Estimating the Extensive Margin of Trade*

J.M.C. Santos Silva† Silvana Tenreyro‡ Kehai Wei§

January 21, 2013

Abstract

Understanding and quantifying the determinants of the number of sectors or firms exporting in a given country is of relevance for the assessment of trade policies. Estimation of models for the number of sectors, however, poses a challenge because the dependent variable has both a lower and an upper bound, implying that the partial effects of the explanatory variables on the conditional mean of the dependent variable cannot be constant and must approach zero as the conditional mean approaches its bounds. We argue that ignoring these bounds can lead to erroneous conclusions due to the model's misspecification, and propose a flexible specification that accounts for the doubly-bounded nature of the dependent variable. We empirically investigate the problem and the proposed solution, finding significant differences between estimates obtained with the proposed estimator and those obtained with standard approaches.

JEL Classification codes: C13; C25; C51; F11, F14.

Key words: Bounded data; Estimation of trade models; Number of sectors.

---

*Santos Silva gratefully acknowledges partial financial support from Fundação para a Ciência e Tecnologia (FEDER/POCI 2010).

†University of Essex and CEMAPRE. Wivenhoe Park, Colchester CO1 1ED, United Kingdom. Fax: +44 (0)1206 872724. E-mail: jmcss@essex.ac.uk.

‡London School of Economics, CEP, and CEPR. Department of Economics, s.579. St. Clement's Building. Houghton St., London WC2A 2AE, United Kingdom. Fax: +44 (0)207 9556018. E-mail: s.tenreyro@lse.ac.uk.

§University of Essex. Wivenhoe Park, Colchester CO1 1ED, United Kingdom. Fax: +44 (0)1206 872724. E-mail: kwei@essex.ac.uk.

1. Introduction

In a landmark paper, Hummels and Klenow (2005) drew attention to the role of the extensive margin in explaining observed international trade patterns, giving origin to a burgeoning literature on its determinants and importance.[1]

Building on Melitz's (2003) model with heterogeneous firms, Helpman, Melitz, and Rubinstein (2008) and Chaney (2008), among others, developed trade models that explicitly consider the decision to export and therefore explicitly model the extensive margin of trade. In parallel, a large number of authors have studied empirically how the extensive margin is affected by factors such as transportation costs, tariffs, or economic and political integration.

The extensive margin can be defined at different levels of aggregation and a variety of definitions have been used in empirical work. For example, Hillberry and Hummels (2008) work at the shipment level, Eaton, Kortum, and Kramarz (2004), and Berthou and Fontagné (2008) work at the firm level, Hillberry and McDaniel (2002), Hummels and Klenow (2005), and Dennis and Shepherd (2007) define the extensive margin at the sector-product level, and Helpman, Melitz, and Rubinstein (2008) consider data at the country level.

Naturally, the econometric methods used in the estimation of models for the extensive margin of trade depend on the level of aggregation that is considered and on the nature of the data available. For example, Berthou and Fontagné (2008), Baldwin and Di Nino (2006), and Helpman, Melitz, and Rubinstein (2008) use binary models to study whether a firm, a sector, or a country exports, while Eaton, Kortum, and Kramarz

[1] The the number of sectors exporting in a country also informs on the degree of specialization of the export base and influences its response to sectoral shocks, affecting the volatility of the economy. For links between the number of sectors producing or exporting and volatility, see Greenwood and Jovanovic (1990), Acemoglu and Zilibotti (1997), Koren and Tenreyro (2007 and 2012), and di Giovanni and Levchenko (2009).

(2004), Hillberry and McDaniel (2002), Flam and Nordström (2006), and Dennis and Shepherd (2007) model the number of firms or sectors that export. While some of the models used in these studies are standard, the specification and estimation of models for the number of sectors exporting raises specific problems and is the focus of this paper.

The number of sectors exporting from origin country j to destination country i is a count variable and therefore it is a non-negative integer. Moreover, if the sectors or products are defined using a classification of economic activities such as the Harmonized Commodity Description and Coding System, the variate of interest has as an upper bound the number of classes in the system. That is, the variate of interest is bounded from below by zero and from above by the number of product categories.

The existence of these bounds implies that the partial effect of the regressors on the conditional mean of the dependent variable (the number of sectors) cannot be constant and must approach zero as the conditional mean approaches its bounds. Therefore, ignoring the nature of the data and simply using OLS, as in Flam and Nordström (2006), is likely to lead to erroneous conclusions because the linear model assumes that the partial effects are constant. Some authors have addressed the existence of the lower bound by using the log of the number of sectors as the dependent variable, see, e.g., Eaton, Kortum, and Kramarz (2004) and Hillberry and Hummels (2008).[2] Alternatively, standard count data models, such as Poisson and negative binomial regressions have been used by Dennis and Shepherd (2007), Berthou and Fontagné (2008), and Persson (2012). However, all these approaches ignore the upper bound and therefore are also unsatisfactory. Indeed, as we will illustrate, these estimators can be even less reliable than the simple linear model, leading to very misleading results.

In this paper we study the estimation of models for the number of sectors exporting from country j to country i. Building on Helpman, Melitz, and Rubinstein (2008) and on the literature on fractional data (see, e.g., Ramalho, Ramalho, and Murteira, 2011),

[2]Naturally, observations for which the number of sectors is equal to zero have to be dropped.

we suggest a flexible specification that takes into account the doubly-bounded nature of the data. In an empirical application we use country-pair data to estimate how different geographic and economic determinants of international trade affect the number of sectors exporting from j to i. The advantage of the proposed model over various alternatives previously used in the literature is clearly illustrated in this application.

2. The economic and statistical models

In the model considered by Helpman, Melitz, and Rubinstein (2008), hereinafter HMR, the operating profits for a firm of country j selling in country i are given by[3]

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

where a is the number of bundles of inputs needed for the firm to obtain one unit of product, c_j is the cost of each bundle in country j, P_i is the price index in country i, Y_i is the income in country i, f_{ij} is proportional to the fixed cost of exporting from j to i, τ_{ij} is the "melting iceberg" variable cost of exporting from j to i, and $\alpha \in (0,1)$ is a parameter such that $\varepsilon = 1/(1-\alpha)$ is the elasticity of substitution across products. The firm exports to market i if $\pi_{ij}(a) > 0$ or, equivalently, if

$$\frac{(1-\alpha)}{c_j f_{ij}}\left(\frac{\tau_{ij}c_j a}{\alpha P_i}\right)^{1-\varepsilon} Y_i > 1,$$

which, taking logs on both sides, leads to

$$\begin{aligned}
0 &< \ln(1-\alpha) - \ln c_j - \ln f_{ij} + \ln Y_i + (1-\varepsilon)(\ln \tau_{ij} + \ln c_j + \ln a - \ln \alpha - \ln P_i), \\
0 &< \theta + \varphi_i + \psi_j - \ln f_{ij} + \frac{\alpha}{\alpha-1}\ln \tau_{ij} + \frac{\alpha}{\alpha-1}\ln a, \\
\ln a &< \frac{1-\alpha}{\alpha}\left(\theta + \varphi_i + \psi_j - \ln f_{ij}\right) - \ln \tau_{ij},
\end{aligned}$$

where $\theta = \ln(\alpha^{\varepsilon-1}\varepsilon^{-1})$, $\varphi_i = \ln\left(Y_i P_i^{(\varepsilon-1)}\right)$, and $\psi_j = -\varepsilon \ln c_j$. Notice that c_j, f_{ij}, and τ_{ij} are assumed not to depend on the identity of the producer, but a is a firm-specific random variable.

[3] See the second equation on page 450 in HMR.

Suppose now that, as in Armenter and Koren (2012), the firms in country j are partitioned into S sectors according to some classification of economic activities, e.g., the Harmonized Commodity Description and Coding System. Then, the condition for sector $s \in \{1, \ldots, S\}$ of country j to export to i is that there is at least one firm in the sector for which $\pi_{ij}(a) > 0$. Therefore, the probability that sector s from country j exports to destination i is given by

$$\Pr\left(\ln a_s < \frac{1-\alpha}{\alpha}\left(\theta + \varphi_i + \psi_j - \ln f_{ij}\right) - \ln \tau_{ij}\right) = \int_{-\infty}^{x'_{ij}\beta} f_{\ln a_s}(z|x_{ij})\, dz = F_s\left(x'_{ij}\beta\right),$$

where a_s denotes the minimum value of a for firms in sector s, $f_{\ln a_s}(\cdot|\cdot)$ is the conditional density of $\ln a_s$ for sector s, $x'_{ij}\beta = (1-\alpha)\left(\theta + \varphi_i + \psi_j - \ln f_{ij}\right)/\alpha - \ln \tau_{ij}$, x_{ij} denotes a vector of regressors including importer and exporter dummies and variables measuring the trade frictions between i and j, β is a conformable vector of parameters, and we let $F_s(\cdot)$ vary with s because the distribution of $\ln a_s$ does not have to be the same for every sector.

Now let T_{ij}^s be an indicator variable that is 1 when at least one firm from sector s in country j exports to country i, being 0 otherwise, and notice that $\mathrm{E}\left(T_{ij}^s|x_{ij}\right) = \Pr\left(T_{ij}^s = 1|x_{ij}\right) = F_s\left(x'_{ij}\beta\right)$. Additionally, define $T_{ij} = \sum_{s=1}^{S} T_{ij}^s$ as the number of sectors exporting from j to i, which is the variable we want to model and is such that $0 \leq T_{ij} \leq S$. Hence, conditioning on x_{ij}, the expected value of the number of exporting sectors is

$$\mathrm{E}\left(T_{ij}|x_{ij}\right) = \sum_{s=1}^{S} F_s\left(x'_{ij}\beta\right). \tag{2}$$

Notice that for $S = 1$ this model is very similar to the first step of the model considered by HMR in which T_{ij} is just an indicator of whether country j exports to i (see equation 12 in HMR). However, we adopt a very different stochastic specification: in our model the unobservable a_s is the source of randomness and we treat the other variables as given; in contrast HMR treat a_s as given and the randomness of the exporting decision appears due to the unobservability of some elements of f_{ij} and τ_{ij},

which are viewed as random variables. In our model the possible presence of these unobserved costs only changes the form of $f_{\ln a_s}(\cdot|\cdot)$.

If sectoral information is available, it may be possible to estimate the functions $F_s\left(x'_{ij}\beta\right)$ and use them to study how the elements of x_{ij} affect $\mathrm{E}\left(T_{ij}|x_{ij}\right)$. However, without access to sectoral data (a constraint we will work with, following HMR), this approach is not available and the expected value of the number of exporting sectors has to be expressed as

$$\mathrm{E}\left(T_{ij}|x_{ij}\right) = SF\left(x'_{ij}\beta\right), \tag{3}$$

where $F\left(x'_{ij}\beta\right) = S^{-1}\sum_{s=1}^{S} F_s\left(x'_{ij}\beta\right)$ is the probability that a randomly drawn sector in country j will export to destination i.[4]

To proceed, it is necessary to specify a functional form for $F\left(x'_{ij}\beta\right)$. The choice of this functional form is an empirical issue that has to be addressed in each particular application. However, we can be guided in the choice of functional form by the fact that $F_s(\cdot)$ is the distribution of a minimum, which suggests that the complementary log-log model is a useful starting point.[5] Because restrictive distributional assumptions are unlikely to be valid in practice, we suggest specifying

$$F\left(x'_{ij}\beta\right) = 1 - \left(1 + \omega\exp\left(x'_{ij}\beta\right)\right)^{\frac{-1}{\omega}}, \tag{4}$$

where $\omega > 0$ is a shape parameter. This model is reasonably flexible and has the complementary log-log model as a limiting case when $\omega \to 0$.[6] Moreover, for $\omega = 1$,

[4]Indeed, $\mathrm{E}\left(T_{ij}|x_{ij}\right) = \sum_{s=1}^{S}\int_{-\infty}^{x'_{ij}\beta} f_{\ln a_s}(z|x_{ij})\,dz = S\int_{-\infty}^{x'_{ij}\beta}\sum_{s=1}^{S} S^{-1} f_{\ln a_s}(z|x_{ij})\,dz$. The result follows by letting $\int_{-\infty}^{x'_{ij}\beta}\sum_{s=1}^{S} S^{-1} f_{\ln a_s}(z|x_{ij})\,dz = F\left(x'_{ij}\beta\right)$, where $\sum_{s=1}^{S} S^{-1} f_{\ln a_s}(\cdot|\cdot)$ is the conditional density of $\ln a_s$ for a randomly picked sector.

[5]The complementary log-log model would be valid under the assumption that $\ln a_s$ follows the Gumbel (extreme value type I) distribution for a minimum.

[6]This choice of functional form corresponds to the assumption that the distribution of a_s for a randomly picked sector is a generalized Pareto with location parameter equal to 0 and scale parameter equal to 1. The form of (2) suggests that $F\left(x'_{ij}\beta\right)$ could also be specified as a mixture model. This approach, however, is computationally and statistically more demanding and therefore we do not pursue it here.

(4) reduces to the logit specification suggested by Papke and Wooldridge (1996) in a related context.[7]

Putting (3) and (4) together we get

$$\mathrm{E}\left(T_{ij}|x_{ij}\right) = S - S\left(1 + \omega \exp\left(x_{ij}'\beta\right)\right)^{\frac{-1}{\omega}}. \tag{5}$$

The model developed by HMR was used to motivate the specification of (5). Alternatively we could have used as starting points the models by Chaney (2008) or Manova (2012), which explicitly consider the existence of different sectors. However, because we consider only the case where no sectoral information is available, starting from the models by Chaney (2008) or Manova (2012) would have led exactly to the same result. Moreover, (5) can be motivated simply from the characteristics of T_{ij}, the random variable of interest. Indeed, T_{ij} is bounded by 0 and S and therefore its conditional expectation has the same bounds. So, it is sensible to specify the expectation of T_{ij} as the product of S, a known constant, by a function bounded by 0 and 1, such as one of the many specifications that have been used in binary choice models.[8]

3. Estimation

Because (5) specifies a conditional expectation and S is known, the model of interest can be written as

$$T_{ij}/S = 1 - \left(1 + \omega \exp\left(x_{ij}'\beta\right)\right)^{\frac{-1}{\omega}} + u_{ij}, \tag{6}$$

[7]Naturally, it is also possible to estimate $F(\cdot)$ nonparametrically, for example using the estimators proposed by Ichimura (1993). However, for typical international trade problems, the implementation of this kind of estimator is too cumbersome to be routinely used.

[8]Models for doubly-bounded count data have been used before (see, e.g., Johansson and Palme, 1996, and Santos Silva and Murteira, 2009). However, to the best of our knowledge, all the estimators used so far are likelihood based, whereas our proposed estimator focus on the conditional expectation and therefore does not require the specification of the likelihood function.**A related estimator, originally used for fractional data, was proposed by Papke and Wooldridge (1996) and will be explored below.**

where T_{ij}/S is bounded between 0 and 1, and u_{ij} is simply defined as $u_{ij} = T_{ij}/S - E\left(T_{ij}/S|x_{ij}\right)$, which implies that $E\left(u_{ij}|x_{ij}\right) = 0$. Estimation of β and ω is standard, but there are several possible consistent estimators of the parameters of interest.

A first approach is simply to use (non-linear) least squares, which is equivalent to using normal pseudo-maximum likelihood (see Gourieroux, Monfort, and Trognon, 1984). This estimator is consistent under very general conditions but it is unlikely to be attractive because it ignores the heteroskedasticity of the error term. Indeed, because T_{ij}/S is bounded between 0 and 1, u_{ij} is necessarily heteroskedastic and a substantial efficiency gain may be obtained by considering a "working" heteroskedasticity pattern, as in Papke and Wooldridge (1996). This heteroskedasticity pattern does not have to be correctly specified but can simply capture the fact that $\mathrm{Var}\left(T_{ij}/S|x_{ij}\right)$ must approach zero as $\mathrm{E}\left(T_{ij}/S|x_{ij}\right)$ approaches either 0 or 1.

Following Papke and Wooldridge (1996), we assume that $\mathrm{Var}\left(T_{ij}/S|x_{ij}\right)$ is proportional to $F\left(x'_{ij}\beta\right)\left(1 - F\left(x'_{ij}\beta\right)\right)$ and estimate the model by Bernoulli pseudo-maximum likelihood, which is a consistent estimator of the parameters of interest under very general conditions (see Gourieroux, Monfort, and Trognon, 1984) and likely to be much more efficient than least squares. That is, β and ω are estimated by maximizing an objective function with individual contributions of the form

$$L\left(\beta, \omega\right) = \left(T_{ij}/S\right) \ln F\left(x'_{ij}\beta\right) + \left(1 - T_{ij}/S\right) \ln\left(1 - F\left(x'_{ij}\beta\right)\right),$$

where $F\left(x'_{ij}\beta\right)$ is given by (4).

One final point is worth emphasizing: given the non-linearity of $F\left(x'_{ij}\beta\right)$ and the fact that we interpret it simply as an approximation to the probability that a randomly drawn sector in country j will export to destination i, the estimates of β are not particularly informative. Therefore, inference should focus on the partial effects of the regressors of interest and not on the parameter estimates per se.

4. Empirical illustration

We have argued for a different method to estimate doubly-bounded variates; whether the use of this approach makes a material difference at the estimation stage is an empirical question.[9] To investigate this matter we estimate a model for the number of sectors exporting from a given country to a destination. The sectors are defined using the 1996 revision of the Harmonized Commodity Description and Coding System at the 6-digit level, which has 5132 categories, and the data were obtained from UN Comtrade for 2001; Table A2 in the Appendix lists the 217 countries and territories for which we were able to obtain data for this study.

Data for the regressors were obtained essentially from the CIA's World Factbook and CEPII. In particular, the CEPII database was used to construct the following regressors: LOG DISTANCE, defined as the natural logarithm of distance between capitals (in kilometres); BORDER, a dummy that equals 1 when the two countries share a land border; COLONIAL TIE, a dummy that equals 1 either if the importer has ever colonized or been a colony of the exporter or if the two countries were once part of the same country; COMMON LANGUAGE, a dummy that equals 1 when the two countries share an official language; BOTH WTO, a dummy that equals 1 when the two countries are members of the WTO; RTA, a dummy that equals 1 if both countries are at least in one common regional trade agreement; COMMON CURRENCY, a dummy that equals 1 if either both countries use the same currency or if the exchange rates between their currencies is fixed. The CIA's World Factbook was used to construct two additional dummies: BOTH ISLANDS, which equals 1 if neither country has land borders; and BOTH LANDLOCKED, which equals 1 if both countries are landlocked. Finally, the variable RELIGION was constructed as in HMR; that is, the variable is the sum of the products of the shares of the population in each of the partners that are

[9] In a set of exploratory simulations we found that, as expected, the size and the direction of the biases of misspecified models depends both on the model and on the design of the experiment, especially on the distribution of the regressors.

Catholic, Muslim, or Protestant.[10] The information used to construct this variable is from multiple sources that include the CIA's World Factbook, Wikipedia, and the work of Kettani (2010a, 2010b, 2010c, 2010d, 2010e). Finally, the model includes importer and exporter dummies; the multilateral resistance terms suggested by Anderson and van Wincoop (2003).

These data are used to estimate six different models. The first model was used by Flam and Nordström (2006) and specifies $E\left(T_{ij}|x_{ij}\right) = x'_{ij}\beta$. The parameters are estimated by least squares and hence these results are labelled OLS. The second model is the one used by Eaton, Kortum, and Kramarz (2004) and by Hillberry and Hummels (2008), and specifies $E\left(\ln T_{ij}|x_{ij}\right) = x'_{ij}\beta$. Estimation is performed by OLS and these results are labelled LogLin. The third model specifies $E\left(T_{ij}|x_{ij}\right) = \exp\left(x'_{ij}\beta\right)$. Estimation is performed by Poisson (pseudo) maximum likelihood as in Dennis and Shepherd (2007), Berthou and Fontagné (2008), and Persson (2010); these results are labelled Poisson. The fourth model uses the same exponential specification for $\mathrm{E}\left(T_{ij}|x_{ij}\right)$ but in this case estimation is performed by negative binomial (pseudo) maximum likelihood as done by Persson (2012); these results are labelled NegBin. The fifth model specifies $\mathrm{E}\left(T_{ij}|x_{ij}\right)$ as in (5) but imposes $\omega = 1$. Estimation is performed by Bernoulli (pseudo) maximum likelihood as described in the previous section. Due to its similarity with the estimator proposed by Papke and Wooldridge (1996), the results for this model are labelled P&W. Finally, the sixth model specifies $\mathrm{E}\left(T_{ij}|x_{ij}\right)$ as in (5) and estimation is again performed by Bernoulli (pseudo) maximum likelihood. The estimates obtained with this more flexible approach are labelled Flex.

Table 1 presents the estimates obtained with the different models and the respective R^2, defined as the square of the correlation between T_{ij} and the corresponding estimate

[10] This variable has the obvious shortcoming of only accounting for three religions; for example, India and Nepal have a low value for RELIGION despite the fact that the majority of the population in both countries is Hindu. However, we include this variable for consistency with HMR. For more on the links between religion and economic activity, see Barro and McCleary (2003).

of $E(T_{ij}|x_{ij})$.[11] Table 2 presents the average across the entire sample of the partial effects of each of the regressors on $E(T_{ij}|x_{ij})$;[12] for the continuous variables (Log distance and Religion) these are just the derivatives of the estimate of $E(T_{ij}|x_{ij})$ with respect to regressors (notice that the derivative is with respect to Log distance, not distance itself), while for the dummy variables the partial effect is the difference between the estimate of $E(T_{ij}|x_{ij})$ with the dummy equal to 1 and with the dummy equal to 0. To provide a visual assessment of the goodness-of-fit of each model, Figure 1 displays the plots of nonparametric fits of $\mathrm{E}(T_{ij}|x_{ij})$ versus the fitted values of $\mathrm{E}(T_{ij}|x_{ij})$ for the each of the six parametric models considered. Each nonparametric fit was obtained by running a kernel regression of T_{ij} on the corresponding parametric fit of $\mathrm{E}(T_{ij}|x_{ij})$.[13] For a correctly specified model the plotted nonparametric fit should lie close to the identity line; the line where the abscissa is equal to the ordinate. For completeness, the identity line and the values of T_{ij} are also included in these plots.

In this example the OLS estimates generally have the expected sign but the magnitudes of some marginal effects appear to be clearly exaggerated. For example, the average increase in the number of sectors exporting from j to i resulting from being part of the same regional trade agreement is estimated to be almost 550, an increase that is more than 10 percent of the total number of sectors considered. The plot in the top-left corner of Figure 1 clearly illustrates the inappropriateness of the linear model in this case. Indeed, we see that the fitted values of $\mathrm{E}(T_{ij}|x_{ij})$ can be below zero and never get close to the upper bound of 5132. As a consequence, the nonparametric fit is far from being a straight line. This implies that the partial effects are mismeasured

[11] For comparability, in the LogLin model the R^2 is the square of the correlation (over the entire sample) between T_{ij} and the exponential of the fitted values of $\ln T_{ij}$.

[12] The results reported for the LogLin model are the partial effects on the exponential of the fitted values of $\ln T_{ij}$, averaged over all observations.

[13] Kernel regressions were performed in Stata 11 (StataCorp., 2009) using the Gaussian kernel and the default bandwidth. For the LogLin model the nonparametric fit is the kernel regression of T_{ij} on the exponential of the fitted values of $\ln T_{ij}$.

Table 1: Parameter estimates (and standard errors)

	OLS	LogLin	Poisson	NegBin	P&W	Flex
Log distance	−72.66	−0.91	−0.60	−1.20	−0.90	−1.07
	(4.72)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Border	444.89	0.49	−0.14	0.96	0.42	0.59
	(55.21)	(0.09)	(0.08)	(0.12)	(0.08)	(0.09)
Both islands	−0.23	0.31	0.41	0.44	0.45	0.53
	(8.61)	(0.06)	(0.07)	(0.07)	(0.07)	(0.08)
Both landlocked	−2.14	0.25	−0.06	0.30	0.04	0.16
	(12.15)	(0.06)	(0.11)	(0.08)	(0.10)	(0.09)
Colonial tie	291.39	0.70	0.49	1.03	0.76	0.97
	(59.15)	(0.08)	(0.07)	(0.09)	(0.07)	(0.08)
Common currency	107.21	−0.09	−0.25	0.74	0.09	0.25
	(54.13)	(0.09)	(0.08)	(0.12)	(0.07)	(0.09)
RTA	547.79	0.36	0.13	0.20	0.24	0.33
	(24.34)	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)
Common language	34.04	0.63	0.39	0.70	0.57	0.64
	(7.19)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
Both WTO	146.61	0.48	0.43	0.19	0.61	0.73
	(6.36)	(0.05)	(0.10)	(0.07)	(0.10)	(0.10)
Religion	0.23	0.40	0.37	0.53	0.35	0.41
	(9.26)	(0.04)	(0.05)	(0.07)	(0.05)	(0.06)
Overdispersion parameter	—	—	—	1.57	—	—
	—	—	—	(0.03)	—	—
ω	—	—	—	—	—	2.50
	—	—	—	—	—	(0.10)
R^2	0.56	0.18	0.76	0.07	0.92	0.92
Sample size	46872	24889	46872	46872	46872	46872

NOTE: All models include importer and exporter dummies.

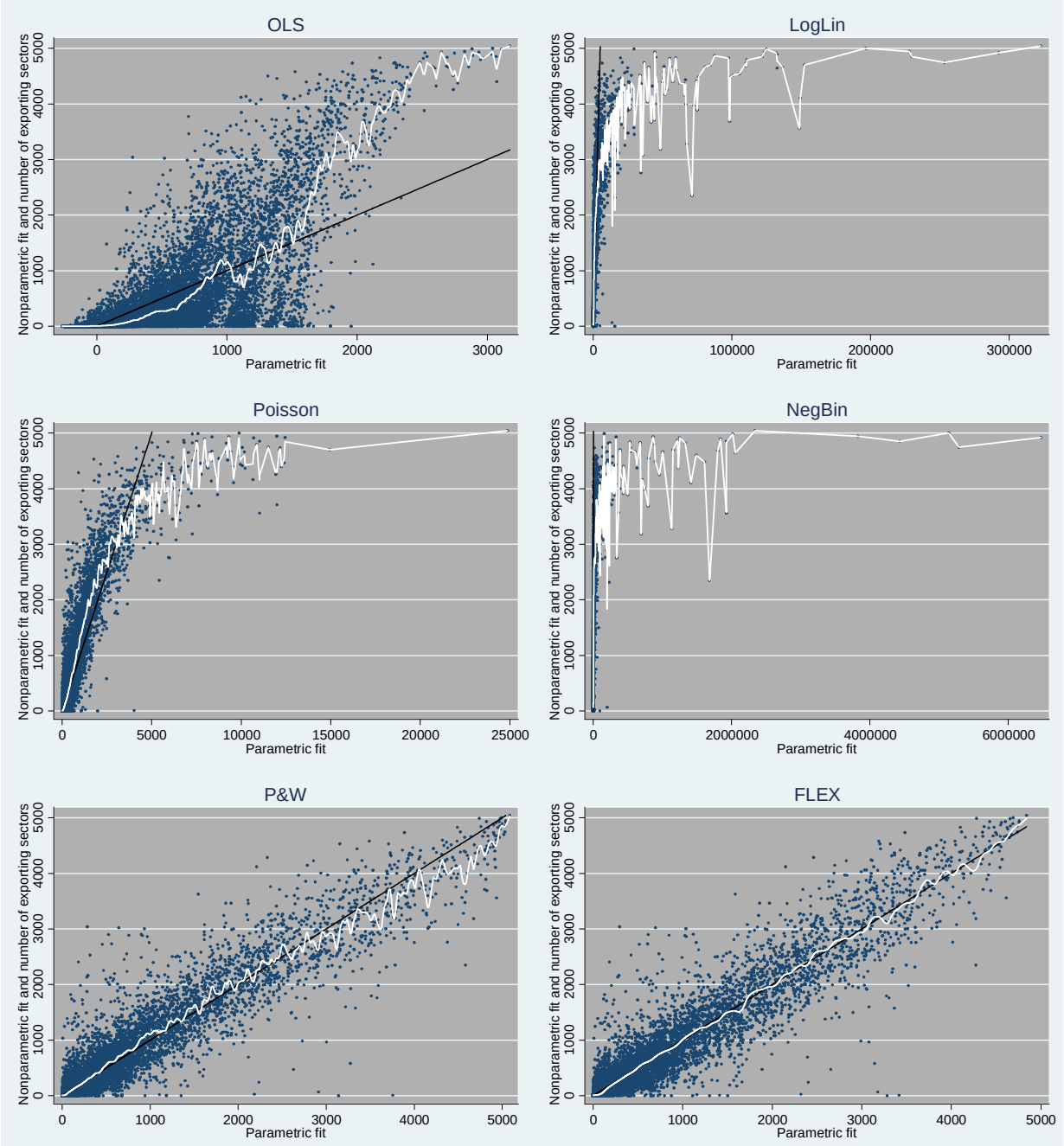

Figure 1: Nonparametric versus parametric fit of $\mathrm{E}(T_{ij}|x_{ij})$ for the six models considered (white line). For a correctly specified model the plotted nonparametric fit should lie close to the identity line. For completeness, the identity line (in black) and the values of T_{ij} (blue dots) are also plotted.

Table 2: Average Partial Effects (and p-values)

	OLS	LogLin	Poisson	NegBin	P&W	Flex
Log distance	−72.66	−263.53	−87.44	−2574.08	−86.86	−86.04
	(0.000)	(1.000)	(1.000)	(1.000)	(0.000)	(0.000)
Border	444.89	152.69	−19.72	1908.71	44.76	53.82
	(0.000)	(1.000)	(1.000)	(1.000)	(0.000)	(0.000)
Both islands	−0.23	106.76	72.68	1192.84	47.79	47.55
	(0.979)	(1.000)	(1.000)	(1.000)	(0.000)	(0.000)
Both landlocked	−2.14	82.87	−8.23	736.35	3.89	13.53
	(0.860)	(1.000)	(1.000)	(1.000)	(0.700)	(0.091)
Colonial tie	291.39	277.10	90.22	3558.86	86.35	95.64
	(0.000)	(1.000)	(1.000)	(1.000)	(0.000)	(0.000)
Common currency	107.21	−26.27	−32.85	1689.78	8.25	20.83
	(0.048)	(1.000)	(1.000)	(1.000)	(0.244)	(0.007)
RTA	547.79	98.72	19.19	402.52	23.66	28.00
	(0.000)	(1.000)	(1.000)	(1.000)	(0.000)	(0.000)
Common language	34.04	209.91	66.37	1617.70	59.92	56.30
	(0.000)	(1.000)	(1.000)	(1.000)	(0.000)	(0.000)
Both WTO	146.61	114.63	55.70	378.50	54.67	55.63
	(0.000)	(1.000)	(1.000)	(1.000)	(0.000)	(0.000)
Religion	0.23	114.79	54.03	1137.58	33.14	33.21
	(0.980)	(1.000)	(1.000)	(1.000)	(0.000)	(0.000)

for most observations and therefore it is not surprising that their average is sometimes quite unrealistic.

Results for the models that do not take into account the upper bound of the data are even less reliable. Indeed, none of the estimated average partial effects for LogLin, Poisson or NegBin is statistically significant and their values vary widely; the results of the NegBin model are particularly erratic. This behaviour is a consequence of the

fact that these models, by ignoring the upper bound, hugely overestimate the partial effects for the upper tail of the distribution, leading these observations to have a disproportionately large influence on the mean partial effect. This fact can be clearly seen in the corresponding plots in Figure 1, which show that the fitted values for LogLin, Poisson, and NegBin can be far above the upper bound of T_{ij}. This problem is particularly severe for the NegBin because, as it is well known, this estimator downweights the observations with large values of T_{ij} and therefore can fit them very poorly. The poor fit of the large observations combined with the exponential specification used for $\mathrm{E}(T_{ij}|x_{ij})$ implies that the partial effects can have extremely large values for many observations, rendering the estimated average partial effects totally unreliable.

The results in Table 2 show that the two models that take into account the upper bound, the one based on the Papke and Wooldridge's (1996) estimator and the more flexible version we propose, give reasonable results. Moreover, the two corresponding plots in Figure 1 clearly illustrate the advantage of using models that recognise the doubly-bounded nature of the data: both for P&W and for Flex the nonparametric fit is much closer to the identity line than for any of the other specifications previously considered.

These plots also show the advantage of the proposed model over P&W. Indeed, the nonparametric fit for Flex is generally much closer to the identity line, especially for the upper part of the distribution. The advantage of the flexible specification is confirmed by noticing that P&W is rejected against the proposed model (the additional parameter ω is significantly different from 1; see Table 1), and that this one is not rejected when tested against a more general specification.[14]

[14]The more general model we consider specifies

$$\mathrm{E}(T_{ij}|x_{ij}) = S\left(1 - \left(1 + \omega \exp\left(x_{ij}'\beta\right)\right)^{\frac{-1}{\omega}}\right)^{\delta},$$

which has as a special case the proposed model when $\delta = 1$. For these data, the estimate of δ is equal to 1.05 with an estimated standard error of 0.10. Therefore we cannot reject the null hypothesis $H_0 : \delta = 1$.

The differences between the results of P&W and Flex are not restricted to their goodness-of-fit. Indeed, although the average partial effects obtained with the two estimators are generally similar, for some of the regressors there are significant differences, as it is the case with COMMON CURRENCY. In particular, the P&W model leads to an estimated average partial effect of COMMON CURRENCY equal to 8 sectors, much smaller than the estimate of 21 sectors obtained with the proposed model. Moreover, the effect of this regressor is not statistically significant in the P&W model, but it is significant in the more flexible alternative.

In short, this example illustrates that the choice of specification used can make a material difference for the results one obtains. In particular it is vital to use models that specifically account for the doubly-bounded nature of the data. In the example presented here the proposed flexible specification clearly outperforms its competitors. This is an encouraging result in that it suggests that the model is flexible enough to describe adequately the type of data we are considering. Although the choice of the appropriate specification to use is an issue that needs to be carefully studied in each application, our results suggest that the proposed specification can be a good starting point.

Conclusions

Understanding and quantifying the factors affecting the number of sectors exporting in a given country is potentially relevant for the assessment of the effects of different trade policies. This paper studies models for the number of sectors exporting from a country to a given destination, when only aggregate country-pair level data are available. We argue that standard estimation methods previously used in the literature are not suitable due to the nature of the dependent variable, the number of sectors, which has both a lower and an upper bound (the latter being the number of classes in the classification system). The existence of these bounds implies that the partial

effects of the explanatory variables on the conditional mean of the dependent variable cannot be constant and must approach zero when the dependent variable approaches its bounds. Ignoring the nature of the data and simply using OLS or count-data models that ignore the upper bound is likely to lead to erroneous conclusions due to the severe misspecification of the models used.

We propose a flexible approach that takes into account the doubly-bounded nature of the dependent variable and, using country-pair data, we compare its performance to that of alterative specifications previously used in the literature. The proposed approach clearly outperforms the traditional estimators and, more importantly, leads to significant differences in the role played by different determinants of the extensive margin for trade. In particular, we argue that while other methods yield economically implausible quantitative effects for various trade determinants (e.g., sharing a border, a common currency or trade agreements), the new method yields economically reasonable effects. We, therefore, suggest that the proposed specification can be useful starting point for the construction of appropriate models identifying the role played by the different determinants of the number of sectors exporting from one country to another.

Appendix

Table A2: List of countries

Afghanistan	North Korea	Lesotho	St. Pierre & Miquelon
Albania	Congo Dem. Rep	Liberia	St. Vincent & the Grenadines
Algeria	Denmark	Libya	Samoa
Andorra	Djibouti	Lithuania	San Marino
Angola	Dominica	Luxembourg	Sao Tome & Principe
Anguilla	Dominican Rep.	Madagascar	Saudi Arabia
Antigua & Barbuda	Ecuador	Malawi	Senegal
Argentina	Egypt	Malaysia	Seychelles
Armenia	El Salvador	Maldives	Sierra Leone
Aruba	Equatorial Guinea	Mali	Singapore
Australia	Eritrea	Malta	Slovakia
Austria	Estonia	Marshall Isds	Slovenia
Azerbaijan	Ethiopia	Mauritania	Solomon Isds
Bahamas	FS Micronesia	Mauritius	Somalia
Bahrain	Faeroe Isds	Mexico	South Africa
Bangladesh	Falkland Isds	Mongolia	Spain
Barbados	Fiji	Montserrat	Sri Lanka
Belarus	Finland	Morocco	Sudan
Belgium	France	Mozambique	Suriname
Belize	French Polynesia	Myanmar	Swaziland
Benin	Gabon	N. Mariana Isds	Sweden
Bermuda	Gambia	Namibia	Switzerland
Bhutan	Georgia	Nauru	Syria
Bolivia	Germany	Nepal	TFYR of Macedonia
Bosnia Herzegovina	Ghana	Neth. Antilles	Tajikistan
Botswana	Gibraltar	Netherlands	Thailand
Br. Virgin Isds	Greece	New Caledonia	Timor-Leste
Brazil	Greenland	New Zealand	Togo
Brunei Darussalam	Grenada	Nicaragua	Tokelau
Bulgaria	Guatemala	Niger	Tonga
Burkina Faso	Guinea	Nigeria	Trinidad & Tobago
Burundi	Guinea-Bissau	Niue	Tunisia
Cambodia	Guyana	Norfolk Isds	Turkey
Cameroon	Haiti	Norway	Turkmenistan
Canada	Honduras	Occ. Palestinian Terr.	Turks & Caicos Isds
Cape Verde	Hungary	Oman	Tuvalu
Cayman Isds	Iceland	Pakistan	USA
Central African	India	Palau	Uganda
Chad	Indonesia	Panama	Ukraine
Chile	Iran	Papua New Guinea	United Arab Emirates
China	Iraq	Paraguay	United Kingdom
Hong Kong	Ireland	Peru	Tanzania
Macao	Israel	Philippines	Uruguay
Christmas Isds	Italy	Pitcairn	Uzbekistan
Cocos Isds	Jamaica	Poland	Vanuatu
Colombia	Japan	Portugal	Venezuela
Comoros	Jordan	Qatar	Viet Nam
Congo Rep.	Kazakhstan	South Korea	Wallis & Futuna Isds
Cook Isds	Kenya	Moldova	Western Sahara
Costa Rica	Kiribati	Romania	Yemen
Croatia	Kuwait	Russia	Zambia
Cuba	Kyrgyzstan	Rwanda	Zimbabwe
Cyprus	Laos	St. Helena	
Czech	Latvia	St. Kitts & Nevis	
Cote D'Ivoire	Lebanon	St. Lucia	

References

Acemoglu, D. and Zilibotti, F. (1997), "Was Prometheus Unbound by Chance? Risk, Diversification, and Growth." *Journal of Political Economy*, 105, 709-751.

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

Armenter, R. and Koren, M. (2012), "A Balls-and-Bins Model of Trade," CEPR Discussion Papers 7783.

Baldwin, R.E. and Di Nino, V. (2006), "Euros and Zeros: The Common Currency Effect on Trade in New Goods," NBER, Working Paper No. 12673.

Barro, R.J. and McCleary, R. M. (2003). "Religion and Economic Growth across Countries," *American Sociological Review*, 68, 760-781.

Berthou, A. and Fontagné, L. (2008), "The Euro Effects on the Firm and Product-Level Trade Margins: Evidence from France," CEPII Working Paper No. 2008-21.

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

Dennis, A. and Shepherd, B. (2007), "Trade Costs, Barriers to Entry, and Export Diversification in Developing Countries," The World Bandk Policy Research Working Paper No. 4368, Washington, D.C.

di Giovanni, J. and Levchenko, A.A. (2009), "Trade Openness and Volatility," *The Review of Economics and Statistics*, 91, 558-585.

Eaton, J, Kortum, S. and Kramarz, F. (2004). "Dissecting Trade: Firms, Industries, and Export Destinations," *American Economic Review*, 94, 150-154.

Flam H. and Nordström, H. (2006), "Euro Effects on the Intensive and Extensive Margins of Trade," *IIES Seminar Paper* No. 750, Institute for International Economic Studies, Stockholm.

Gourieroux, C., A. Monfort and A. Trognon (1984). "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681-700.

Greenwood, J. and Jovanovic, B. (1990), "Financial Development, Growth, and the Distribution of Income." *Journal of Political Economy*, 98. 1076.1107.

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

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

Hillberry, R. and McDaniel, C. (2002), "A Decomposition of North American Trade Growth since NAFTA," Working Papers 15866, United States International Trade Commission, Office of Economics.

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

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Singe-Index Models," *Journal of Econometrics*, 58, 71-120.

Johansson, P. and Palme, M. (1996), "Do Economics Incentives Affect Work Absence: Empirical Evidence Using Swedish Micro Data". *Journal of Public Economics*, 59, 195-218.

Kettani, H. (2010a), "Muslim Population in the Americas: 1950 – 2020," *International Journal of Environmental Science and Development*, 1, 127-135.

Kettani, H. (2010b), "Muslim Population in Africa: 1950 – 2020," *International Journal of Environmental Science and Development*, 1, 136-142.

Kettani, H. (2010c), "Muslim Population in Asia: 1950 – 2020," *International Journal of Environmental Science and Development*, 1, 143-153.

Kettani, H. (2010d), "Muslim Population in Europe: 1950 – 2020," *International Journal of Environmental Science and Development*, 1, 154-164.

Kettani, H. (2010e), "Muslim Population in Oceania: 1950 – 2020," *International Journal of Environmental Science and Development*, 1, 165-170.

Koren, M. and Tenreyro, S. (2007), "Volatility and Development," *The Quarterly Journal of Economics*, 122, 243-287.

Koren, M. and Tenreyro, S. (2012), "Technological Diversification," *American Economic Review*, forthcoming.

Manova, K. (2012), "Credit Constraints, Heterogeneous Firms, and International Trade," *The Review of Economic Studies*, forthcoming.

Melitz, M.L. (2003), "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71, 1695-1725.

Papke, L.E. and Wooldridge, J.M. (1996), "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates," *Journal of Applied Econometrics*, 11, 619-632.

Persson, M. (2012), "Trade Facilitation and the Extensive Margin," *The Journal of International Trade & Economic Development: An International and Comparative Review*, forthcoming.

Ramalho, E.A., Ramalho, J.J.S. and Murteira, J.M.R. (2011), "Alternative estimating and testing empirical strategies for fractional regression models," *Journal of Economic Surveys*, 25, 19-68.

Santos Silva, J.M.C. and Murteira, J.M.R. (2009), "Estimation of Default Probabilities Using Incomplete Contracts Data," *Journal of Empirical Finance*, 16, 457-465.

StataCorp. (2009). *Stata Release 11. Statistical Software.* College Station (TX): StataCorp LP.