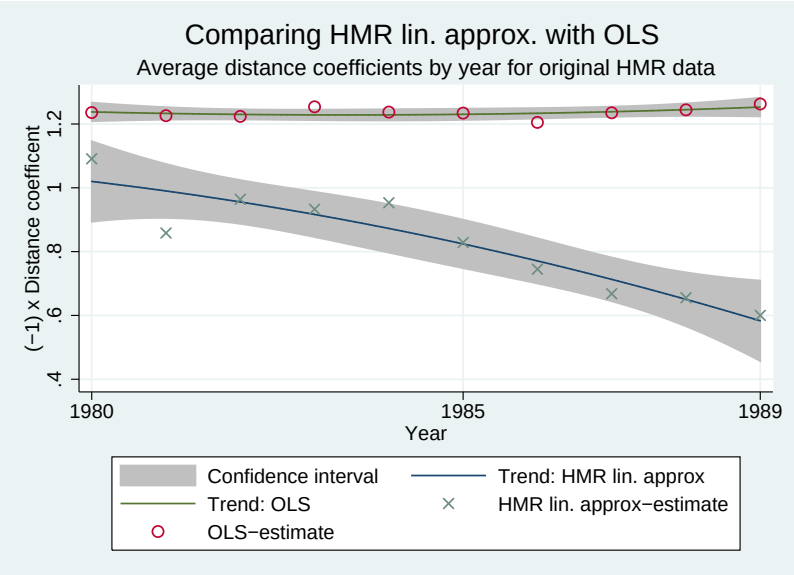

Figure 4: Comparing estimates of linear approximation of HMR with OLS for original HMR data.

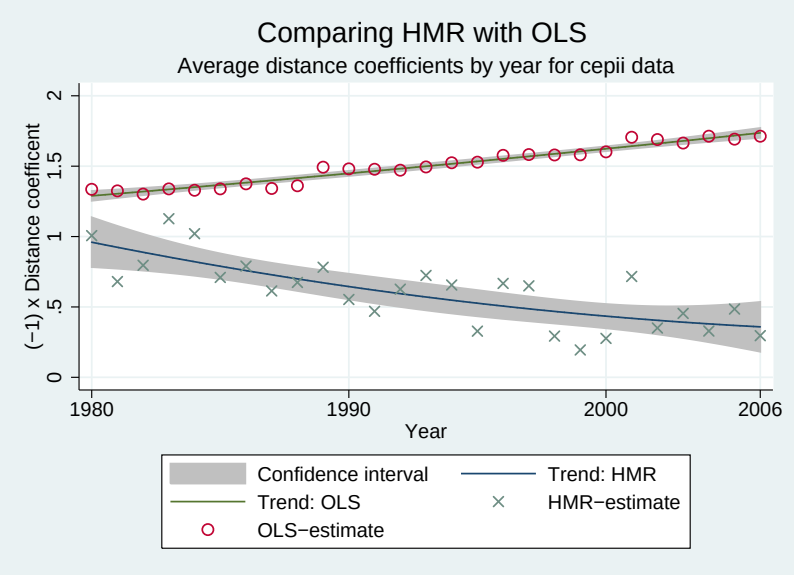

Figure 5: Comparing estimates of HMR with OLS for CEPII data.

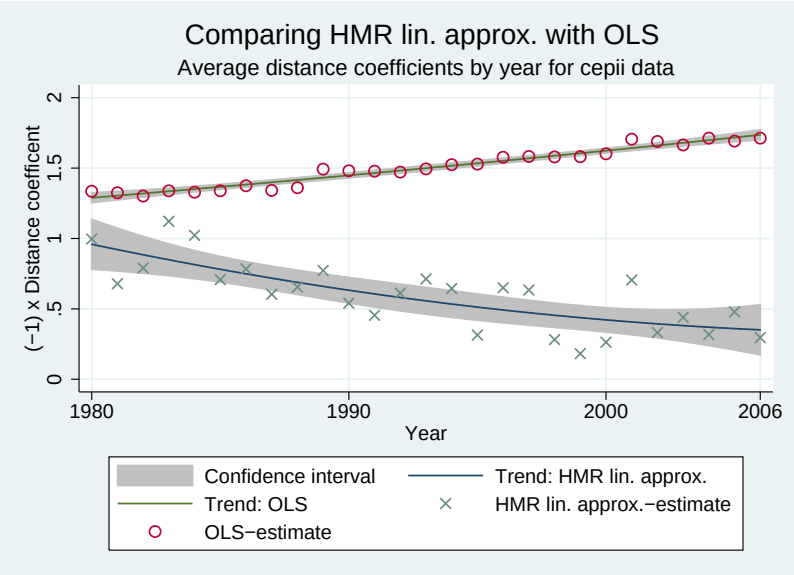

Figure 6: Comparing estimates of linear approximation of HMR with OLS for CEPII data.

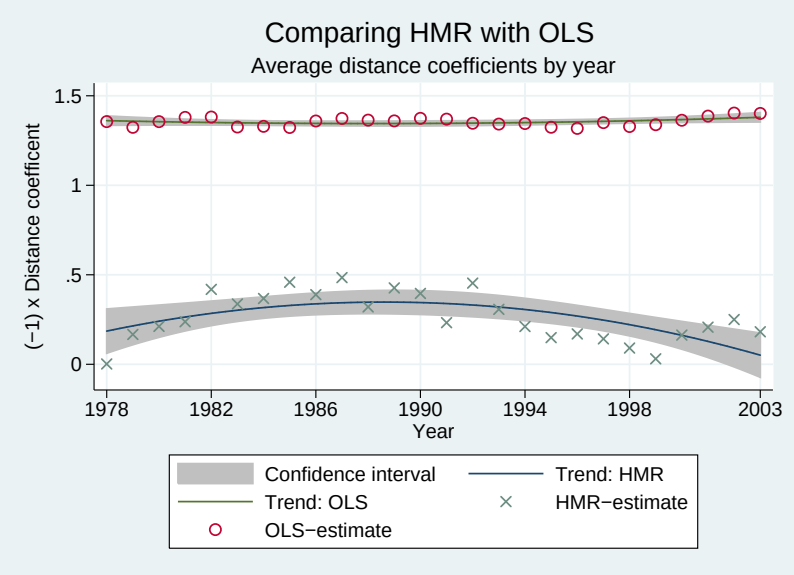

Figure 7: Comparing estimates of HMR with OLS for industry-level data (averaged).

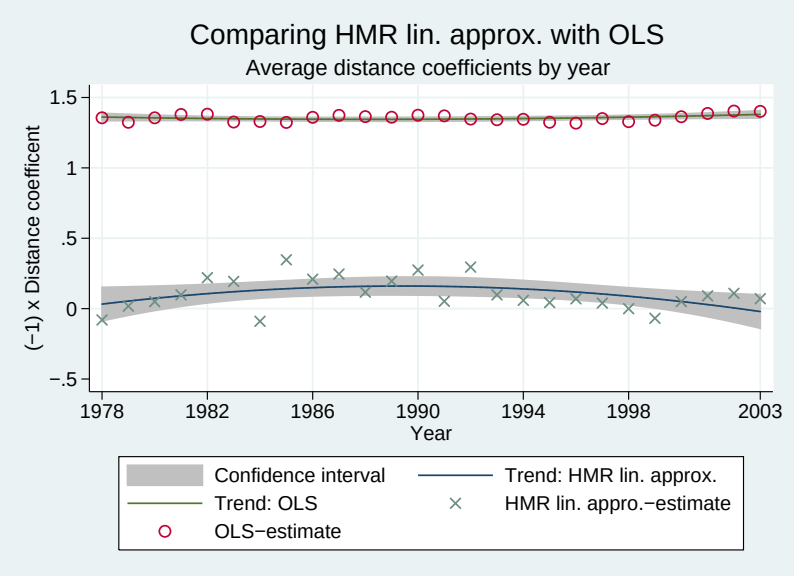

Figure 8: Comparing estimates of linear approximation of HMR with OLS for industry-level data (averaged).

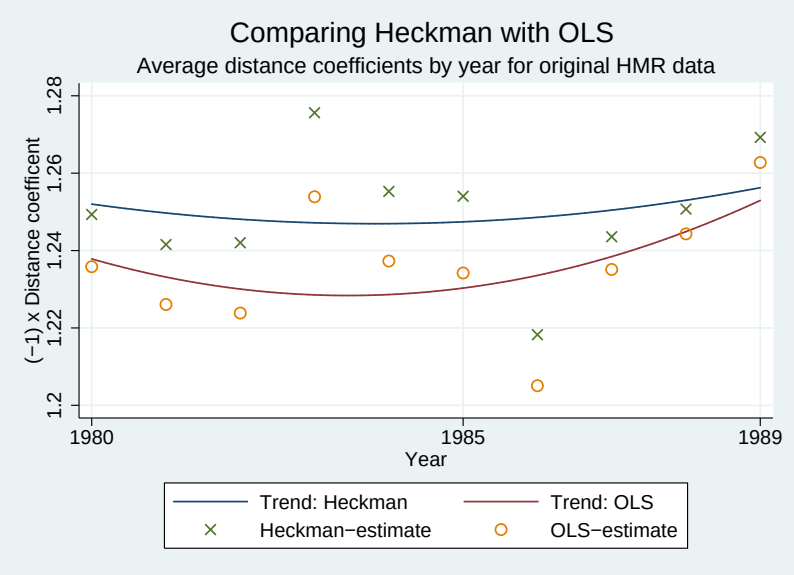

Figure 9: Comparing estimates of Heckman with OLS for original HMR data.

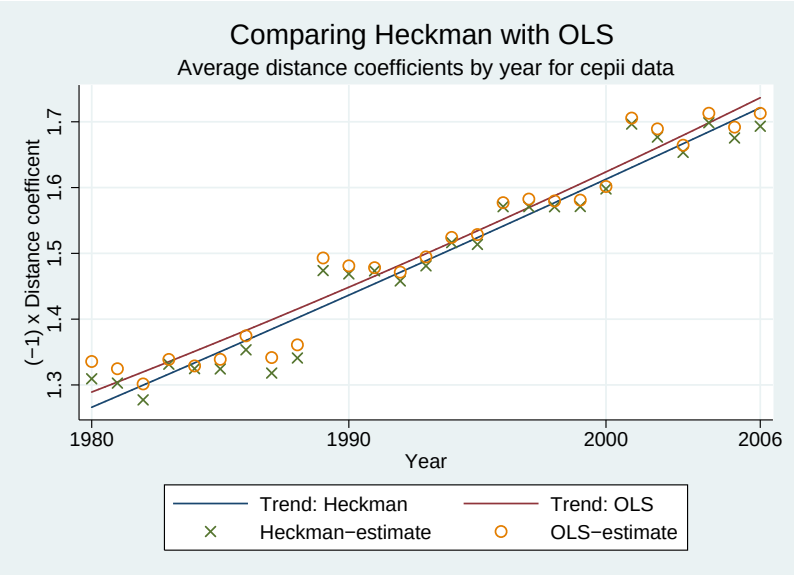

Figure 10: Comparing estimates of Heckman with OLS for CEPII data.

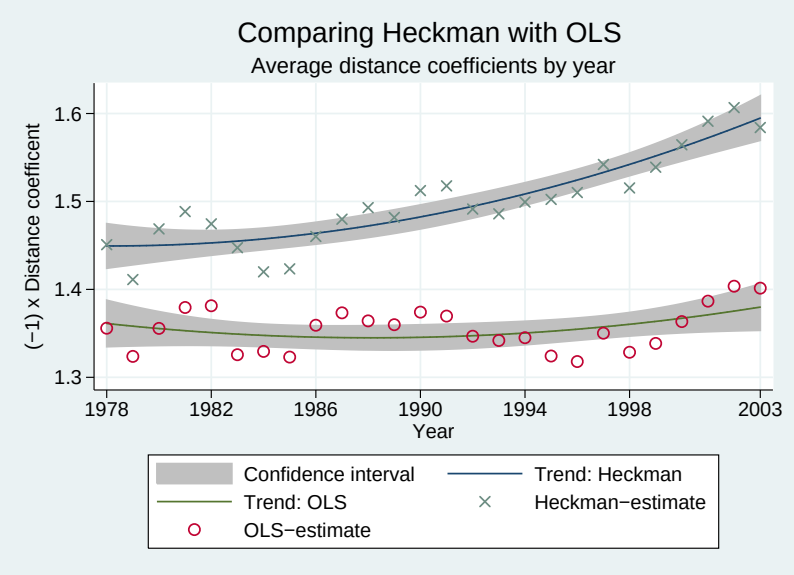

Figure 11: Comparing estimates of Heckman with OLS for industry-level data (averaged).

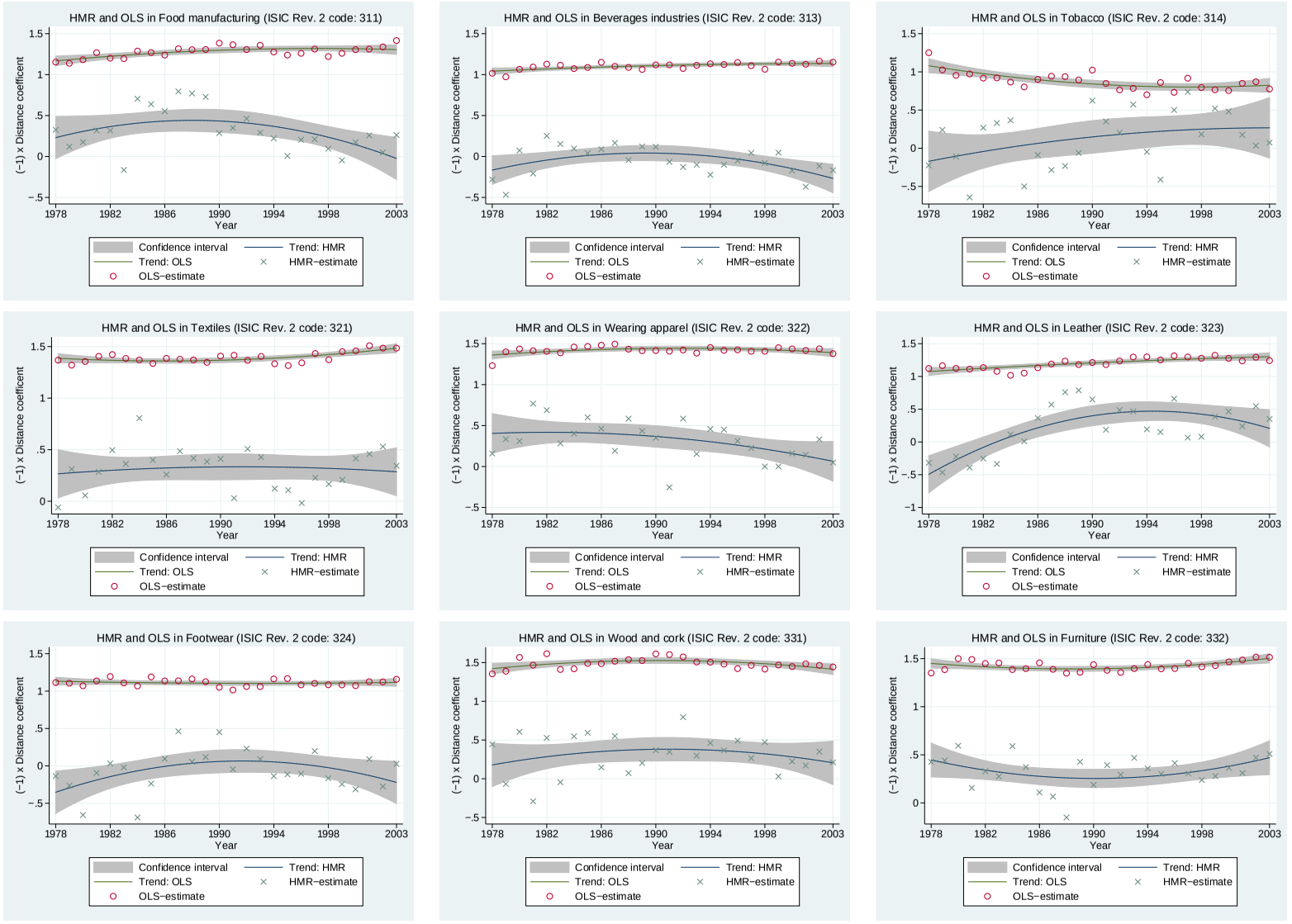

Figure 12a: Comparing estimates of HMR with OLS for different ISIC Rev. 2 industries.

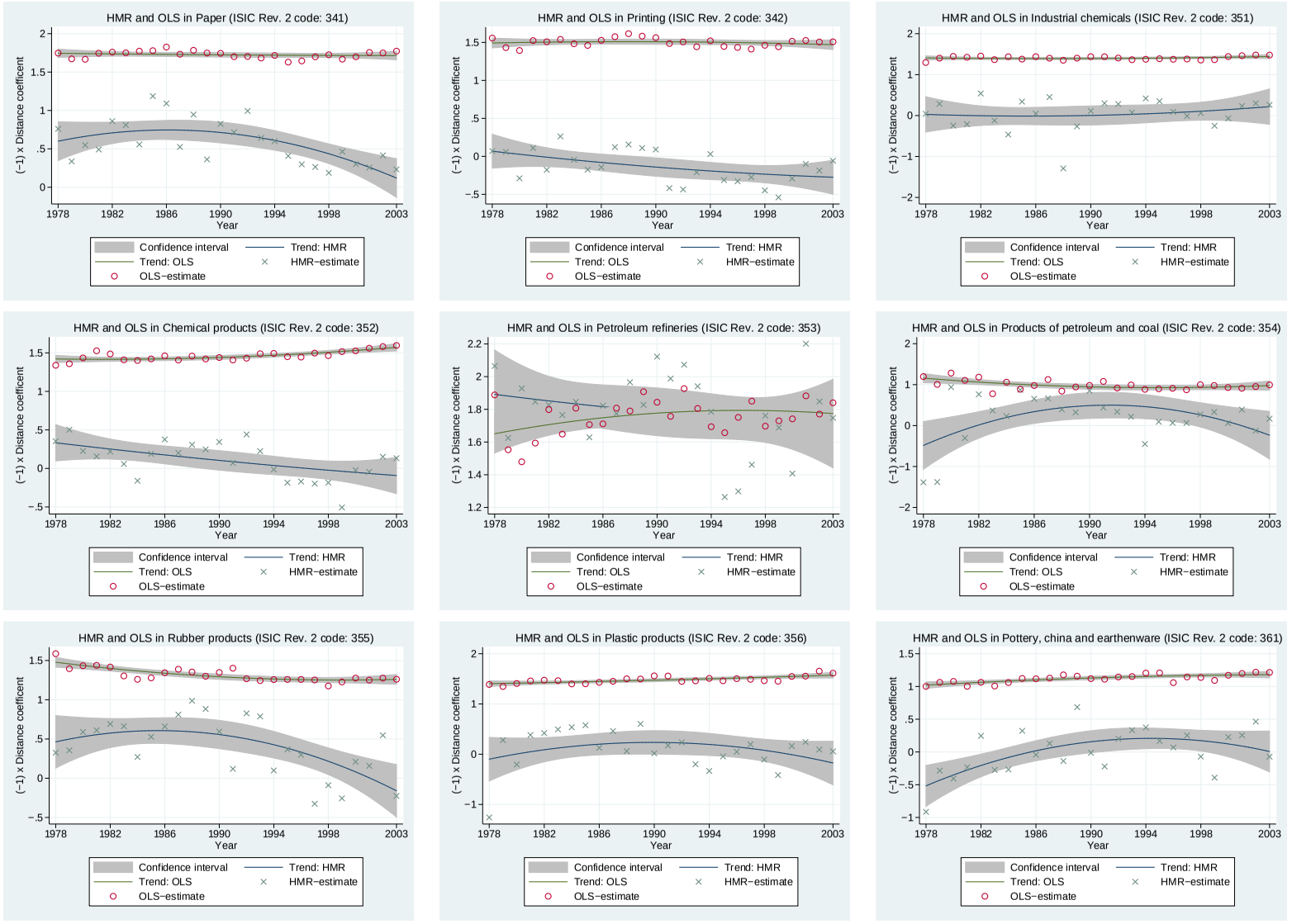

Figure 12b: Comparing estimates of HMR with OLS for different ISIC Rev. 2 industries.

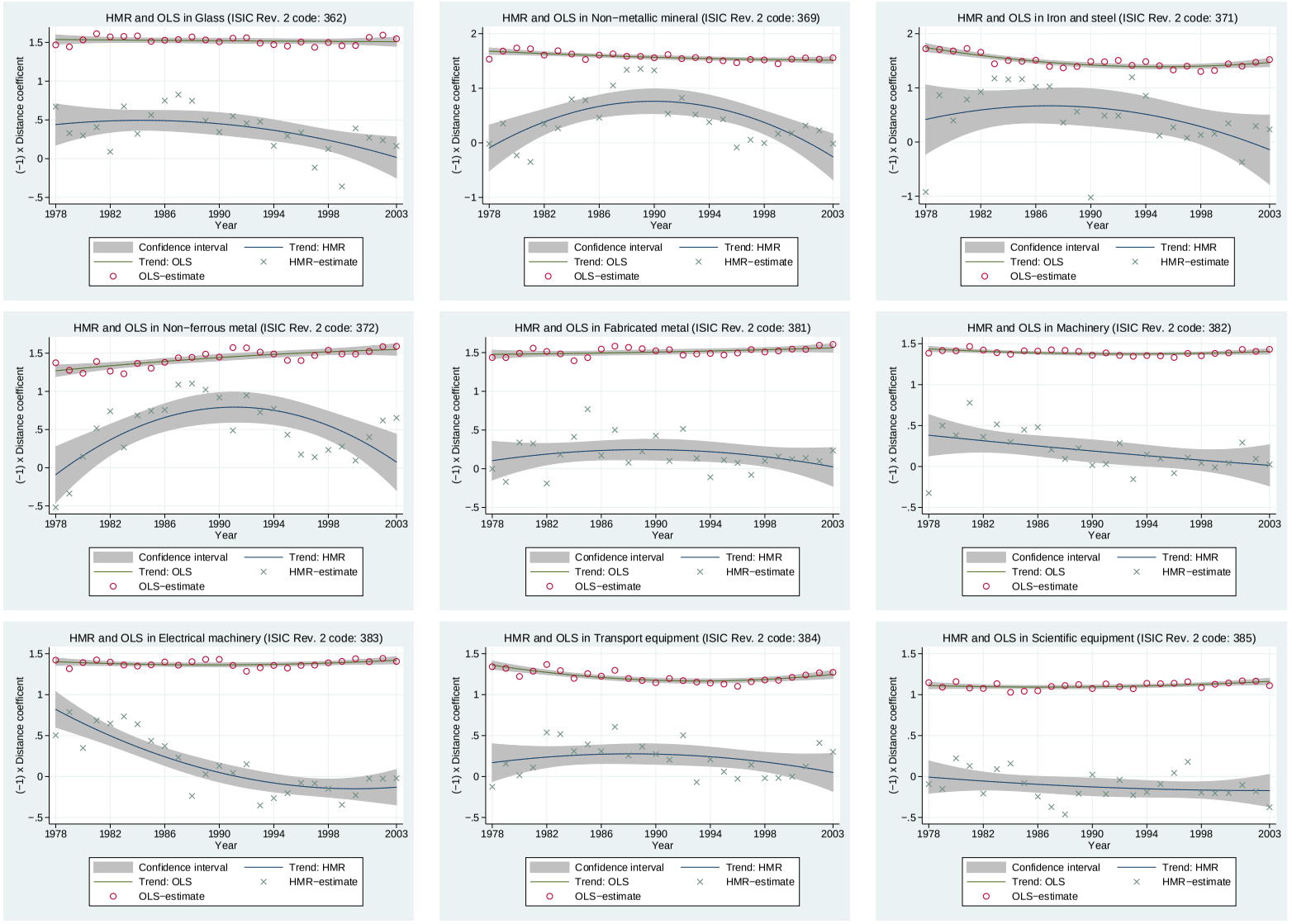

Figure 12c: Comparing estimates of HMR with OLS for different ISIC Rev. 2 industries.

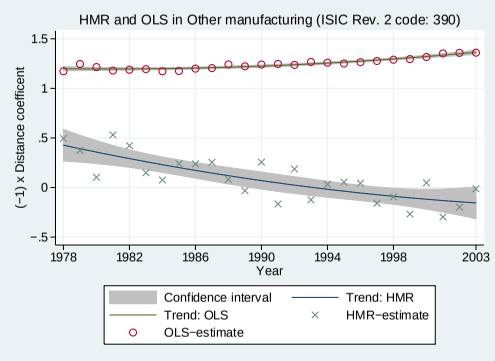

Figure 12d: Comparing estimates of HMR with OLS for different ISIC Rev. 2 industries.

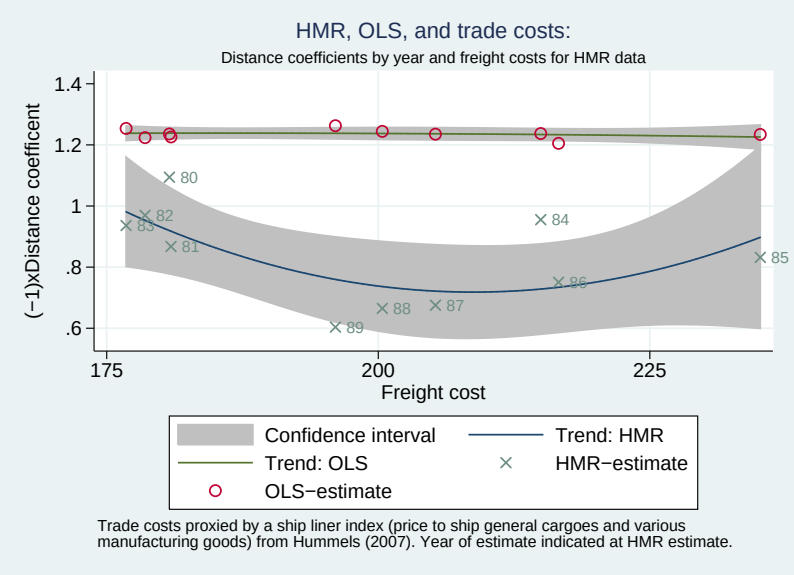

Figure 13: HMR, OLS and freight costs for original HMR data.

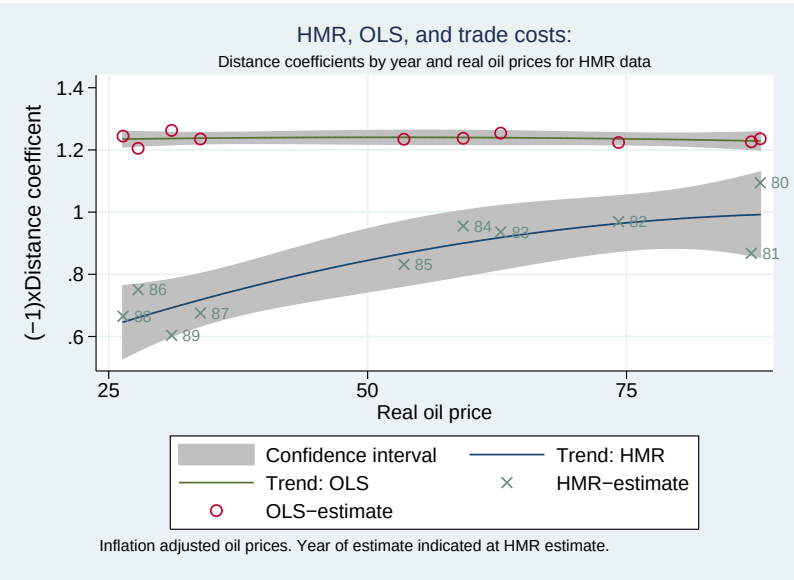

Figure 14: HMR, OLS and oil prices for original HMR data.

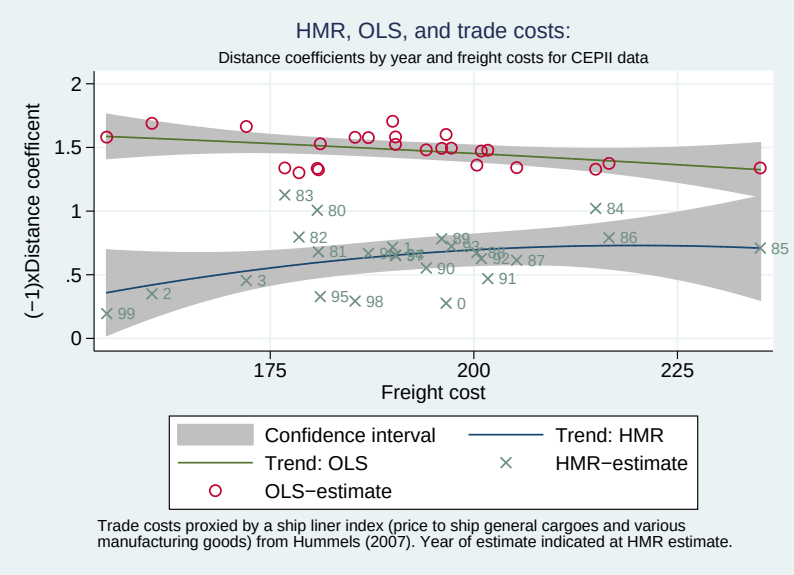

Figure 15: HMR, OLS and freight costs for CEPII data.

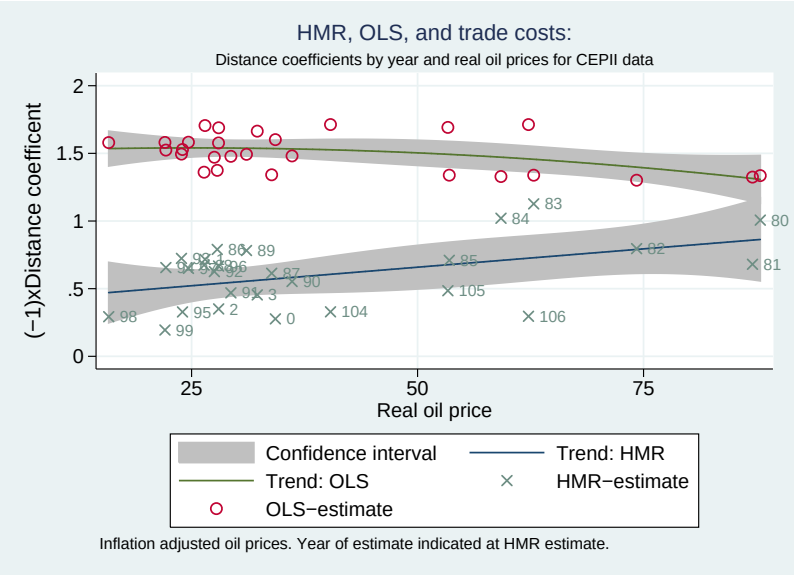

Figure 16: HMR, OLS and oil prices for CEPII data.

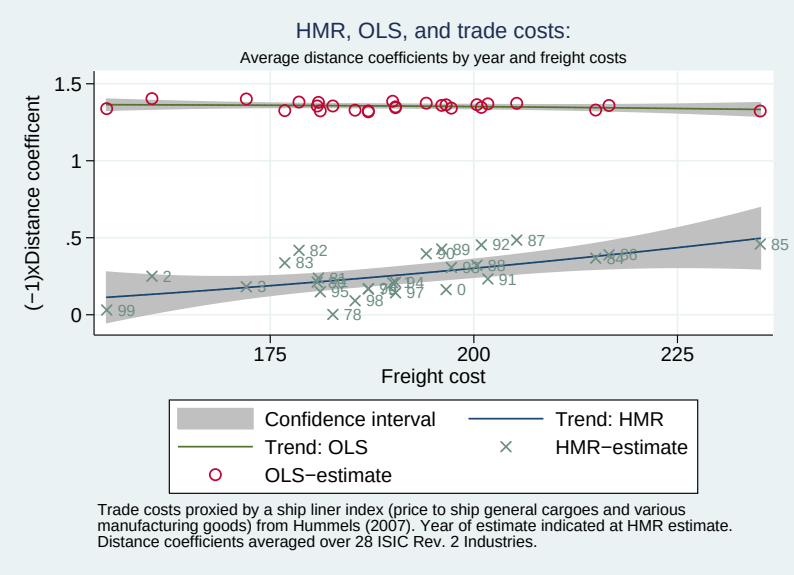

Figure 17: HMR, OLS and freight costs for industry-level data (averaged).

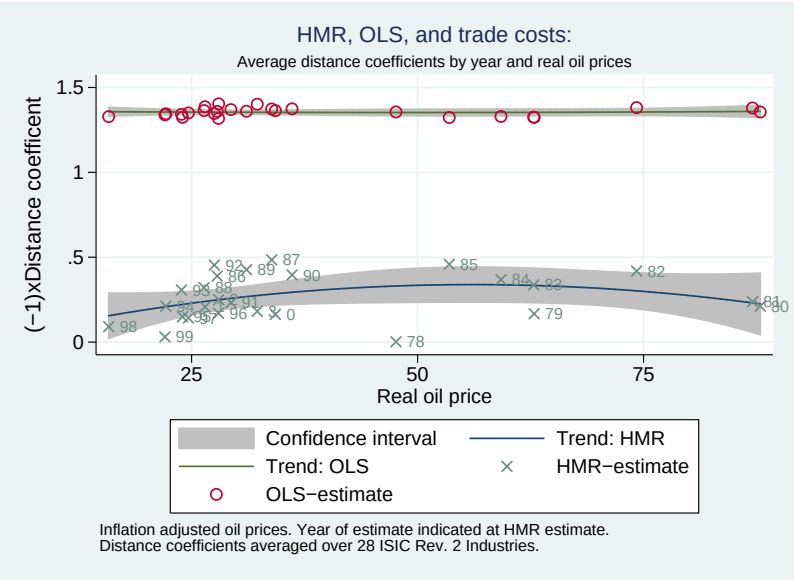

Figure 18: HMR, OLS and oil prices for industry-level data (averaged).

ISIC Code	Industry	Increase in bias	Share of differentiated industries
311	Food	yes	0.28
313	Beverage	yes	0.2
314	Tobacco	no	0
321	Textiles	yes	0.68
322	Wearing apparel	yes	0.96
323	Leather	yes	0.88
324	Footwear	yes	1
331	Wood and cork	no	0.7
332	Furniture	no	1
341	Paper	yes	0.18
342	Printing	yes	1
351	Industrial chemicals	no	0.19
352	Chemical products	yes	0.92
353	Petroleum refineries	no	0.13
354	Products of petroleum and coal	yes	0
355	Rubber products	yes	1
356	Plastic products	yes	1
361	Pottery, china and earthenware	yes	1
362	Glass	yes	1
369	Non-metallic mineral	yes	0.8
371	Iron and steel	yes	0.33
372	Non-ferrous metal	yes	0
381	Fabricated metal	yes	1
382	Machinery	yes	1
383	Electrical machinery	yes	1
384	Transport equipment	yes	1
385	Scientific equipment	yes	1
390	Other manufacturing	yes	0.92

Table 1: 28 ISIC Rev. 2 manufacturing industries, where *yes* corresponds to a dummy which is equal to 1 if we do find an increase in the bias $\hat{\gamma}^{OLS} - \hat{\gamma}^{HMR}$ from Figures 12a-12d. *Share of differentiated industries* is the share of differentiated SITC 4-digit industries according to Rauch (1999) within the ISIC industry.

A Appendix

A.1 Proof of Proposition 1

Proof: Rewrite (2.10) as $m_{ij} = \beta' \mathbf{X}_{ij} + \omega_{ij} + u_{ij}$ and z_{ij}^* in (2.16) as $z_{ij}^* = \varphi^{*\prime} \mathbf{X}_{ij} + \eta_{ij}^*$ where $\beta \equiv (\beta_0, \lambda_j, \chi_i, -\gamma)'$, and $\varphi^* = \left(\gamma_0^*, \xi_j^*, \zeta_i^*, -\gamma^*\right)', -\kappa^*$. Let, $\hat{\beta}^{OLS}$ denote the OLS estimator of β ignoring the sample selection and omitted variable corrections. We then obtain:

$$\mathrm{E}\left(\hat{\beta}^{OLS}\right) = \beta + \left[\mathbf{X}_{ij}\mathbf{X}_{ij}'\right]^{-1} \mathbf{X}_{ij} \mathrm{E}\left[\omega_{ij} + u_{ij} \left| z_{ij}^* > 0\right.\right], \tag{A.1}$$

where we have exploited that the $\mathbf{X}_{ij}$ variables contain only geography information and are therefore deterministic. To evaluate (A.1), examine the conditional expectations $\mathrm{E}\left[\omega_{ij} \left| z_{ij}^* > 0\right.\right]$ and $\mathrm{E}\left[u_{ij} \left| z_{ij}^* > 0\right.\right]$. Using formula (16.36) on p. 549 in Cameron and Trivedi (2005), we first obtain:

$$\begin{aligned} & E\left[u_{ij} \left| z_{ij}^* > 0\right.\right] = cov\left(u_{ij}, \eta_{ij}^*\right) \mathrm{E}\left[\eta_{ij}^* | \eta_{ij}^* > \varphi^{*\prime}\mathbf{X}_{ij} - \kappa^*\phi_{ij}\right] \qquad (A.2) \\ = \; & corr\left(u_{ij}, u_{ij} + \nu_{ij}\right) \frac{\sigma_u}{\sigma_\eta} \frac{\phi\left(\varphi^{*\prime}\mathbf{X}_{ij} - \kappa^*\phi_{ij}\right)}{\Phi\left(\varphi^{*\prime}\mathbf{X}_{ij} - \kappa^*\phi_{ij}\right)} \\ \equiv \; & \beta_{u\eta}\bar{\eta}_{ij} > 0, \end{aligned}$$

where $\beta_{u\eta} = corr\left(u_{ij}, u_{ij} + \nu_{ij}\right)\sigma_u/\sigma_\eta$ and $\bar{\eta}_{ij} = \frac{\phi(\varphi^{*\prime}\mathbf{X}_{ij} - \kappa^*\phi_{ij})}{\Phi(\varphi^{*\prime}\mathbf{X}_{ij} - \kappa^*\phi_{ij})}$. Further, we have assumed that u_{ij} and η_{ij}^* are bivariate normally distributed. Note that this implies that $u_{ij} = cov(u_{ij}, \eta_{ij}^*)\ \eta_{ij}^*/\sigma_\eta^2 + \varrho_{ij}$, where ϱ_{ij} is independent of η_{ij}^* and has zero mean. Hence, $\mathrm{E}\left[u_{ij} \left| \eta_{ij}^* > \varphi^{*\prime}\mathbf{X}_{ij} - \kappa^*\phi_{ij}\right.\right] = cov(u_{ij}, \eta_{ij}^*)\ 1/\sigma_\eta^2 \mathrm{E}\left[\eta_{ij}^* \left| \eta_{ij}^* > \varphi^{*\prime}\mathbf{X}_{ij} - \kappa^*\phi_{ij}\right.\right]$ and $cov\left(u_{ij}, \eta_{ij}^*\right) = corr\left(u_{ij}, u_{ij} + \nu_{ij}\right)\sigma_u\sigma_\eta$. To proceed, use a linear approximation of $\omega_{ij} = \ln\left[\left(Z_{ij}^*\right)^\delta - 1\right]$ for $z_{ij}^* > 0$. We can then write $\omega_{ij} = \ln\left[\left(Z_{ij}^*\right)^\delta - 1\right] = \ln\left[\exp\left(\delta z_{ijl}^*\right) - 1\right] \approx \delta z_{ij}^* > 0$, where $\delta = \sigma_\eta \frac{k-\varepsilon+1}{\varepsilon-1}$ is defined as above.[24] We then obtain:

$$\begin{aligned} & \mathrm{E}\left[\omega_{ij} \left| z_{ij}^* > 0\right.\right], \qquad (A.3) \\ = \; & \mathrm{E}\left[\delta z_{ij}^* \left| z_{ij}^* > 0\right.\right] = \delta \mathrm{E}\left[\left\{\mathrm{E}\left[z_{ij}^* \left| \mathbf{X}_{ij}\right.\right] + \eta_{ij}^*\right\} \left| z_{ij}^* > 0\right.\right] \\ = \; & \delta \mathrm{E}\left[z_{ij}^* \left| \mathbf{X}_{ij}\right.\right] + \delta \mathrm{E}\left[\eta_{ij}^* \left| z_{ij}^* > 0\right.\right], \\ = \; & \delta \mathrm{E}\left[\gamma_0^* + \xi_j^* + \zeta_i^* - \gamma^* d_{ij} - \kappa^*\phi_{ij} \left| \mathbf{X}_{ij}\right.\right] + \delta \mathrm{E}\left[\eta_{ij}^* \left| z_{ij}^* > 0\right.\right], \\ = \; & \delta\left[\gamma_0^* + \xi_j^* + \zeta_i^* - \gamma d_{ij}^* - \kappa^*\phi_{ij} + \bar{\eta}_{ij}^*\right], \\ = \; & \delta\varphi^{*\prime}\mathbf{X}_{ij} + \delta\bar{\eta}_{ij}^*. \end{aligned}$$

Noting that $\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X}\varphi^*\delta = \varphi^*\delta$, we obtain:

$$\mathrm{E}\left(\hat{\beta}^{OLS}\right) = \beta + \varphi^*\delta + \left[\mathbf{X}_{ij}\mathbf{X}_{ij}'\right]^{-1}\mathbf{X}_{ij}\delta\bar{\eta}_{ij}^* + \left[\mathbf{X}_{ij}\mathbf{X}_{ij}'\right]^{-1}\mathbf{X}_{ij}\beta_{u\eta}\bar{\eta}_{ij}^* \gtreqless 0. \tag{A.4}$$

[24] It can be shown that this approximation works very well in the range of ω_{ij} from $[0.5, \infty]$ and estimated values of δ around 1.

Since country dummies in $\mathbf{X}_{ij}$ are not correlated by construction and distance is hardly correlated with country dummies the matrix $\mathbf{X}'\mathbf{X}$ can be viewed as diagonal. But then:

$$\mathrm{E}\left(-\hat{\gamma}^{OLS}\right) = -\gamma - \gamma\delta + \frac{\sum_i \sum_j d_{ij}}{\sum_i \sum_j \left(d_{ij}\right)^2} \left[\delta + \beta_{u\eta}\right] \bar{\eta}^*_{ij}, \tag{A.5}$$

and hence

$$\mathrm{Bias}(\hat{\gamma}^{OLS}) = \gamma\delta - \frac{\sum_i \sum_j d_{ij}}{\sum_i \sum_j \left(d_{ij}\right)^2} \left[\delta + \beta_{u\eta}\right] \bar{\eta}^*_{ij}.$$

■

A.2 Proof of Proposition 2

Proof: From (2.22), we have $\mathrm{Bias}(\hat{\gamma}^{OLS}) = \gamma\delta - \Xi\left[\delta + \beta_{u\eta}\right]\bar{\eta}^*_{ij}$. Thus, it follows that $\frac{\partial \mathrm{Bias}(\hat{\gamma}^{OLS})}{\partial t} = \delta\frac{\partial\gamma}{\partial t} - \Xi\left[\delta + \beta_{u\eta}\right]\frac{\partial\bar{\eta}^*_{ij}}{\partial t}$. The change of the omitted variable bias over time is simply given by:

$$\frac{\partial\left(\delta\gamma\right)}{\partial t} = \delta\frac{\partial\gamma}{\partial t} < 0.$$

The sign of the change of the sample selection bias depends on the sign of

$$\begin{aligned}
\frac{\partial\bar{\eta}^*_{ij}}{\partial t} &= \frac{\partial\left(\frac{\phi\left(z^*_{ij}\right)}{\Phi\left(z^*_{ij}\right)}\right)}{\partial t} \\
&= \frac{1}{\Phi(z^*_{ij})^2}\left[\left(\phi'\left(z^*_{ij}\right)\cdot\Phi\left(z^*_{ij}\right) - \phi\left(z^*_{ij}\right)^2\right)\right]\frac{\partial z^*_{ij}}{\partial t} \\
&= \left[\frac{-z^*_{ij}\phi\left(z^*_{ij}\right)}{\Phi\left(z^*_{ij}\right)} - \left(\frac{\phi\left(z^*_{ij}\right)}{\Phi\left(z^*_{ij}\right)}\right)^2\right]\frac{\partial z^*_{ij}}{\partial t} \\
&= \left[-z^*_{ij}\bar{\eta}^*_{ij} - \left(\bar{\eta}^*_{ij}\right)^2\right]\frac{\partial z^*_{ij}}{\partial t} \\
&= -\bar{\eta}^*_{ij}\left[z^*_{ij} + \bar{\eta}^*_{ij}\right]\frac{\partial z^*_{ij}}{\partial t}.
\end{aligned} \tag{A.6}$$

Note that

$$\frac{\partial z^*_{ij}}{\partial t} = -d_{ij}\frac{\partial\gamma\left(t\right)}{\partial t} > 0.$$

The derivative of the mills ratio $\frac{\partial\bar{\eta}^*_{ij}}{\partial z^*_{ij}} = -\bar{\eta}^*_{ij}\left[z^*_{ij} + \bar{\eta}^*_{ij}\right]$ is negative. This can be shown by noting that

$$E\left[\eta^*_{ij}\left|\eta^*_{ij} > -\varphi'\mathbf{X}\right.\right] = \frac{\phi\left(\varphi'\mathbf{X}\right)}{\Phi\left(\varphi'\mathbf{X}\right)} = \frac{\phi\left(-\varphi'\mathbf{X}\right)}{1-\Phi\left(-\varphi'\mathbf{X}\right)}, \tag{A.7}$$

and using the result derived in Sampford (1953) and also given in Theorem 19.2 on page 876 in Greene (2012), that for $\phi\left(x\right)/\left(1-\Phi\left(x\right)\right)$ the derivative with respect to x is given

by

$$\frac{\phi(x)}{1-\Phi(x)}\left[\frac{\phi(x)}{1-\Phi(x)}-x\right], \tag{A.8}$$

and bounded between zero and one. Using the equality given in equation (A.7), we may write this as:

$$\frac{\phi(\varphi'\mathbf{X})}{\Phi(\varphi'\mathbf{X})}\left[\frac{\phi(\varphi'\mathbf{X})}{\Phi(\varphi'\mathbf{X})}+\varphi'\mathbf{X}\right]=\bar{\eta}_{ij}\left[z_{ij}+\bar{\eta}_{ij}\right]. \tag{A.9}$$

Hence, this expression differs from our derivative of $\bar{\eta}^*_{ij}$ only by the multiplication with -1. Hence, the derivative of $\bar{\eta}^*_{ij}$ with respect to z^*_{ij} is bounded between -1 and 0. But then $\frac{\partial\bar{\eta}^*_{ij}}{\partial t}=\partial\left[\phi\left(z^*_{ij}\right)/\Phi\left(z^*_{ij}\right)\right]/\partial t<0$. The change in the bias for OLS is therefore ambiguous, depending on whether the change in the sample selection bias or the change in the omitted variable bias is larger:

$$\frac{\partial\text{Bias}(\hat{\gamma}^{OLS})}{\partial t}=\delta\frac{\partial\gamma}{\partial t}-\Xi\left[\delta+\beta_{u\eta}\right]\frac{\partial\bar{\eta}^*_{ij}}{\partial t}\gtreqless 0. \tag{A.10}$$

■