Appendix A to "A Balls-and-Bins Model of Trade:" Nesting Balls-and-Bins in a Structural Model

Roc Armenter and Miklós Koren*

October 2013

For online publication

1 Data reference

Description of U.S. export data

Export data in the U.S. are based on Shipper's Export Declaration (SED) forms filed by exporters with the Customs and Border Protection and the Census Bureau. Filing a separate SED is mandatory for each shipment valued over $2,500. A *shipment* is defined as "all merchandise sent from one USPPI [firm] to one foreign consignee, to a single foreign country of ultimate destination, on a single carrier, on the same day."[1]

Each shipment is assigned a unique product code out of 8,988 potential "Schedule B" codes (of which 8,880 had positive exports in 2005). The Schedule B classification is based on the Harmonized System; the first six digits are HS codes. The remaining 4 digits are specific to U.S. exports. For convenience, we refer to these product codes in the paper as 10-digit HS codes.

We drop all 15 product codes in Chapter 98 (Special Classification Provisions). These categories are for products that are not identified by kind, either because of their low value, or some other reason.

There are 231 potential destination countries. Some of these entities are not countries but territories within countries (for example, Greenland has its own country code). We drop the country code 8220 (Unidentified Countries) and 8500 (International Organizations).

The Census Bureau publishes product–country aggregates based on this shipment-level dataset in "U.S. Exports of Merchandise." For each statistic, it also reports the number of SEDs (hence the number of shipments) that statistic is based on.

Armenter: Federal Reserve Bank of Philadelphia. E-mail: roc.armenter@phil.frb.org. *Koren*: Central European University, IECERS and CEPR. E-mail: korenm@ceu.hu

[1] "Correct Way to Complete the Shipper's Export Declaration," February 14, 2001 version.

We calculate the average shipment size for a product–country pair as the total value of exports divided by the total number of shipments in 2005. For each product, we then take the median shipment size across destination countries.

Baldwin and Harrigan (2007)

Baldwin and Harrigan (2007) use data on U.S. imports and exports with all trading partners in 2005 in their analysis. This data comes from the U.S. Census, which reports value, quantity, and shipping mode for imports and exports and shipping costs and tariff charges for imports by trading partner and 10-digit HS commodity code. The Census does not report import trade values less than $250 for imports and $2,500 for exports, so small trade values are treated as zeroes. For imports, their dataset contains 228 trading partners (countries for which at least one good had a nonzero import value) for goods in 16,843 different 10-digit HS categories. For exports, there are 230 trading partners for goods in 8,880 different 10-digit HS categories (see Table 2).

Baldwin and Harrigan also use data on trading partner distance from the United States from Jon Haveman's website:

`http://www.macalester.edu/research/economics/PAGE/HAVEMAN/Trade.Resources/Data/Gravity/dist.txt`.

Macro variables (GDP, GDP per worker) are from the Penn World Tables.

Helpman, Melitz, and Rubinstein (2007)

Helpman, Melitz and Rubinstein (2007) use annual trade data on bilateral trade flows for 158 countries (see Table A1 for a list) from Feenstra's "World Trade Flows, 1970-1992" and "World Trade Flows, 1980-1997".

They also use data on population and GDP per capita from the Penn World Tables and the World Bank's World Development Indicators. They use data from the CIA World Factbook on whether a country is landlocked or an island, along with each country's latitude, longitude, legal origin, colonial origin, GATT/WTO membership status, primary language and religion.

Data from Rose (2000) and Glick and Rose (2002) is used to identify whether a country pair belonged to a currency union or the same FTA, and data from Rose (2004) to identify whether a country is a member of the GATT/WTO.

The variable capturing regulation costs of firm entry is derived from data reported in Djankov et al. (2002).

Bernard, Jensen, and Schott (2007)

Bernard, Jensen, and Schott (2007) use a dataset that links individual trade transactions to information on the U.S.-based firms involved in the transactions. Data on trade transactions for exports in 1993 and 2000 is collected by the U.S. Census Bureau, and includes information on export value, quantity, destination, date of transaction, port, and mode of transport at the 10-digit HS code level. Shipments data are collected for all export shipments above $2,500.

Transaction-level data on imports are collected by U.S. Customs and Border Protection for all import shipments above $2,000. Detailed firm data comes from the Longitudinal Business Database of the Census Bureau. This dataset includes employment and survival information for all U.S. establishments, though the linked dataset does not include establishments in industries outside the scope of the Economic Census.

Hummels and Klenow (2005)

Hummels and Klenow (2005) use data from the United Nations Conference on Trade and Analysis (UNCTAD) Trade Analysis and Information System (TRAINS) CD-ROM for 1995. This dataset consists of bilateral import data for 5,017 goods, 76 importing countries and all 227 exporting countries. Goods are classified by 6-digit HS code. They also use matching employment and GDP data for a subset of 126 exporters and 59 importers from Alan Heston et al. (2002). More detailed U.S. trade data comes from the "U.S. Imports of Merchandise" CD-ROM for 1995 from the U.S. Bureau of the Census. This dataset reports value, quantity, freight paid, and duties paid for 13,386 10-digit commodity classifications and 222 countries of origin, 124 of which have matching data on employment and GDP.

Bernard and Jensen (1999)

This paper uses firm-level data from the Longitudinal Research Database of the Bureau of the Census from 1984-1992. Their dataset includes all plants that appear in the Census of Manufactures for 1987 and 1992. For comparisons which involve more than one year, the set of firms is further restricted to those which also appear in the the Annual Survey of Manufactures for the inter-census years. The result is an unbalanced panel of between 50,000 and 60,000 plants for each year.

Bernard, Eaton, Jensen and Kortum (2003)

Bernard, Eaton, Jensen and Kortum (2003) use data from the 1992 U.S. Census of Manufactures in the Longitudinal Research Database of the Bureau of the Census. This dataset covers over 200,000 plants, and records the value of their shipments, production and non-production employment, salaries and wages, value-added, capital stock, ownership structure, and value of exports.

Bernard, Jensen, Redding and Schott (2007)

Bernard, Jensen, Redding and Schott (2007) use transaction-level U.S. data from the 2002 U.S. Census of Manufactures. This paper also looks at more detailed data from the Linked-Longitudinal Firm Trade Transaction Database, which is based on data collected by the U.S. Census Bureau and the U.S. Customs Bureau. The dataset reports the product classification, value and quantity shipped, data of shipment, trading partner, mode of transport, and participating U.S. firm for all U.S. trade transactions between 1992 and 2000.

Eaton, Kortum, and Kramarz (2004)

Eaton, Kortum, and Kramarz (2004) use French firm-level data on type and destination of exported goods from 1986. This dataset is constructed by merging customs data with tax-administration data sets from Bénéfices Réel Normal (BRN)-Système Unifié de Statistiques d' Entreprises (SUSE) data sources, and contains information on over 200 export destinations and 16 SIC industries.

Eaton, Kortum, and Kramarz (2007)

Eaton, Kortum, and Kramarz (2007) use sales data of over 200,000 French manufacturing firms to 113 markets in 1986. As in Eaton, Kortum, and Kramarz (2004), this dataset is constructed by merging customs data with tax-administration data sets from Bénéfices Réel Normal (BRN)-Système Unifié de Statistiques d' Entreprises (SUSE) data sources.

2 Robustness analysis

We have explored a host of alternative calibrations. We detail here a selected few that shed light on the key determinants of our results. In the first set of calibrations we vary the total number of shipments (and thus observations) by assuming a counterfactual average shipment size. This illustrates how the model predictions depend on the sparsity of the data. Second we document the role of the skewness on trade flows, focusing on the calibration for firm-level facts.

2.1 Total number of shipments

We redo here our results for different numbers of shipments. We specify a "ball size" (in dollars) and convert trade flows into a discrete number of shipments by dividing the trade flow by the assumed "ball-size." We report results for ball sizes equal to \$2,500, \$18,000, \$36,000, \$100,000, and \$500,000. The smallest ball size is the lowest observed value of export transactions given the reporting rules of the Census Bureau, while \$36,000 is the actual average shipment size, and thus the value that ensures that the total number of shipments in the exercise is the same as in the data. Table 1 reports the implied number of shipments, with the corresponding number in the data highlighted in boldface.

Ball size	Number of shipments (10^6)
\$2,500	311.1
\$18,000	43.2
\$36,000	**21.6**
\$100,000	7.7
\$500,000	1.5

Table 1: Ball-size calibrations and number of shipments (in millions)

First we experiment with ball sizes equal and larger than the average size of an export. From an economic point of view, it may well be the case that the relevant decisions involve multiple transactions simultaneously and a calibration with a larger ball size would be appropriate. Table 2 shows our quantitative results for ball sizes between \$36,000 and \$500,000. We also included the corresponding data value for each of the stylized facts.

		Ball size		
Moment	Data	\$36k	\$100k	\$500k
HS10-level product×country U.S. export flows				
Share of zeros	82%	72%	80%	90%
OLS coefficient of nonzero flow on GDP	0.08	0.10	0.09	0.06
Firm×country U.S. export flows				
Share of zeros	98%	96%	98%	99%
Gravity for firms, GDP OLS coefficient	0.71	0.56	0.61	0.68
Single-product exporters				
Fraction over total exporters	42%	43%	57%	76%
Share of total exports	0.4%	0.3%	1.1%	7.4%
Single-destination exporters				
Fraction over total exporters	64%	44%	58%	77%
Share of total exports	3.3%	0.3%	1.1%	7.5%
Single-destination, single-product exporters				
Fraction over total exporters	40%	43%	57%	76%
Share of total exports	0.2%	0.3%	1.1%	7.4%
Exporters in U.S. manufacturing				
Fraction over total firms	18%	74%	61%	41%
Size premium of exporters	4.4	34	25	16

Table 2: Ball-size calibrations: \$36,000; \$100,000; and \$500,000

The changes in the magnitudes are intuitive. First, as we calibrate to a larger ball size, there are fewer balls overall and the incidence of empty product bins increases. This applies equally for zeros in trade or the fraction of single-product, single-destination exporters. Single-product and single-country exporters also increase their export share. With fewer shipments overall, most firms will end up with just one ball and would necessarily be single-product, single-country exporters. A larger ball-size calibration also reduces the fraction of exporting firms, closer to the one we see in the data. This is because if firms are taken to have fewer balls, it is less likely that any one of them comes from exports. However, even the \$500,000 ball-size calibration would predict significantly more exporters (41%) than in the data (18%). This suggests that economies of scale in deciding whether or not to export are rather strong.

Incidentally the column under the actual average shipment size, \$36,000, provides an additional check as it shuts down all the systematic variation in shipment size in trade flows. The resulting predictions from the model are virtually undistinguishable from those using the number of shipments per trade flow.

Table 3 shows our quantitative results for smaller ball sizes, between $36,000 and $2,500. These calibrations illustrate neatly the slow rate of convergence to a dense data set: the number of shipments under the smaller ball-size calibrations is orders of magnitude larger than the documented evidence yet sparsity still gives rise to zeros. The last column describes the limit as ball size shrinks to zero and the data is perfectly dense.

		Ball size			
Moment	Data	$36k	$18k	$2,500	none
HS10-level product×country U.S. export flows					
Share of zeros	82%	72%	66%	45%	0
OLS coefficient of nonzero flow on GDP	0.08	0.10	0.10	0.09	0
Firm×country U.S. export flows					
Share of zeros	98%	96%	94%	86%	0
Gravity for firms, GDP OLS coefficient	0.71	0.56	0.53	0.42	0
Single-product exporters					
Fraction over total exporters	42%	43%	35%	15%	0
Share of total exports	0.4%	0.3%	0.1%	0.0%	0
Single-destination exporters					
Fraction over total exporters	64%	44%	35%	15%	0
Share of total exports	3.3%	0.3%	0.1%	0.0%	0
Single-destination, single-product exporters					
Fraction over total exporters	40%	43%	35%	14%	0
Share of total exports	0.2%	0.3%	0.1%	0.0%	0
Exporters in U.S. manufacturing					
Fraction over total firms	18%	74%	81%	95%	100%
Size premium of exporters	4.4	34	67	337	n.a.

Table 3: Ball-size calibrations: $36,000; $18,000; and $2,500

As expected, smaller ball sizes imply fewer empty bins, both for products and for firms. Note, however, that even with a $2,500 ball size the majority of product and firm bins remain empty. This means that even a very small degree of indivisibility leads to a large number of empty bins. (In the limit, of course, there will be no empty bins.) Smaller balls also imply less action on the "extensive margin." Because most bins are filled, it is unlikely for new balls to fall in empty bins – hence the coefficient of country size on number of product or firm bins is smaller.

2.2 Skewness in export sales

The skewness in trade flows and categories plays an important role in our results. The gravity equation naturally generates skewness across destinations as some of them are large or small, close or far. Similarly heterogeneity in products (turnips or airplanes) generates a large variation across product categories.

We focus our robustness analysis on the skewness present in the distribution of export sales across exporters. The bottom line is that the balls-and-bins model matches the firm-level export patterns whenever the calibration accounts for the left-tail in the export distribution—in short, the small exporters.

In the first set we retain the use of the lognormal distribution and vary the parameter σ. By adjusting σ downwards we reduce the skewness of the distribution. The location parameter μ remains constant so we match the median exporter sales. Doing so preserves the left tail properties of the distribution; the skewness is reduced by thinning the right tail of the distribution. We maintain the calibration of the bin size distribution as detailed in the paper.

Table 4 reports the results. The first row is the calibration used in the paper. We focus on the fraction of exporters that are single-product and single-country exporters. These are 42 % and 60 % in the data. We explore parameters down to $\sigma = 2.3$, which is well below any estimate for the distribution of *domestic* sales.

		Fraction of Exporters	
σ	μ	Single-product	Single-country
3	11	43.3 %	44.1 %
2.7	11	42.6 %	43.5 %
2.5	11	42.0 %	43.0 %
2.3	11	41.3 %	42.4 %

Table 4: Lognormal distribution: Alternative calibrations.

As Table 4 makes clear, the model predictions are very robust. The predicted fractions decrease, but only very slowly. The fraction of single-product, single-country exporters (not reported) remains very close to 40 % for all calibrations.

The reason why the model is so robust is that the right tail of the distribution is irrelevant. Virtually all firms selling more than $100,000 are predicted to be multi-country, multi-product exporters. It does not really matter how exporters above this threshold are distributed.

It is instead the small exporters—the left tail—that drives our results. We can make this point explicitly by choosing a calibration with few small exporters. We reduce σ to 2.7 and *increase* the location parameter to 12.5 so we cut by half the number of exporters below $100,000. In this case, the fraction of single-product exporters falls to 31.7 %, and the fraction of single-country exporters falls to 32.6 %. The predicted fractions now fall sharply if we keep lowering σ and increasing μ, collapsing the distribution from both sides. With $\sigma = 2.3$ and $\mu = 13.5$ the fraction of single-product exporters is just 16 %.

We also experimented with alternative distributions as the Pareto distribution and the Yule-Simon distribution (drawing the number of shipments directly). The results reinforced the conclusion that matching the left-tail is a necessary and sufficient condition for the balls-and-bins model to match the facts.

3 Aggregation and size premiums

3.1 Aggregate statistics

There is a total of T trade flows (countries, firms) in the dataset, each indexed by t and comprised of n_t shipments. The distribution of shipments across trade flows, $n_1, n_2, ..., n_T$, is taken as given. We find it useful to describe the distribution of shipments across trade flows as a probability distribution over $\mathbb{N}$, denoted π_n.[2] Each shipment can be classified into one of K categories.

The expected number of non-empty bins across all trade flows is given by

$$E(k|n_1, n_2, ..., n_T) = \sum_{n=1}^{N} \pi_n \sum_{i=1}^{K} [1 - (1 - s_i)^n] = \sum_{i=1}^{K} \sum_{n=1}^{N} \pi_n [1 - (1 - s_i)^n]. \tag{1}$$

Let $G(z)$ denote the *probability generating function* (PGF) corresponding to the distribution $\{\pi_n\}$:

$$G(z) = \sum_{n=1}^{N} \pi_n z^n.$$

Then the number of non-empty bins can be written as

$$E(k|n_1, n_2, ..., n_T) = \sum_{i=1}^{K} [1 - G(1 - s_i)].$$

Since $G(z)$ is strictly convex, uneven bin-size distributions will have a smaller expected number of non-empty bins.

What about the proportion of single-bin trade flows? For each trade flow of size n, the probability is $\sum_{i=1}^{K} s_i^n$. The conditional probability is then

$$\Pr(k = 1|n_1, n_2, ..., n_T) = \sum_{n=1}^{N} \pi_n \sum_{i=1}^{K} s_i^n = \sum_{i=1}^{K} \sum_{n=1}^{N} \pi_n s_i^n.$$

We can also express it in terms of the PGF as

$$\Pr(k = 1|n_1, n_2, ..., n_T) = \sum_{i=1}^{K} G(s_i).$$

It then becomes clear that the convexity of $G(z)$ also preserves the properties of each flow with respect to the fraction of single bins. In particular, we can now assert that more even bin-size distributions induce a lower fraction of single-bin flows.

Finally we can also calculate the fraction of *balls* that have fallen into a single bin. This corresponds to, for example, the fraction of *sales* attributed to single-product firms.

$$\sum_{n=1}^{N} \pi_n n \sum_{i=1}^{K} s_i^n = \sum_{i=1}^{K} \sum_{n=1}^{N} \pi_n n s_i^n.$$

[2] To be precise, we assume that the support is bounded by some finite N.

With the use of the PGF notation,

$$\sum_{n=1}^{N} \pi_n n s_i^n = G'(s_i)s_i.$$

And we can easily have the average size of trade flows that all fall in bin i is

$$\frac{\sum_{n=1}^{N} \pi_n n s_i^n}{\sum_{n=1}^{N} \pi_n s_i^n} = \frac{G'(s_i)s_i}{G(s_i)}.$$

It is important to note that, unless the number of trade flows is infinite, the actual fractions will be a random variable. Since all distributions are known it is actually possible to derive the actual distribution for each moment. It is, however, often unpractical to do so and one can use Monte Carlo methods to derive the distribution as needed.

3.2 Exporter's size premium

Let π_n be the unconditional size distribution of firms. The firm-size distribution conditional on not exporting is

$$\Pr(n|\text{no export}) = \frac{\Pr(\text{no export}|n)\pi_n}{\Pr(\text{no export})}.$$

The average sales (number of balls) of non-exporters is

$$E(n|\text{no export}) = \sum_{n=1}^{\infty} \frac{\pi_n n(1-s)^n}{\Pr(\text{no export})}.$$

The average sales for the population of firms is

$$E(n) = \sum_{n=1}^{\infty} \pi_n n.$$

We can express the expected sales of non-exporters in terms of the probability generation function $G(z)$ of the firm size distribution.

$$E(sales|\text{no export}) = \frac{(1-s)G'(1-s)}{G(1-s)},$$

the elasticity of G evaluated at $1-s$. Note that G is differentiable. The unconditional mean is given by the same formula but evaluated at $z = 1$:

$$E(sales) = \frac{1G'(1)}{G(1)}.$$

A sufficient condition for non-exporters being smaller than the average if the elasticity of G is increasing in z.

To see how the skewness in the firm size distribution leads to a large exporter premia, we parameterize the distribution as a *zeta distribution*. This is the discrete analogue to Pareto distribution, and its probability mass function is

$$\pi_n = \frac{n^{-\alpha}}{\zeta(\alpha)}.$$

Here α is the tail exponent, and is estimated to be about 2.06 by Axtell (2001). The probability generating function of the zeta distribution is

$$G(z) = \frac{\mathrm{Li}_\alpha(z)}{\zeta(\alpha)},$$

where Li_α is the (non-analytic) polylogarithm function. By properties of polylogarithm, the elasticity of $G(z)$ is given by

$$\frac{zG'(z)}{G(z)} = \frac{\mathrm{Li}_{\alpha-1}(z)}{\mathrm{Li}_\alpha(z)}.$$

With $\alpha = 2.06$, this implies that exporters are about 18 times as big as non-exporters. If we lower α closer to 2, we are putting more mass of the distribution on its upper tail. For $\alpha = 2.02$, exporters are 27 times as big as non-exporters.