Foreign Affiliate Sales Data - Work in progress

Eddy Bekkers
World Trade Organization
eddy.bekkers@wto.org

Erwin Corong
ecorong@purdue.edu
GTAP Center

Csilla Lakatos
World Bank
clakatos1@worldbank.org

Indre Macskasi
indre.macskasi@gmail.com

Jeanne Metivier
World Trade Organization
jeanne.metivier@wto.org

Steen Wettstein*
World Trade Organization
steen.wettstein@wto.org
†

Wen Jin "Jean" Yuan
USITC
WenJin.Yuan@usitc.gov

2021-04-15

Abstract

This paper describes the approach to produce a complete and reconciled global dataset of foreign affiliate sales which can be merged with the existing GTAP database and can thus be used in computable general equilibrium (CGE) models extended with affiliate sales, building on the approach in Fukui and Lakatos (2012).

To construct the affiliate sales dataset we proceed in five steps. First, foreign affiliate sales data are collected from Eurostat, OECD, and national sources. Second, inward and outward data are reconciled. Third, three-dimensional affiliate sales data (by host country, home country and sector) are used to estimate a gravity equation (with Poisson Pseudo Maximum Likelihood, PPML) to be able to generate values for home-host-sector pair combinations for which reported data are missing. The final specification is selected based on out-of-sample performance of various specifications. Fourth, the estimated affiliate sales data are reconciled with reported data at higher levels of aggregation, using quadratic optimization techniques. Preference is generally given to the activities reported by and within a country's territory (inward flow) when available, because of the overall better quality of those figures. Fifth, and finally, the affiliate sales data are integrated in the GTAP database, adjusting the affiliate sales data for possible inconsistencies with gross output in the GTAP database.

Two affiliate sales datasets are produced: (1) a 2014 dataset, compatible with the GTAP 10 Data Base, and (2) a 2017 dataset, compatible with the GTAP 11 Data Base.

Keywords: affiliate sales; FATS; global affiliate sales data; database

1 Introduction

There is a lack of reported foreign affiliate sales data, which complicates incorporation of multinational activity and affiliate sales in computable general equilibrium models. Various scholars have in the past compensated for a lack of foreign affiliate sales data by estimating the missing information. Fukui and Lakatos (2012) produced global foreign affiliate sales dataset spanning the years 2003-2007 using EUROSTAT FATS data. The final dataset generated by these scholars includes 128 host and home countries and 31 sectors. In more recent work, Wettstein et al. (n.d.) construct a dataset on trade in

*This contribution reflects a description of a methodology and does not represent the position or opinions the WTO or its respective Members, nor the official position of any staff members.

†Corresponding author; Email: steen.wettstein@wto.org

Table 1: Number of reported observations by source

source	Inward	Outward
Eurostat	1406303	526434
National source	4664	8984
OECD	87242	35656

services by mode of supply. Mode 3 services trade measures trade by commercial presence which corresponds with the local sales of foreign affiliates. The final dataset spans 2005 to 2017 and consists of about 200 economies with flows from/to partner world and 13 service sectors. Our aim is to build on this previous work and create a comprehensive dataset of bilateral affiliate sales for all regions and all sectors (manufacturing and services) in the GTAP Data Base for the years 2008-2018.

The objective of this paper is to describe into detail the methodology used to creating the bilateral global affiliates sales dataset. To create complete '3-dimensional,' home-host country by sector, foreign affiliate sales dataset, we collect all the available FATS data that we are aware of. Moreover, we make use of data on the number of foreign establishments and FDI stocks.

The final dataset ranges from 2008 to 2018 and consists of ?? host countries, ?? home countries and 35 sectors. Values will be proportionally allocated to cover all sectors in GTAP.

The rest of the paper is organized as follows. The next section presents the available FATS data and data needed to estimate missing foreign affiliate sales data. Section 3 presents the steps undertaken to fill in the missing zeros and estimate the missing data. The quadratic optimisation procedure used to to reconcile estimated data at the three-dimensional level with reported data at the one- and two-dimensional level is described in section 4. Finally, section 5 concludes.

2 Data

This section describes the available FATS data and other data such as FDI and value added data used to construct a 3-dimensional global foreign affiliate sales database.

2.1 Foreign affiliate sales data

We collected the foreign affiliate data on sales and number of enterprises from all sources known to us. This includes data from Eurostat, OECD and national sources as statistic offices of United States, Canada, Israel, Malaysia, Serbia, Thailand, Viet Nam, Zambia and North Macedonia. Table 1 summarizes the number of observations available from different sources.

3 Data imputation and derivations

3.1 Strategies to fill in missing zeros

The specificity of the 3-dimensional FATS data is its large number of zeros. The structure of the available data allows to identify and fill missing zeros.

Firstly, when reported data with the partner world is zero, new country-to-country entries are created and filled with zeros. Then, whenever reported aggregated sectors (e.g. 'B-N_S95_X_K' and 'B-S_X_O') are zero, new entries with every sector are created and filled with zeros.

Secondly, FDI stocks data are used to fill some additional gaps. Sources are summarized in (table) and come from the:

(i) UNCTAD World Investment Report the IMF Coordinated
(ii) Direct Investment Survey (CDIS) database
(iii) Eurostat database
(iv) OECD database
(v) Various national sources

If there is no Foreign direct investment (reported zeros), the assumption is that there is no (or negligible) commercial presence of foreign entities. Therefore, for theses cases, if FDI data exist and entries do not exist in the FATS, entries are created and filled with zeros.

Thirdly, if the reported number of foreign enterprises in FATS data is zero, new entries are created and filled with zeros.

3.2 Mirroring

Only a few countries collect FATS data. A way to increase the number of countries in the dataset is to take advantage of the information available in the opposite flow. If no values are reported within a time series between two countries and the information in the opposite flow is available, this info is used as estimates. In this exercise, missing inward FATS are estimated using outward FATS reported by the home country. Before this step inward FATS data contain 36 reporting countries and goes to - 202 countries after exploting mirroring values.

Table 2 summarizes the results of the steps described above and reports the number of observations by each methodology flag. At this point the dataset contains 16.93 % reported values, 9.72 % imputed zeros, and 72.01 % values derived from the mirroring flow. Table 3 summarizes the data between North and South. One can observe that in value terms most of foreign affiliate sales take place in developed countries with the headquarters in another developed country, while the biggest number of observations are

Table 2: FATS data: number of data points per methodology flag

Methodology	Description	Inward	Inward_share	Outward	Outward_share
agr	NA	18493	1.33	19102	1.41
E0	Simple derivation	9747	0.70	631634	46.78
E0.1	World derivation	46890	3.37	252188	18.68
E0.2	FDI	74510	5.36	42259	3.13
E0.3	Number of enterprises	4060	0.29	1967	0.15
E6	Correction negative	1	0.00	NA	NA
M_E0	Mirroring	620717	44.63	5175	0.38
M_E0.1	Mirroring	227295	16.34	46357	3.43
M_EUR	Mirroring	149646	10.76	143960	10.66
M_NAT	Mirroring	2556	0.18	998	0.07
M_OECD	Mirroring	1364	0.10	2390	0.18
R_EUR	Reported	230311	16.56	198915	14.73
R_NAT	Reported	1758	0.13	3893	0.29
R_OECD	Reported	3349	0.24	1337	0.10

Table 3: North-South Statistics

flow	DEV	n	Value	n_share	Value_share
Inward	North-North	345529	102279572	0.25	0.82
Inward	North-South	124665	3251847	0.09	0.03
Inward	South-North	906828	19361171	0.65	0.15
Inward	South-South	13081	220018	0.01	0.00
Outward	North-North	334628	79329299	0.25	0.79
Outward	North-South	920258	18589271	0.68	0.19
Outward	South-North	80422	2185097	0.06	0.02
Outward	South-South	12033	220018	0.01	0.00

of sales in developing countries with the headquarters in a developed country. Our initial dataset has relatively small number of observations between 'South-South.' This is due to the lack of FATS data collection practices in developing countries.

When the data is reported by both home and host country, in some cases very large discrepancies are observed. In trade literature to reconcile the import and export data Gehlhar (1996) developed a methodology where reporters are ranked based on how accurate their reporting is. Shaar (2019) pointed out that this accuracy measure is biased and depending on whether import or export flow is larger it produces different results even when the absolute difference between the flows is the same. He proposed a new measure to calculate the similarity between the flows:

$$S_{ij} = 1 - |M_{ij} - X_{ji}|/(M_{ij} + X_{ji}) \tag{1}$$

where M_{ij} is import value from country j reported by country i, and X_{ji} is export value of reporting country j to country i. To get the reliability of the reporter the similarity measure is weighted by the import share in country i's total imports and summed across partners. Thus, the average similarity for country i's imports can be written as follows:

$$\bar{S}_i^M = \sum_{j=1}^{N} M_{ij}/M_i * S_{ij}^M \tag{2}$$

We apply this methodology on FATS data by exchanging the inward data for import value and outward for export. Generally, inward data is thought to be better quality, thus we give preference to inward data and select outward data only when inward data is not available or the reporter of inward flow gets very low reliability ranking.

3.3 Backasting, forecasting and interpolation

This step exploits the information contained within the time series to complete gaps in the bilateral data set. If information is available for more than one year, and contained at least one missing value, either in the middle, at the beginning or at the end of the time series, the missing data point is interpolated[1], backcasted or forcasted using three-year moving averages, respectively.

3.4 Gravity estimation

In this section, results are presented for different specifications of gravity regressions. These regressions are run using pooled sample and estimated sector by sector. The predictive power of different specifications is tested using cross-validation simulations.

3.4.1 Data to estimate missing foreign affiliate sales

This sub-section explanatory variables are described. Variables are chosen according to both their availability (number of countries) and their predictive power.

For most countries, GDP per capita is pulled out from the UNdata website. Countries for which GDP per capita is missing are taken from national source. This is the case for the Faeroe islands and Chinese Taipei. Tanzania is sourced from the World Bank data.

Developed/developing country status is based on the WTO self-declaring list.

As commonly used in the gravity literature, the bilateral data, which includes distance, common language, contiguity and colonial link, is from CEPII (Mayer and Zignago (2011)).[2] The sector specific value added data is from UN national accounts, see appendix for more details.

[1]More specifically, the Fritsch & Carlson algorithm is used for monotonic cubic interpolation. It is a variant of cubic spline that preserves monotonicity of the interpolation function

[2]We approximate data for Curacao and Sint Maarten with data for Netherlands Antilles and data for Serbia, Montenegro and Kosovo with data for Yugoslavia.

No explicit policy-related variables are used to predict sales due to their poor coverage and, more importantly, to limit endogeneity problems in future analytical applications. Nevertheless, their impact and predictive power is tested.

We use a FTA dummy from DESTA database which stores information on various types of trade agreements such as customs unions, free trade agreements or partial free trade agreements. We also include Bilateral Investment Treaty (BIT) from An Electronic Database of Investment Treaties (EDIT) and a dummy variable for intra-EU flow. Finally, we source the data on Government Effectiveness and Regulatory Quality from The Worldwide Governance Indicators (WGI). We follow Marel and Shepherd (2020) and take a simple average of two indicators to get an aggregate measure for the capacity of the government to effectively formulate and implement sound policies. Individual governance indicators are reported for over 200 countries and territories over the period 1996-2019. However, the data is missing for 11 countries that are in our initial FATS dataset.[3]

3.4.2 Pooled regressions

To check the explanatory power of collected variables we run

The first step is to run pooled regressions using all available observations. To account for large number of zeros in our data, the models are fitted using the Poisson Pseudo Maximum Likelihood (PPML) estimator as proposed by Silva and Tenreyro (2006).[4] We start with the benchmark model specified as:

$$\begin{aligned} FAS_{ijst} = exp(&\alpha_1 + \beta_1 ln(dist_{ij}) + \beta_2 colony_{ij} + \beta_3 contig_{ij} + \beta_4 comlang_{ij} + \\ &\beta_5 ln(VA_{ist}) + \beta_6 ln(VA_{jst}) + \beta_7 ln(GDPpc_{it}) + \beta_8 ln(GDPpc_{jt})) * \epsilon_{ijst} \end{aligned} \tag{3}$$

where FAS_{ijst} is foreign affiliate sales in sector s of country i with headquarters of the affiliates in country j in year t. $dist_{ij}$ stands for physical distance between the countries i and j. $colony_{ij}$ is a dummy equal to 1 if countries i and j have ever had a colonial link. $contig_{ij}$ is a dummy indicating whether the two countries are contiguous. $comlang_{ij}$ dummy is set to if the same language is spoken by at least 9% of the population in both countries. VA_{ist} and VA_{jst} stands for sector specific value added in sector s and year t of host country i and home country j, respectively. $GDPpc_{it}$ and $GDPpc_{jt}$ are GDP per capita variables for both host country i and home country j. ϵ_{ijst} is the error term.

In the second specification we add additional explanatory variables:

[3]Missing countries: Netherlands Antilles, Cura?ao, Faeroe Islands, Kosovo, Montserrat, New Caledonia, PSE,French Polynesia, Saint Martin, Turks and Caicos Islands, Serbia and Montenegro.

[4]Fukui and Lakatos (2012) compare the Poisson Pseudo Maximum Likelihood (PPML), the Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB) and find that PPML outperform the other two estimators while used on affiliate sales data.

$$FAS_{ijst} = exp(Benchmark + \beta_9 DevStatus_i + \beta_{10} DevStatus_j + \beta_{11} FTA_{ijt} + \beta_{12} BIT_{ijt} + \beta_{13} WGI_{it} + \beta_{14} EU_{ijt}) * \epsilon_{ijst} \quad (4)$$

where $DevStatus_i$ and $DevStatus_j$ are dummy variables indicating whether host country i and home country j, respectively, are developing countries. FTA_{ijt} is a dummy equal to 1 if countries i and j share a foreign trade agreement (or $DESTAdepth$ variable which ranges from 1 to 7 and describes how comprehensive the foreign trade agreement is between the two countries). BIT_{ijt} is a bilateral investment treaty dummy. WGI_{it} is a variable ranging from -2.5 to 2.5 and measuring the government effectiveness of the host country i. EU_{ijt} is a dichotomous dummy variable taking the value 1 if both countries are EU member states.

Table 4: Pooled regresssions

	value											
	0: Benchmark .	0: Reported .	0: excl.deriv .	0: excl.mir. .	1: Benchmark .	1: Reported .	1: excl.deriv .	1: excl.mir. .	2: Benchmark .	2: Reported .	2: excl.deriv. .	2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	
CONSTANT	−33.800***	−25.552***	−32.513***	−29.791***	−25.297***	−18.132***	−24.283***	−22.427***	−31.103***	−21.479***	−29.943***	
	(1.735)	(1.983)	(1.807)	(2.091)	(1.461)	(1.797)	(1.498)	(1.905)	(2.226)	(3.191)	(2.258)	
LOG(DIST)	−0.454***	−0.493***	−0.437***	−0.506***	−0.475***	−0.513***	−0.456***	−0.523***	−0.294***	−0.413***	−0.295***	
	(0.068)	(0.065)	(0.069)	(0.064)	(0.068)	(0.064)	(0.069)	(0.063)	(0.095)	(0.105)	(0.095)	
CONTIG	0.143	0.192	0.156	0.211	0.056	0.148	0.071	0.163	0.220	0.241*	0.218	
	(0.157)	(0.152)	(0.153)	(0.153)	(0.168)	(0.161)	(0.166)	(0.162)	(0.142)	(0.150)	(0.140)	
COMLANG_ETHNO	0.129	0.007	0.128	−0.001	0.108	−0.048	0.107**	−0.055	0.275**	0.193	0.258**	
	(0.195)	(0.232)	(0.188)	(0.231)	(0.190)	(0.229)	(0.185)	(0.227)	(0.121)	(0.190)	(0.120)	
COLONY	0.226	0.181	0.224	0.210	0.313	0.278	0.306**	0.305	0.287**	0.119	0.282**	
	(0.212)	(0.237)	(0.208)	(0.239)	(0.197)	(0.229)	(0.193)	(0.229)	(0.128)	(0.178)	(0.129)	
LOG(GDP_REP)	0.800***	0.853***	0.769	0.868								
	(0.050)	(0.057)	(0.051)	(0.057)								
LOG(GDP_PAR)	0.976***	0.930***	0.957	0.932								
	(0.029)	(0.039)	(0.030)	(0.038)								
LOG(VAI_REP)					0.813***	0.846***	0.783***	0.860***	0.867***	0.878***	0.838***	
					(0.047)	(0.053)	(0.048)	(0.052)	(0.039)	(0.053)	(0.040)	
LOG(VAI_PAR)					0.966***	0.924***	0.949***	0.924***	0.999***	0.926***	0.983***	
					(0.031)	(0.043)	(0.032)	(0.042)	(0.033)	(0.045)	(0.034)	
LOG(GDPPC_REP)	0.344***	−0.040	0.336***	−0.065	0.302***	0.051	0.299**	0.028	0.259**	−0.088	0.252**	
	(0.088)	(0.113)	(0.089)	(0.106)	(0.084)	(0.107)	(0.085)	(0.102)	(0.113)	(0.150)	(0.115)	
LOG(GDPPC_PAR)	1.105***	1.072***	1.059***	1.124***	1.062***	1.005***	1.017***	1.063***	1.336***	1.304***	1.295***	
	(0.089)	(0.096)	(0.091)	(0.092)	(0.089)	(0.099)	(0.091)	(0.096)	(0.125)	(0.155)	(0.127)	
DEVING_WTO_REP									0.550**	−0.340**	0.548**	
									(0.282)	(0.412)	(0.279)	
DEVING_WTO_PAR									0.270	0.356	0.335	
									(0.345)	(0.430)	(0.345)	
FTA_DESTA									−0.141	−0.510**	−0.150	
									(0.215)	(0.255)	(0.216)	
BIT									0.330**	0.180	0.286**	
									(0.133)	(0.183)	(0.135)	
WB_WGI									0.423***	0.285**	0.424***	
									(0.138)	(0.127)	(0.137)	
EU									0.900***	0.832***	0.858***	
									(0.132)	(0.191)	(0.128)	
Country FE	NO	NO	NO	NO	NO	NO	NO	NO				
Industry FE	YES	YES	YES	YES	YES	YES	YES	YES				
Year FE	YES	YES	YES	YES	YES	YES	YES	YES				
R2	0.49	0.528	0.496	0.52	0.539	0.569	0.541	0.562	0.589	0.61	0.583	
R2 adj	0.49	0.528	0.496	0.52	0.539	0.569	0.541	0.562	0.589	0.61	0.583	
# Zeros	1306710	184769	337171	326791	1306710	184769	337171	326791	1306710	184769	337171	
# of Mirror	993024	0	150267	0	993024	0	150267	0	993024	0	150267	
Observations	1,370,288	235,185	403,884	379,912	1,370,233	235,185	403,884	379,911	1,314,374	235,185	403,883	

¡em¿ Notes : ¡em¿ ¡em¿ Standard errors are robust for ppml ¡br¿ p-values in parenthesis ¡br¿ * p¡.1, ** p¡.05, As shown by table 4, running regressions with and without imputations, derivations and the use mirroring values does not impact the coefficients, nor the R2. As expected, Distance has always a significant and negative impact on the prediction of sales. Contiguity, Common language, and Colony are positive but not significant. Value added by sector always has a positive and significant impact like GDP. GDP per capita of the home country is always positive and highly significant. No country fixed effects are introduced in this example because predictions are needed for countries outside the sample.

3.4.3 Sector by sector regressions

To allow the heterogeneity between sectors we estimate gravity equation for each sector separately.

To be completed.

3.4.4 Cross-validation

In this subsection we describe the cross-validation exercise which we carry out to measure the out-of-sample accuracy of the models used to predict the affiliate sales. In particular, the predictive power of the different specifications are compared in terms of error. The sample used for the cross-validation contains information available for all specifications, i.e. the sample of model M1.1. The algorithm took the following steps: 1) The sample was randomly split into two datasets: the training set (70%) and the test set (30%). Each reporter-partner pair was represented in both sets. 2) Each model was fit on the training set and then the coefficients were used to predict the test set. 3) For each specification M1.1, M1.2, M1.3, M1.4, and M1.5, the prediction accuracy was computed using the mean absolute error (MAE) criterion. 4) Steps 1. to 3. were repeated 200 times.

To be completed.

4 Reconciling the dimensions

4.1 Quadratic Optimization

The final consistency of the FATS sales will be ensured by a quadratic optimization technique proposed by Silva and Tenreyro (2006). This allows to incorporate and reconcile information from different sources.

To be completed.

5 Concluding remarks

To be completed.

6 Appendix: Completing value added data

The value added data by sector is sourced from UN national accounts. It is mostly reported at 1 digit sector level. The data is not available for all country-year-sector combinations and, thus this appendix describes the procedure used to fill in the gaps. This is done in two steps:

1) When country-sector combination is available at least for one year, we use the share of value added in GDP of that sector in that country to approximate the data for the missing years.

2) When there are no observations for a specific country-sector combination, an average sector share in GDP across regions is used, except for middle East and Commonwealth of Independent States where global shares are used (due to a small number of available countries). Moreover, to account for structural change we multiply the regional shares of each service sector in GDP by total service share in GDP of individual country sourced from World bank or fitted value of total service share in GDP regressed on GDP per capita (see equation 5) for countries which are not present in the World Bank data.

$$ServiceShare_{rt} = \alpha + \beta * GDPcap_{rt} + \epsilon_{rt} \tag{5}$$

To account for regional differences we run the regressions by region. The estimated value added for an individual service sector is:

$$VA(ser)_{rit} = \widehat{ServiceShare}_{rt} * sectorShare_{region,it} * GDP_{rt} \tag{6}$$

Similarly, the value added in agriculture, mining and querying, manufacturing, electricity, water supply, and construction is estimated as:

$$VA(ABCDE)_{rit} = (1 - \widehat{ServiceShare}_{rt}) * sectorShare_{region,it} * GDP_{rt} \tag{7}$$

To fill in the World Bank data that is missing for some years for country-sector combination linear extrapolation is used:

$$log(ServiceShare(WB)_{rt}) = \alpha + \beta(year_t) + \gamma_r + \varepsilon_{rt} \tag{8}$$

where γ_r stands for country fixed effects.

Table 5 presents the number of observations by the way the data was estimated or reported. After filling in the missing data, value added is greater than GDP for Singapore, Thailand (2018 only), and Samoa. The estimated values for these countries are re-scaled so that sectoral value added adds up to GDP.

Table 5: Flags in value added data

flag	n	Share
E1 - Est. using sector shares from other years	1987	5.01
E20 - Est. using UN data on services share	99	0.25
E21 - Est. using WB data on sercices share	14432	36.39
E22 - Est. using linear extrap. of missing years in WB data	462	1.17
E23 - Est. using fitted values of service share	2342	5.91
R - Reported	20332	51.27

References

Fukui, Tani, and Csilla Lakatos. 2012. "A Global Database of Foreign Affiliate Sales." GTAP Research Memoranda 4009. Center for Global Trade Analysis, Department of Agricultural Economics, Purdue University. `https://ideas.repec.org/p/gta/resmem/4009.html`.

Gehlhar, Mark J. 1996. "Reconciling Bilateral Trade Data For Use In Gtap." Technical Papers 28714. Purdue University, Center for Global Trade Analysis, Global Trade Analysis Project. `https://doi.org/10.22004/ag.econ.28714`.

Marel, Erik van der, and Ben Shepherd. 2020. "Trade Facilitation in Services : Concepts and Empirical Importance." No. 9234. *Policy Research Working Paper*. World Bank, Washington, DC.`%20https://openknowledge.worldbank.org/handle/10986/337431`.

Mayer, Thierry, and Soledad Zignago. 2011. "Notes on CEPII's Distances Measures: The GeoDist Database." Working Papers 2011-25. CEPII. `http://www.cepii.fr/CEPII/en/publications/wp/abstract.asp?NoDoc=3877`.

Shaar, Karam. 2019. "Reconciling International Trade Data." EconStor Preprints 206629. ZBW - Leibniz Information Centre for Economics. `https://ideas.repec.org/p/zbw/esprep/206629.html`.

Silva, J. M. C. Santos, and Silvana Tenreyro. 2006. "The Log of Gravity." *The Review of Economics and Statistics* 88 (4): 641–58. `https://ideas.repec.org/a/tpr/restat/v88y2006i4p641-658.html`.

Wettstein, Steen, Antonella Liberatore, Joscelyn Magdeleine, and Andreas Maurer. n.d. "A Global Trade in Services Data Set by Sector and by Mode of Supply (TiSMoS)."