A Global Foreign Direct Investment Stocks Database

Eddy Bekkers*
Csilla Lakatos†
Roger So‡
Wen Jin "Jean" Yuan§
Fabio Traverso¶

PRELIMINARY DRAFT PLEASE DO NOT QUOTE

Abstract

We construct a three-dimensional (bilateral, by sector) global FDI stock database for the baseline year of 2014 and 2017 which is compatible with the GTAP version 10 and version 11 database. To do so, we perform a gravity estimation on bilateral and by sector FDI stock data collected from the Eurostat and the ITC Investment Map which is subsequently used to extrapolate for a complete dataset by filling in all missing data. Next, the extrapolated dataset is reconciled under a quadratic optimisation procedure with higher dimension FDI data collected from the OECD, UNCTAD and IMF. Finally, we adopt a novel computational method based on the probabilistic theory of absorbing Markov chains to further adjust the reconciled database, its purpose is to correct for geographical distortions in the FDI statistics caused by the presence of tax havens. This paper serves as an update to the previous global FDI database constructed by Gouel et al. (2012) which is based on the GTAP version 7 database with the baseline year of 2004.

Keywords: Computable general equilibrium models, Foreign Direct investment, Databases, Conduit FDI

JEL codes: C68, C82, F21

Print date: 15th April 2021

*Corresponding Address: Eddy Bekkers, World Trade Organization, Rue de Lausanne 154, 1202 Geneva, Switzerland. email: eddy.bekkers@wto.org. *Disclaimer:* The opinions expressed in this article should be attributed to its authors. They are not meant to represent the positions or opinions of the WTO and its Members and are without prejudice to Members' rights and obligations under the WTO.

†clakatos1@worldbank.org – World Bank
‡yu.so@wto.org – World Trade Organization
§Wen.Yuan@usitc.gov – United States International Trade Commission
¶fabio.e.traverso@outlook.it

1 Introduction

From productivity spillover to technological transfer, the vital role of foreign direct investment (FDI) in the global contemporary economy have long been the focus of numerous quantitative economic analysis. Its importance has also led to the development of computable general equilibrium (CGE) models that explicitly incorporate FDI (Petri, 1997; Dee and Hanslow, 2000), thus providing scholars and policymakers an useful tool in stimulating changes in FDI and its wider impacts to the rest of the economy. Nevertheless, early research under the new framework often suffered from poor inputs due to the limited availability of global FDI data. To address this issue, Gouel et al. (2012) collected and compiled global FDI data sourced from various multilateral databases such as those maintained by the International Monetary Fund (IMF) and Eurostat. Subsequently, any remaining missing values are extrapolated under a gravity estimation and the completed dataset has further undergone a quadratic optimisation procedure to produce the first fully balanced, three-dimensional (multi-sector, bilateral multi-region) FDI database. Finally, adjustments made by Lakatos and Walmsley (2011) brought the database to be compatible with the Global Trade Analysis Project (GTAP) v7 regional and sectoral aggregation of 113 regions and 57 sectors under the reference year of 2004.

The database has since become a valuable input for research under the GTAP-FDI framework (Lakatos and Fukui, 2014), and it was adapted into later GTAP database versions with more recent reference years by assuming a proportionate growth of FDI to the host country's GDP. However, as time passes this ad-hoc solution is becoming increasingly inadequate: take the GTAP database as an example, its geographical coverage in version 10 has been expanded to 141 regions. Meanwhile, now with 65 sectors, the GTAP sectoral classification has gone through an overhaul in order to be compatible with the International Standard Industrial Classification (ISIC).

Turning to the original data sources used in Gouel et al. (2012), there the coverage and the data quality of FDI databases since 2004 has been greatly improved. In particular, the IMF's Coordinated Direct Investment Survey (CDIS) began reporting bilateral FDI stock data on top of one dimensional (total FDI with World) data in 2009 and it is now the only comprehensive, global database available at the bilateral level. In terms of data quality, the latest 4th edition Benchmark Definition of Foreign Direction Investment (BMD4) was widely implemented in the OECD and Eurostat database since 2013, and this new reporting standard introduced several key improvements to more accurately reflect the the direction and magnitude of FDI movement between countries.

Aside from changes in the GTAP database and improvements in FDI data, the rise of developing countries as both the destinations and sources of cross-border investment has upended the previous trend in which global FDI was dominated by North-North investment until the early 2000s. According to the World Investment Report (2019), less than half (46%) of the total FDI went to developed economies in 2018, and at a record low, only 55% of total World FDI originated from developed economies in the same year. In addition, as the global economy becomes more and more digitized, the report points to a sectoral shift of investment from heavy assets towards tech industries with intangible assets. Together, these new developments underline the need for a revision to the original GTAP FDI database.

Against the backdrop, this project follows closely the procedures adopted in Gouel et al. (2012) and updates the FDI database to be fully compatible with the GTAP v10 and v11 database with the reference year of 2014 and 2017 respectively. Nevertheless, there are two major changes introduced in this project: first, our work only concerns the update of FDI **stock** (position) database and we omit FDI **flow**, this is because most International Economics and International Business literature tends to adopt stock values as proxy of foreign affiliate activities due to the volatile nature of FDI flow. In addition, the frequent occurrence of negative FDI flow as a results of reverse investment or disinvestment pose a further challenge to our gravity estimation which relies on the Poisson pseudo maximum likelihood (PPML) estimator, negative values are also incompatible with the GTAP database which is designed to only accept strictly positive inputs. Second, we additionally adopt a novel computation method developed by Casella (2019) that adjust FDI data based on an estimated matrix of ultimate investor country share. The motivation behind the adjustment is to addressed some of the main criticisms and shortcomings of FDI statistic that remain unresolved despite the passage of time. More specifically, almost all FDI data currently available are reported based on immediate investor country rather than ultimate investor country, these data do not account for the complex financing structure of multinationals and are therefore heavily distorted by the presence of tax havens/offshore financial centres with favourable tax and regulatory regimes, the proposed remedy above seeks to identify these 'phantom FDI' and reallocate them according to their ultimate investor country. Finally, other minor adjustments are also made to the original procedures so that they are better suited for the current data at hand, these will be discussed in greater details at various points of the paper.

The rest of the paper is organised as followed: Section 2 introduces the original data sources and details the adjustments made in the cleaning and merging of these FDI data; the gravity model is presented in Section 3 and the extrapolation results is subsequently discussed in Section 4; Section 5 describes the optimisation procedures and the phantom FDI adjustment is carried out in Section 6; finally, we concludes in Section 7.

2 Data Collection and Cleaning

2.1 Original Data Sources

a) United Nations Conference on Trade and Development (UNCTAD) FDI Statistic

The UNCTAD FDI Statistic is released annually in the Annex section of its flagship 'World Investment Report'. With 203 reporting countries, the database has the widest country coverage, but at the same time it is a one-dimensional database with no bilateral or sectoral breakdown, meaning that there are only two observations per reporting country per year - total inward FDI stock and total outward FDI stock[1].

b) IMF Coordinated Direct Investment Survey

The IMF database is a two-dimensional bilateral database to which each reporting country submits their annual inward and outward FDI position with respect to various partner countries. Technically speaking, each reporting country can submit bilateral stock information with more than 200 partner countries/territories, however in reality the data availability varies greatly by country and by year. In terms of country coverage, the database began with 92 reporting countries in 2009 and slowly expanded to around 110 participants in recent years, the coverage is therefore not universal but it is still the most comprehensive database available at the bilateral level.

c) Organisation for Economic Co-operation and Development (OECD) FDI Statistics

At first glance, the OECD FDI database is also two-dimensional but with a much limited country coverage of 36 which are mostly comprised of OECD Member states, and as a matter of fact, all 36 reporting countries are already covered in the IMF database. However, the OECD database further contains FDI stock information at the two-dimensional sectoral level. In other words, aside from a bilateral dataset without sectoral breakdown similar to the IMF database, additionally there is a sectoral-inward and a sectoral-outward dataset. Note that these sectoral datasets refers to total inward and outward FDI stock with World which is then breakdown by sector, crucially they do not contain bilateral information.

Nevertheless, the importance of the OECD FDI database mainly lies in the new reporting guideline introduced in BMD4 (OECD, 2014). More specifically, BMD4 recommends breaking down total FDI into FDI made by special purpose entities (SPEs) and by normal

[1]Note that in a separate project, the UNCTAD's 'Bilateral FDI Statistic 2014' contains bilateral stock information for 206 reporting countries, however we do not consider these data as they are only available up to 2011 and there is no update since

operating units (Non-SPEs). While SPEs do not have a strict definition, generally they refer to enterprises with few employees or little activities and their main purpose is to serve as financial conduits for transmitting 'pass through capital' and 'transit-FDI'. Such distinction is useful as the existence of SPEs heavily 'distorts the country patterns of FDI statistics and cause double-counting in the statistics'. This separate distinction of FDI made by SPEs and non-SPEs is relevant for the construction of FDI dataset used in this project and the identification of 'phantom FDI' which we shall return to in Section 3(c) and Section 6.

In addition, the new guideline makes further recommendations to better capture the influence exerted from investment made between fellow enterprises[2] as well as to measure FDI stocks by their market values which is more intuitive and consistent over time. Unlike the distinction of SPEs and Non-SPEs above, these improvements are already incorporated into the raw FDI statistic and do not affect our data cleaning and adjustment procedures, as a result we do not discuss them in details.

d) Eurostat FDI Statistic

As a regional database, the Eurostat also has a limited country coverage: it is consist of only 39 reporting countries which are mostly European nations. However, just as in the original data construction more than a decade ago, Eurostat remains to be the only three-dimensional database with FDI statistic available both by partner country and by sector, therefore it continues to be the backbone of the gravity estimation. Similar to the OECD FDI database, the reported FDI follows the BMD4 guideline and they are broken down into SPEs and non-SPEs.

e) International Trade Centre (ITC) Investment Map

The Investment Map is a project undertaken by the ITC to collect and merge FDI data from the four aforementioned databases, in this sense, it is itself a secondary database with the exact same inputs as ours[3]. Instead, the novelty of the Investment Map lies with its effort in collecting three-dimensional and two-dimensional data from national sources (e.g. central banks) of countries that are not covered by the OECD or Eurostat, therefore the Investment Map should be considered as a supplement to these two databases. In the end, national sources data collected by the ITC contribute an additional 26 and 11 reporting countries at the two-dimensional sectoral level and three-dimensional level respectively.

[2]Enterprises that have no direct investment relationship themselves but have a direct investor in common

[3]The output of the Investment Map project is not directly adopted as our FDI datasets due to many fundamental differences in how data is collected and treated

Table 1: Dimensional Coverage of FDI Databases

Database	One-Dimensional	Two-Dimensional		Three-Dimensional
		Bilateral	Sectoral	
UNCTAD	✓			
IMF	✓	✓		
OECD	✓	✓	✓	
Eurostat	✓	✓	✓	✓
ITC*	✓	✓	✓	✓

*The dimensional coverage of ITC database varies depending on the reporting country

2.2 Discrepancy Analysis

a) Discrepancy between databases

Table 1 summarises the dimensional coverage of the five FDI databases discussed above, note that in every database with two-dimensional coverage or higher, FDI data is also available at the lower-dimension. It must be emphasised here that these lower-dimension data are separate values submitted by the reporting countries rather than a simple aggregation of the higher-dimensional data, this is reasonable since many higher-dimensional entries were withheld due to confidentiality reasons. While greater data availability is certainly welcomed, nonetheless these overlapping dimension give rise to the concern over discrepancies between the databases.

After crosschecking the data between our original sources, we find high level of consistency among FDI statistic from the IMF, OECD and Eurostat database, this is expected since all three databases follows similar reporting standard. More specifically, according to the databases' respective metadata, the OECD and Eurostat essentially adopt a common framework for reporting FDI statistic, meanwhile the OECD's BMD4 is developed jointly with the IMF's 6th edition Balance of Payments and International Investment Position Manual (BPM6) and are therefore compatible. Indeed, in the OECD's metadata all but one of the reporting countries specifically indicate that their data is fully consistent with those submitted to the IMF Survey. In spite of this, minor discrepancies are still present for a variety of reasons such as later data revision as a result of differences in submission timeline.

Regarding UNCTAD's FDI data, various data adjustments adopted in the database meant that we observe large discrepancy for individual countries. For instance, the UNCTAD database excludes SPEs and adopts non-SPE values for specific countries, whereas the IMF strives to report total FDI for every reporting country. As an illustration, comparing to the IMF database, this adjustment in the UNCTAD database leads to the total inward FDI stock received by the Netherlands and Luxembourg in 2017 to be 67% and 95.6% smaller respectively.

Finally, since FDI data collected from national sources by the ITC are strictly compliment to the OECD and Eurostat database, there is no discrepancy by construction. However, for some of these reporting countries at the bilateral level, their FDI data differs from the ones submitted to the IMF Survey, in most cases, this is a result of differences in national agencies from which the data is collected.

b) Discrepancy between inward and outward stock

The discrepancies between inward and outward reporting is a persistent issue in FDI data. In the original exercise, Gouel et al. (2012) had already provided an summary of the root causes: to paraphrase, these discrepancies are a result of failures in accounting for reinvested earnings, short-term financing, investment made by SPEs, cross-border real estate transactions, and investment made by affiliates.

Although the BMP4 brought forward a number of improvements including the identification of SPEs and the tracing of investment made between foreign affiliates, it is not clear whether they are adopted universally outside of the OECD/Eurostat framework. Furthermore, the quality of data still varies depending on the capacity of national institutions that are in charge of collecting and processing data in the first place. Consequently, global inward FDI still differs from global outward FDI. Similarly at the bilateral and three-dimensional level, the inward FDI stock as reported by the host country rarely matches with the outward FDI stock data reported by the source country, and very often the differences are significant.

In summation, the discrepancies across dimensions and between inward and outward stock underline the need for reconciliation through a quadratic optimisation procedure which we will return to in Section 5.

2.3 Constructing FDI Datasets

We now follow Gouel et al. (2012) and construct various one- and two-dimensional datasets as optimisation constraints as well as a three-dimensional dataset for the subsequent gravity estimation and extrapolation. In the original exercise, the authors had directly adopted the UNCTAD/IMF database as their one-dimensional constraint[4], similarly the OECD database was also directly adopted for the two-dimensional constraints. However, we have decided to deviate from their approach since the discrepancy analysis and the overview in Table 1 above highlight the fact that when compiling FDI dataset at the one and two-dimensional level, there are now multiple databases available and therefore the selection of which data sources to give preference to would unavoidably involve making value judgements based on observed discrepancies.

Aside from merging databases, we have also introduced other changes in the construction

[4]The IMF had only started collecting bilateral FDI stock data from 2009 onwards

procedures in order to achieve maximum data coverage while ensuring overall consistency, these will be discussed in greater details in the sub-sections below. Finally, note that for the one- and two-dimensional datasets, we are only collecting data for 2014 and 2017 which correspond to the reference year of GTAP v10 and v11. Whereas at the three-dimensional level, we strive to cover multiple years and construct a panel dataset for the gravity estimation.

a) One-Dimensional datasets

Beginning at the one-dimensional level, as mentioned above we are compiling two datasets as constraints for the quadratic optimisation procedure: total inward FDI stock and total outward FDI stock. Since IMF Survey is consistent and fully covers all the reporting countries in the OECD and Eurostat database, it boils down to choosing between the IMF and UNCTAD database. In comparison, the UNCTAD database applies different reporting principles to different countries: beside from the adoption of non-SPE values discussed above, in some countries the values are estimated or derived by adding the reported flow value to previous year's stock. Moreover, some entries are themselves secondary and are originated from the IMF. In the end, favouring a more consistent approach, we give preference to inputs from the IMF Survey which include around 120 and 90 countries reporting inward and outward FDI respectively, and they are supplemented with a further 60 and 70 countries from the UNCTAD database.

In order to ensure consistency, we refrain from mixing sources between the two reference years[5]. However, we treat inward and outward as two separate datasets and therefore do not impose the restriction that for a given country, its inward and outward data must come from the same source. Finally, we carry out an additional check for data coming from the UNCTAD database and discard any FDI values that are estimated or derived.

b) Two-Dimensional datasets

Moving up to the two-dimensional level, here we are separately compiling a bilateral dataset, a sectoral-inward dataset, and a sectoral-outward dataset. Similarly, they are also used as constraints for the quadratic optimisation.

At the bilateral level, once again the IMF Survey fully covers the OECD, Eurostat as well as national sources data collected by the ITC, meaning that no additional data is available beyond the original 121 reporting countries. To expand our coverage, we adopt mirroring values in the IMF dataset which essentially treat outward FDI stock reported by the source country as inward FDI stock of the host country. Not surprisingly, the procedure would generate

[5] As an example, Argentina reports inward FDI stock to the IMF Survey in 2017 only, we therefore treat 2014 as missing even though UNCTAD data is available for both 2017 and 2014

many overlapping observations whereby both mirrored values and inward FDI are available, and as analysed previously the discrepancies between them are large. For reconciliation, we adopt a rather simple approach by giving preference to inward FDI and use mirror values to fill in the gaps since we believe host countries have greater incentive to accurately measure inward FDI for tax and regulatory purposes. Furthermore, mirrored values are only adopted when there is no inward FDI data for both 2014 and 2017 within a given country-pair, this additional condition is meant to prevent volatility caused by mixing inward and mirrored values within time series. Ultimately, the adoption of mirrored values increased the number of observations by more than 40% and expanded the country coverage to 245.

Regarding the sectoral-inward dataset, the OECD database provides information on 33 reporting countries and it is supplemented with an additional 6 and 24 reporting countries from the Eurostat and ITC database. On the other hand, the sectoral-outward dataset is made up of 33, 6, and 11 reporting countries from the OECD, Eurostat and ITC database respectively. Since many European nations are also OECD Members, there are many overlapping observations between the two database but their values are virtually identical. As a result, given that national sources data from the ITC Investment Map are strictly supplementary, the merging of databases will not lead to conflicting data or discrepancy. Note that mirroring values are not available to dataset with no bilateral dimension therefore in the end, the country coverage for the the sectoral-inward and -outward dataset are limited to just 63 and 50.

c) Three-Dimensional dataset

Our three-dimensional dataset is constructed with 39 reporting countries from the Eurostat and an additional 11 reporting countries from the ITC Investment Map. The Eurostat database contains over 200,000 observations for both inward and outward FDI stock whereas data from national sources collected by the ITC are of poor coverage, contributing only around 9,000 inward and 3,000 outward FDI observations. However, extra information from the ITC is still useful in mitigating the economic and geographical bias of the Eurostat database in which most reporting countries are developed European nations.

Before further adjustments, note that Eurostat is the only database that reports FDI values in Euro, for our purposes they are converted to US Dollar using the end-of-year exchange rate published by the IMF.

Next, the 400,000 observations in the Eurostat database refers to FDI made by all resident units or 'total FDI', among them the database separately distinguishes investment made by SPEs for around 170,000, or 40% of the total observations. Following OECD (2014)'s recommendation, we subtract SPEs and adopt non-SPE values where possible to limit distortions and to better capture FDI movements across countries.

All outward FDI observations are subsequently transformed to mirrored values and we follow the exact same principles adopted in the construction of two-dimensional bilateral database where we give preference to inward FDI and avoid mixing within time series. In the end, our three-dimensional database contains just over 300,000 observations with a country coverage of 214. With the most detailed sectoral and bilateral information, the constructed three-dimensional dataset forms the backbone of our data in the subsequent gravity estimation and extrapolation.

d) Negative values

Although rare in FDI stock statistics, negative values do exist in the two-dimensional and three-dimensional datasests but they only account for less than 5% of the total observations across all datasets. According to the OECD, 'negative FDI positions largely result when the loans from the affiliate to its parent exceed the loans and equity capital given by the parent to the affiliate'. As a result, as pointed out in Gouel et al. (2012), they do not carry significant economic meaning and we follow their approach in treating negative values as zeros.

e) Time, sector & country coverage

The implementation of OECD's BMD4 was first applied to FDI data in 2013. In order to make full use of the improvements introduced in the new FDI definition as well as to avoid structural breaks in the data, this necessarily constrains the time dimension of our three-dimensional dataset to a total of 6 years (2013 - 2018), note that this does not affect the constraint datasets at the one- and two-dimensional level as they are only collected for the year 2014 and 2017.

In terms of sector coverage, the Eurostat, OECD, and ITC dataset are in accordance with the 4th Revision of the ISIC[6] whereas the GTAP database adopts its own GTAP Sectoral Classification (GSC). An one to one mapping is not feasible given the differences between the two classifications and data availability. Instead, based on the concordance table between the two classifications, we are able to map out a more aggregated set of 34 sectors that fully incorporates the original 65 GTAP sectors bar some minor discrepancies. For more information and the complete list of sectors, refer to Table A1 in the Annex section.

The complete list of countries and their availability at each dimension can be found in Table A2, the table also contains correspondence between the International Organization for

[6] Technically speaking, the Eurostat data is based on the 2ed Revision of the statistical classification of economic activities in the European Community (NACE 2) which is derived from the 4th Revision ISIC, however sector classification between the two are identical at the first and second level

Table 2: Summary of Constraint Datasets

FDI Dataset	Benchmark Data	Supplement Data	No. of countries		No. of observations		Total FDI Value ($ tn.)	
			2014	*2017*	*2014*	*2017*	*2014*	*2017*
One-Dimensional								
Inward	IMF	UNCTAD	170	168	170	168	30.2	37.0
Outward	IMF	UNCTAD	149	154	149	154	31.7	39.1
Two-Dimensional								
*Bilateral**	IMF Inward	IMF Mirroring	245	245	22,073	24,880	31.0	37.6
Sectoral-Inward	OECD	Eurostat + ITC	62	63	1,398	1,402	21.0	24.2
Sectoral-Outward	OECD	Eurostat + ITC	50	50	1,157	1,185	24.8	28.9

*After applying mirroring technique

Standardization (ISO) and GTAP country code which will be relevant for the final reconciliation in ??.

We conclude the section with Table 2 which displays a summary of the constraint datasets in respect of their data components, country coverage, the number of observations, and total FDI value across all countries and sectors in the year 2014 and 2017. Through merging databases and adopting mirrored values, we are able to achieve near universal country coverage for all datasets except at the two-dimensional sectoral level. In terms of total FDI values, it is somewhat surprising to see that they are fairly consistent between the two one-dimensional datasets and the bilateral dataset (after adopting mirrored values). This was not case for the 2004 data in the original exercise[7] and we believe it reflects the improvements in both data coverage and quality over the years. Finally, it is also worth pointing out that although the one-dimensional and two-dimensional sector-outward datasets both has less country coverage than their inward counterparts, the reported total FDI values are always higher. This reflects the fact that most missing countries in the datasets are developing nations which are typically net importer of foreign capital.

3 Gravity Estimation

As introduced in the last section, our gravity estimation relies on the constructed three-dimensional dataset with more than 300,000 observations covering FDI position in 34 sectors between 214 countries over a period of six years (2013-2018)[8]. Note that the 300,000 observations represent just a tiny fraction of the complete dataset with more than 9 million possible entries[9], therefore the aim of our estimation is to acquire a set of reliable coefficients on various bilateral and country-specific characteristics which are subsequently used to extrapolate for all missing values. To proceed, we begin the section by taking a closer look at

[7]See Appendix A, Table 4 in Gouel et al. (2012)
[8]After the adoption of mirrored values
[9]$214 \times 213 \times 34 \times 6 = 9,298,728$

the three-dimensional dataset, focusing specifically on the distribution of observations and FDI stock across time, countries and sectors.

3.1 Data Description

a) Dependent variable - Three-Dimensional FDI Stocks

Table 3: Distribution by Year

	2013	2014	2015	2016	2017	2018
No. of Observations	61,224	60,625	53,008	53,787	53,021	27,413
Total FDI Values ($ tn.)	12.1	12.5	13.2	13.6	16.1	10.1

As shown in Table 3, we observe a fairly even distribution of observations across the 6-year time period while the total value of FDI stocks slowly increases over time. The year 2018 stands out to be the exception: there is a sharp drop of reported observations which naturally leads to a decrease in total FDI values. Being the most recent year in which data is available, this is a result of confidential values withheld by the reporting countries.

Table 4: Distribution by Country Group

	North-North*	North-South	South-North	South-South
No. of Observations	135,546	73,467	71,448	28,617
Total FDI Values ($ tn.)	63.1	8.96	5.11	0.56

*'North' refers to 'Advanced Economies' as classified by the IMF

When it comes to the distribution of observations by country group, the geographical bias of the Eurostat database is most evidently displayed in Table 4 whereby South-South FDI positions only account for less than 10% of the total observations. Similarly, the value of FDI stock with developing countries as the host country, or in other words the combination of South-North and South-South FDI positions, stands at just $5.67 trillion which is again less than 10% of the total value. However, the extent of underrepresentation is not immediately clear: while investments received by developing countries are certainly the dominant trend in recent years, FDI stocks as cumulative values of all past investment are inherently poorer in reflecting changes in recent periods.

Finally, moving on to Table 5, the decision to merge the 65 GTAP sectors into a more aggregated set of 34 sectors has proven to be fairly balanced since there are more than 7,000 observations for almost every sector. However, regarding the total FDI values, there is large heterogeneity across sectors. This is expected as each individual sector possesses widely different characteristics that determine their attractiveness to potential investors. For instance, it is perhaps not surprising to see that the finance sector (GTAP 57) takes the

Table 5: Distribution by Sector (GTAP Sector Classification)

	1-14 (ag)	15-18 (coaoilgas)	19-26 (foodbt)	27-28 (texwap)	30-31 (lumppp)	32 (p_c)
No. of Observations	7,098	12,737	7,954	7,983	7,490	8,144
Total FDI Values ($ tn.)	0.125	2.565	3.496	0.165	0.392	2.128
	33 (chm)	**34 (bph)**	**35 (rpp)**	**37-39 (metals)**	**40 (ele)**	**42 (ome)**
No. of Observations	7,961	7,050	7,235	7,469	11,947	7,853
Total FDI Values ($ tn.)	4.007	0.997	0.350	0.759	1.083	1.222
	43 (mvh)	**44 (otn)**	**29,36,41,45 (otherman)**	**46-47 (elygas)**	**48 (wtr)**	**49 (cns)**
No. of Observations	7,551	7,590	7,433	12,777	12,505	12,591
Total FDI Values ($ tn.)	1.209	0.202	2.074	1.348	0.061	0.644
	50 (trd)	**51 (afs)**	**52 (otp)**	**53 (wtp)**	**54 (atp)**	**55 (whs)**
No. of Observations	12,208	12,765	7,742	7,671	7,745	7,070
Total FDI Values ($ tn.)	7.088	0.296	0.158	0.213	0.014	0.252
	56 (cmn)	**57 (ofi)**	**58 (ins)**	**59 (rsa)**	**60 (obs)**	**61 (ros)**
No. of Observations	13,052	7,439	7,577	8,356	12,701	13,310
Total FDI Values ($ tn.)	2.155	31.857	1.345	2.662	7.287	0.209
	62 (osg)	**63 (edu)**	**64 (hht)**	**65 (dwe)**		
No. of Observations	6,329	7,973	8,022	7,750		
Total FDI Values ($ tn.)	0.031	0.015	0.055	1.239		

lion's share of total FDI values with more than $30 trillion. On the other hand, the public administration (GTAP 62) and education (GTAP 63) sector are both typically nationalised or with high entry barriers, therefore only accounting for $31 billion and $15 billion of global FDI respectively.

b) Explanatory Variables

We adopt the same explanatory variables as in Gouel et al. (2012). More specifically, we regress FDI stocks on the log of bilateral distance, log of GDP & GDP per capita for both the host and source country, and four dummy variables that take the value of 1 if a country-pair shares the same language; if a country-pair has a colonial relationship; if the host country is a developing country; and if the source country is a developing country.

Our data on GDP and GDP per capita is mainly sourced from the UN National Accounts[10] and they are included under a gravity equation as proxy for a country's market size (GDP)

[10]Exceptionally, data from Taiwan, China and Faroe Islands is supplemented from their respective national agencies and data for Tanzania is sourced from the World Bank

and the level of development (GDP per capita). The bilateral variables (distance, colonial relationship and common language) are sourced from the CEPII database and they are meant to capture bilateral transaction cost and cultural proximity.

Finally, we rely on the 'Advanced Economies' status as classified by the IMF to categorise countries into the developing and developed group. Note that there isn't a single, universally adopted definition for developing/devolped countries and often they are self-declared (e.g. Developed Country Members of the WTO) or determined solely based on income (e.g. World Bank high income countries). Ultimately, the IMF classification represents a more balanced approached which is based on a combination of per capita income, export diversification and degree of integration into the global financial system, therefore it has the unique advantage of excluding resource-rich exporters and accounting for the developed of financial markets which play a crucial role in facilitating FDI.

3.2 Methodology and Model Specification

The gravity model is probably one of the most frequently used 'workhorse tools' in analysing the determinants international trade flow, even at its simplest form, it is capable of returning consistent results with high goodness-of-fit. Outside of the trade literature, the model has been extended and remained empirically robust in other fields such as international migration and in this case, FDI. Aside from pure empirical application, theoretical models of FDI under a general equilibrium framework have also been developed by Bergstrand and Egger (2007) and Head and Ries (2008). Moreover, in recent years this is complemented by ever more sophisticated model from Ramondo and Rodríguez-Clare (2013) that explored interactions between trade and multinational productions which was in turn further expanded to include domestic innovation (Arkolakis et al., 2018) and sector heterogeneity (Alviarez, 2019).

With a well-established theoretical underpinning, we begin by replicating the previous exercise with identical estimation strategy which implies the adoption the PPML estimation technique with year (η_t) and sector (μ_k) fixed effects. Together with the same explanatory variables updated to 2018, the pooled gravity equation is shown below with k, t, i, j representing sector, time, host country and source country respectively:

$$\begin{aligned} FDI_{ijtk} = \exp(&\beta_1 + \beta_2 \ln(GDP_{ti}) + \beta_3 \ln(GDP_{tj}) + \beta_4 \ln(GDPpc_{ti}) \\ &+ \beta_5 \ln(GDPpc_{tj}) + \beta_6 \ln(DIST_{ij}) + \beta_7 COMLG_{ij} \\ &+ \beta_8 COL_{ij} + \beta_9 DEV_i + \beta_{10} DEV_j + \eta_t + \mu_k + \epsilon_{ijtk}) \end{aligned} \tag{1}$$

Upon evaluating the original methodology, we believe that it remains suitable for this update.

Given that zero observations account for just over 80% of our total observations and will be lost when taking logs, the PPML estimator allows the inclusion of these observations while also mitigating potential bias arises from heteroskedasticity in the data (Santos Silva and Tenreyro, 2006).Note that although alternative estimators such as the Negative Binomial Pseudo Maximum Likelihood (NBPML) has been proposed in recent years[11], there is no consensus in the literature and the PPML estimator remains one of the 'golden rules' of gravity estimation.

In terms of the sector and year fixed effects, they are included in consideration of sector heterogeneity and to pick up common shocks over time. Nevertheless, the above combination of fixed effects does not adequately account for the level of transaction cost or entry barriers relative to third countries, and the omission of these multilateral resistance terms could lead to bias in the estimated coefficients (Anderson and van Wincoop, 2003). Instead, the standard approach proposed by Olivero and Yotov (2012) is to include importer-year and exporter-year fixed effects under a gravity panel setting. However, by doing so one can no longer identify country specific variables such as GDP as they will be absorbed by the fixed effects. Given that our primary objective is to extrapolate for missing countries and country-pairs, this necessarily constrains our ability in limiting estimation bias.

After the initial replication, we make two extensions to the model specification before proceeding to extrapolation. First, we include three country dummies as separate regressors which each takes the value of 1 if either the host or source country is Luxembourg, the Netherlands or Hong Kong, China. As discussed in Section 3(c), during the construction of this three-dimensional dataset all investments made by SPEs are subtracted as a mean to exclude transit-FDI. However, non-SPE values are not available for more than half of the observations and therefore significant distortions remain, especially in the three aforementioned economies. As an illustration, 89 out of the highest 100 reported values (FDI_{ijtk}) occurs with one of these three economies as the host or source country, and they are mostly concentrated in the financial sector (GTAP 57). Consequently, they are included as an additional control for transit-FDI with the aim to limit potential distortions on the estimated coefficients.

Second, with ample observations available in each sector as well as the large variation in FDI values (See Table 5), we deviate from the original approach of pooled estimation and instead regressing FDI separately for each one of the 34 sectors. Not only is having sector-sepcific coefficients more intuitive than sector fixed effects, it also introduce sector variation between country-pairs in the subsequent extrapolation, whereas under pooled estimation, a host country's inward FDI in a specific sector as a share of total FDI from any given partner is always the same regardless of the partner country. Together with the addition of the three

[11]See, for example, (Burger et al., 2009) and (Egger and Staub, 2016)

country dummies, the final estimation equation is as followed:

$$
\begin{aligned}
FDI_{ijt,k} = \exp(&\beta_1 + \beta_2 \ln(GDP_{ti}) + \beta_3 \ln(GDP_{tj}) + \beta_4 \ln(GDPpc_{ti}) + \beta_5 \ln(GDPpc_{tj}) \\
&+ \beta_6 \ln(DIST_{ij}) + \beta_7 COMLG_{ij} + \beta_8 COL_{ij} + \beta_9 DEV_i + \beta_{10} DEV_j \\
&+ \beta_{11} LUX_{ij} + \beta_{12} NLD_{ij} + \beta_{13} HKG_{ij} + \eta_{t,k} + \epsilon_{ijt,k})
\end{aligned} \tag{2}
$$

3.3 Estimation Results

The results under the pooled estimation is shown in Table 6, column (1) displays the original estimation output based on FDI stock data from 1994 to 2004, note that the coefficients and standard errors are extracted directly from the regression table in Gouel et al. (2012). Column (2) displays the replicated results under the same gravity equation but based on the updated three-dimensional dataset introduced in the last section. Finally, column (3) displays the results with the three additional country dummies for Luxembourg, the Netherlands, and Hong Kong, China.

Looking at the original and replicated results in the first two columns, it is reassuring that almost all the variables remain statistically significant to the 1% with theory-consistent signs. The positive and significant source developing country dummy and insignificant common language dummy from the replicated result may be counter-intuitive, but as we will show later, these variables are more sensitive to changes in model specification. Turning to the magnitude of coefficients, we observe smaller absolute values for distance and GDP-related variables which suggest a decrease in investment barriers and transaction cost as well as a greater presence of smaller and less developed nations (with lower GDP and GDP per capita). However, it must be emphasise that the comparisons above must be taken with a grain of salt given the expansion of data coverage in the Eurostat database and the improvements in data quality under the BMD4 guideline, moreover the construction of our three-dimensional dataset also differs from the original approach in procedures such as the treatment of FDI made by SPEs and the adoption of mirrored values[12].

Moving on to column (3), the only difference there is the inclusion of the three aforementioned country dummies. However, the benefits from addressing these three outliers are clearly visible from the regression table: aside from being positive and highly significant, the regression returns a much higher fit with the pseudo r-squared increasing from 0.573 to 0.688. Regarding the other explanatory variables, we observe an increase in the magnitude of the GDP coefficients whereas on the other hand, there is a drop in the GDP per capita

[12]In the previous exercise, separate distinction of investments made by SPEs are not available in the Eurostat database, and discrepancies between inward and mirrored observations are reconciled by taking the average of the two

Table 6: Pooled Estimation (PPML Estimator)

VARIABLES	(1) Original†	(2) Replication	(3) Country Dummies
ln_DIST	-0.756***	-0.588***	-0.428***
	(0.01)	(0.0411)	(0.0468)
ln_GDP_host	0.858***	0.531***	0.673***
	(0.009)	(0.0270)	(0.0275)
ln_GDP_source	0.775***	0.615***	0.827***
	(0.009)	(0.0435)	(0.0435)
ln_GDPpc_host	2.417***	1.195***	0.611***
	(0.061)	(0.0892)	(0.0877)
ln_GDPpc_source	2.995***	1.718***	1.211***
	(0.061)	(0.132)	(0.109)
Developing Country_host	1.558***	1.282***	0.887***
	(0.086)	(0.122)	(0.139)
Developing Country_source	-0.345	1.533***	1.255***
	(0.215)	(0.223)	(0.239)
Common Language	0.683***	0.0569	0.694***
	(0.029)	(0.130)	(0.0715)
Colonial Relationship	0.136***	0.744***	0.0793
	(0.033)	(0.181)	(0.109)
LUX Dummy	-	-	2.467***
			(0.121)
NLD Dummy	-	-	2.074***
			(0.0872)
HKG Dummy	-	-	2.272***
			(0.259)
Observations	68,230	307,621	307,621
Year F.E.	Yes	Yes	Yes
Sector F.E.	Yes	Yes	Yes
Pseudo R-squared	-	0.573	0.688

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

† Source: Directly extracted from Gouel et al. (2012)

variables, this corrects for the distortions brought by the three financial centres which are some of the most developed economies in the world while being relatively small in terms of their absolute sizes. Note that the common language dummy becomes significant again whereas the colonial dummy is no longer significant, this is also intuitive as we believe that they reflect the colonial history of Hong Kong and the Netherlands.

In the final specification, we run the estimations separately for the 34 sectors, the results are too numerous to be shown here and they are instead listed in Table B1, B2, B3, and B4 of Annex B. Generally speaking, the country dummies, distance and GDP-related coefficients remain robust across all sectors with only a few exceptions. The common language, colonial relationship and the two developing country dummies are also fairly consistent but with higher frequency of returning insignificant or counter-intuitive (opposite sign) coefficients across various sectors. Considering the fact that multinationals face widely different consumer preferences, transaction cost and entry barriers across sectors, we are satisfy with the overall results of the sector-specific coefficients and they are adopted for the subsequent extrapolation.

3.4 Sensitivity Analysis (Summary)

Despite the inclusion of the three outlier country dummies and estimating sector by sector, the final model specification is similar to the original regression in Gouel et al. (2012) and it remains parsimonious in nature. However, this specific gravity equation is not chosen *a priori*, instead we have sought to test its robustness by including additional explanatory variables meant for limiting potential omitted variables bias as well as using different versions of the three-dimensional dataset by slightly adjusting its the construction procedures. In the end, the performance of these additional variables are poor and we do not observe a noticeable increase in the goodness of fit, plus the different datasets return similar results which all point to the reliability of our model specification and of the way we have constructed the three-dimensional dataset.

Aside from looking at the fit and coefficients under different model specifications, we have also explored their predictive power by conducting out-of-sample validations. More specifically, the observations are first randomly and repeatedly spilt into pairs of training and testing groups, we then run the different specifications under the training groups and extrapolate for the testing groups. Finally, the mean absolute errors (MAEs) are acquired by taking the difference between the extrapolated values and actual reported data from the testing groups. Upon comparison, a lower MAE would suggest higher predictive power and our results indicate that both the inclusion of countries dummies and sector by sector estimation significantly decrease the MAE.

For a walkthrough of the set up of the robustness checks and out-of-sample validations as well as a more extensive discussion of the results, they can be found in the Annex Section C.

4 Extrapolation Results

Using the sector-specific coefficients attained above, we extrapolate to fill in and fully complete the three-dimensional dataset which is now comprised of 204 countries (each with 203 partners), all 34 sectors and for all 6 years. In summary, there are now more than 8 million observations in the dataset, the vast majority of them are extrapolated values and original reported FDI made up just 3.4% of the total observations. Note that the CEPII database and GDP information is not available for a few island countries and oversea territories, therefore the final coverage is slightly smaller than the original list of 214 countries.

Following the initial extrapolation, we compare the dataset with the previously constructed one-dimensional and two-dimensional bilateral datasets that are used as optimisation constraints (See Section 2.3), and we specifically focus on the year 2014 and 2017 as they are the reference years for GTAP v10 and v11 respectively. The comparison reveals various features of the extrapolated dataset while also serving as a rough gauge to the extend of adjustments needed in the subsequent quadratic optimisation procedures.

Table 7: Total World FDI by dataset (Unit: Trillion USD)

	1-D Inward	**1-D Outward**	**2-D Bilateral**	**3-D Extrapolate**
2014	30.1	30.7	30.8	38.1
2017	36.9	37.2	37.4	42.3

Beginning at the top, we can aggregate upwards to acquire the total world FDI for both the three-dimensional extrapolated dataset and the constraint datasets[13], and the corresponding values in the year 2014 and 2017 are listed in Table 7. Upon comparison, we can observe that in terms of total FDI, the extrapolated dataset slightly overshoot against the target, but overall it still matches fairly well with the three constraint datasets at the one- and two-dimensional level.

We now proceed to compare the datasets at the one-dimensional level, meaning that for the extrapolated and the two-dimensional bilateral datasets, we aggregate to acquire the total inward and outward FDI stock for each country. Next, as shown in Figure 1, we have ranked the 10 smallest and biggest FDI recipients in the extrapolated dataset, and once again they are compared against actual reporting in the constraint datasets. On the left-hand side of the figure, it is clear that our model over-extrapolated for countries with small inward FDI, and sometimes by a big margin. The difficulty in extrapolating for small values is not surprising given the fact that our gravity estimation is primarily based on data collected from the Eurostat database which over-represents developed, European nations. Moreover, this most likely explains the overshoot in the total World FDI stock discussed above. On the other hand, the extrapolated results are a lot more accurate for bigger nations with high inward FDI, the only exceptions from the top 10 list are two oil-rich country – Qatar and United Arab Emirates. This is also expected since their high GDP and GDP per capita mainly reflect rents from natural resources rather than overall integration with the global

[13]We exclude the two-dimensional sectoral-inward and sectoral-outward datasets in this particular comparison since the total FDI is much lower due to their lack of country coverage

Figure 1: Total Inward FDI by country (Year 2017)

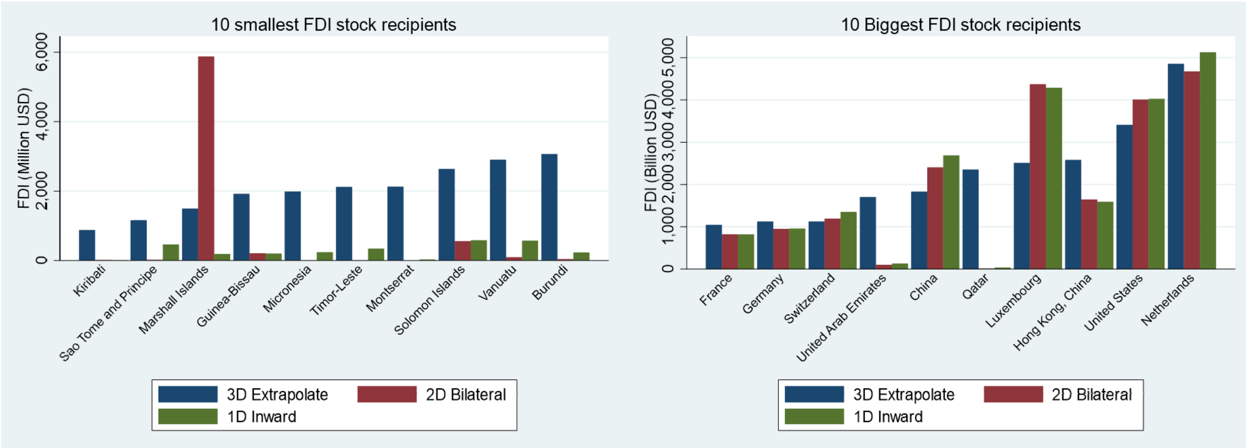
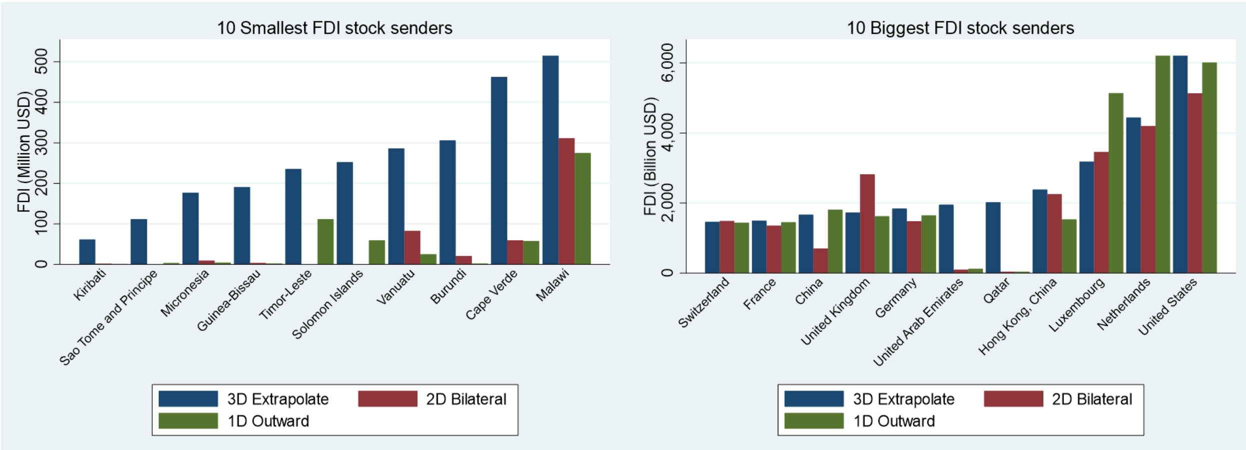


Figure 2: Total Outward FDI by country (Year 2017)



economy. Owing to a parsimonious approach, our model specification is unable to capture these country-specific characteristics.

5 Quadratic Optimisation

After filling in the missing values using econometric extrapolation, the final consistency of the database is ensured using a quadratic optimisation technique that allows us to incorporate and reconcile information from different constraint datasets and produce a balanced database. We adopt a similar approach as described in Gouel et al. (2012) whereby the objective function of the quadratic optimisation is to minimise the difference between initial estimates and final values subject to adding up constraints. For a given sector k, host country i and source country j and reliability weights w, the quadratic optimisation is implemented as follows:

Minimise:

$$\sum_{kij}(FDI^1_{kij} - FDI^0_{kij})^2/w_{kij} + \sum_{ij}(FDI^1_{ij} - FDI^0_{ij})^2/w_{ij} + \sum_{ki}(FDI^1_{ki} - FDI^0_{ki})^2/w_{ki}$$

$$+ \sum_{kj}(FDI^1_{kj} - FDI^0_{kj})^2/w_{kj} + \sum_{i}(FDI^1_{i} - FDI^0_{i})^2/w_{i} + \sum_{j}(FDI^1_{j} - FDI^0_{j})^2/w_{j}$$

Subject to:

$$\sum_{kij} FDI^1_{kij} = FDI_i^{\text{1D\_Inward}}$$

$$\sum_{kij} FDI^1_{kij} = FDI_j^{\text{1D\_Outward}}$$

$$\sum_{k} FDI^1_{kij} = FDI_{ij}^{\text{Bilateral}}$$

$$\sum_{j} FDI^1_{kij} = FDI_{ki}^{\text{Sectoral\_Inward}}$$

$$\sum_{i} FDI^1_{kij} = FDI_{kj}^{\text{Sectoral\_Outward}}$$

where the FDI^0 variables denotes the initial sector-host-source specific FDI data constructed in the previous step and FDI^1 denotes the final values resulting from the optimisation. In addition to the three-dimensional data, we complement the dataset with the five constraint datasets constructed in Section 2.3 – the one-dimensional inward/outward, the bilateral, and the two-dimensional sectoral inward/outward datasets. The constraints of the optimisation are aimed to target these aggregate values in the five aforementioned datasets. Reliability weights are chosen to reflect our confidence in the dependability of the underlying data: higher weights increase the penalty for deviating from the initial values and so are used with data in which we have greater confidence; correspondingly, lower weights are used for less certain data.

6 Phantom FDI

When looking at National FDI statistics, some countries appear to be large investors despite their relatively small economic size. This distortion in the data has been defined as *Phantom FDI*. A methodology to accommodate for this flaw consists in taking into account the likelihood that a country is a tax haven and thus tracing back the FDI stock to the original investors. Thanks to OECD data on ultimate investors, we can compare the results of this estimation method with actual data: the results are extremely accurate, as discussed in Casella (2019) seminal paper. The aims of this section are the replication and extension of Casella's framework to exploit more granular data on tax havens.

6.1 The Framework

Suppose that, starting from FDI bilateral data on m countries , we can build the following square matrix:

$$FDI = \begin{bmatrix} FDI_{11} & FDI_{12} & \dots & FDI_{1m} \\ FDI_{21} & FDI_{22} & \dots & FDI_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ FDI_{m1} & FDI_{m2} & \dots & FDI_{mm} \end{bmatrix}, \tag{3}$$

where each row represent a country j's share of inward FDI coming from country i. By default, the diagonal of this matrix is equal to zero: that is, the reported share of a country's FDI into itself is zero. This is a first shortcoming of the standard FDI reporting: it rules out double counting. However, the main concern regarding the *FDI* matrix constructed above is that it does not take into account the fact that a large part of FDI investment from a special set of countries, which we will denote as conduit, is not actually coming from the conduit countries themselves but is only *passing through* them for fiscal and financial purposes. Therefore, the *ultimate* investors differ from the *reported* ones: it is natural that in the latter the role of conduit countries in global FDI is overstated. How can one take into account this source of bias? The entries of 3 can be seen as the "probability" that a country j has received FDI from country i: therefore, Casella argues, we can look at FDI shares as if they were stochastic processes, i.e. random variables observed in different points in time. To address the presence of conduit countries, Casella proposes to represent FDI flows as stochastic processes with two important properties: Markovianity and homogeneity. The ultimate investors matrix can be thus represented as follows:

$$\underset{m \times m}{P^*} = \frac{R}{I - Q}, \tag{4}$$

where I is the identity matrix. Furthermore,

$$Q = p_d(j,i)p_c(i)_{j=1...m,i=1...m}, \tag{5}$$

and

$$R = p_d(j,i)(1 - p_c(j,i))_{j=1...m,i=1...m}. \tag{6}$$

$p_d(j,i)$ represents the probability that a country j's inward FDI comes from country i. On the other hand, $p_c(i)$ represents the probability that country i is a conduit jursidiction for country j. Whereas $p_d(j,i)$ can be estimated naturally as the share of j's reported inward FDI coming from i like in Equation 7 , $p_c(i,j)$ needs further assumptions.

6.2 The baseline replication

In order to obtain P^*, I have computed

$$p_d(j,i) = \frac{FDI(j,i)}{\sum_{i=1}^{m} FDI(j,i)}, \tag{7}$$

whereas $p_c(i,j) = 1 \forall j$ for a list of 37 countries[14], $p_c(j,i) = \frac{SPE_{ij}}{FDI_i^{OUT}j}$ for countries with available bilateral data on special purpose entities in the OECD database. For countries with only unilateral data on SPEs outward FDI, $p_c(i) = \frac{SPE_i}{FDI_i^{OUT}}$. [15] Since Hong Kong, Ireland and Singapore do not report SPEs data but are known as SPEs hubs, we approximated their conduit probabilities with the *implied investment approach*. That is, exploiting UNCTAD rich dataset on Outward FDI, we run the following specification on the top quartile of countries in terms of FDI:

$$FDI_i = \beta_1 * GDP_i + \varepsilon_i. \tag{8}$$

After this, we took $F\tilde{D}I_i = F\hat{D}I_i + 1.645 * sig\hat{m}a_i$, the upper bound of the 90% confidence interval of the linear prediction and computed

$$p_c(i) = \frac{F\tilde{D}I_i - FDI_i}{FDI_i}. \tag{9}$$

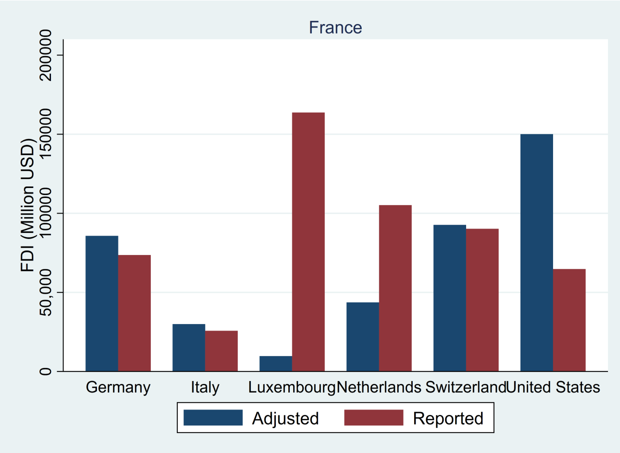
This heuristic methodology allows thus to obtain unilateral conduit probabilities as the excess outward FDI of a country when SPEs data are not available; this is the case for Singapore, Hong Kong and Ireland.

[14]Anguilla, Antigua and Barbuda, Aruba, Bahamas, Bahrain, Belize, Bermuda, Virgin Islands, British Cayman Islands, Cook Islands, Cyprus, Dominica, Gibraltar, Grenada, Guernsey, Isle of Man, Jersey, Liberia, Liechtenstein, Malta, Marshall, Islands Mauritius, Monaco, Montserrat, Nauru, Netherlands, Antilles, Niue, Panama, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Samoa, San Marino, Seychelles, Turks, and Caicos Islands, U.S. Virgin Islands, Vanuatu which are all deemed as 'pure' tax heavens

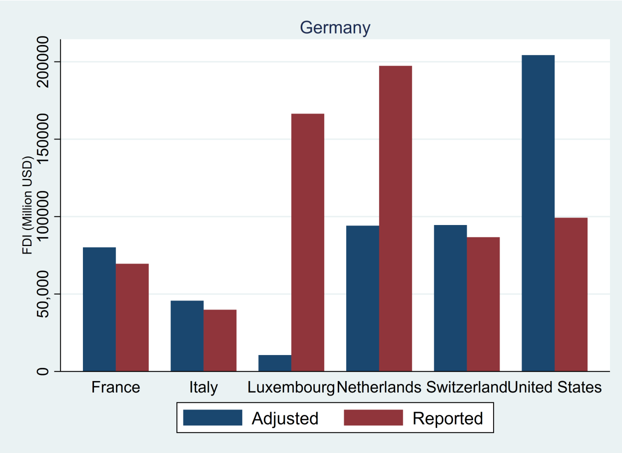
[15]Note that the introduction of bilateral conduit estimation is a novelty in the literature.

Figure 3: Reported vs. Adjusted comparison

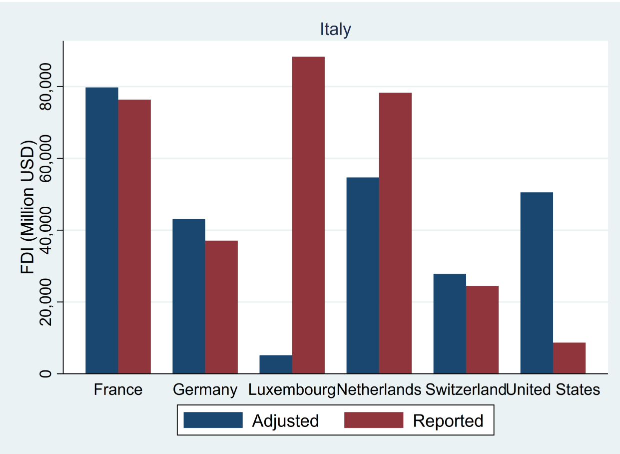
(a) France Inward FDI



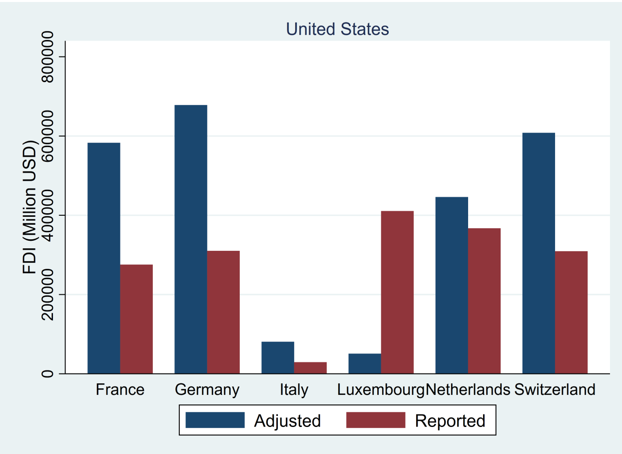
(b) Germany Inward FDI



(c) Italy Inward FDI



(d) USA Inward FDI



6.3 Results

The results of the adjustment downsize significantly the role of Netherlands and Luxembourg, which is the main goal of this methodology. Symmetrically, the adjustment spills FDI to Western Countries that invest *in conduit jurisdictions* and are thus under-reported in official data.

7 Conclusion

The purpose of this paper is to build a three-dimensional global FDI stock database with the most recent FDI stock data for CGE analysis. We follow closely the procedure as adopted in Gouel et al. (2012). We collected raw FDI stock data from various data sources, including IMF, UNCTAD, OECD, EuroStat and ITC investment map, adopt an econometric framework to fill in the missing values, and uses a quadratic optimisation procedure to balance the dataset.

There are a couple differences introduced in our project: first, we only update the FDI stock data and omit the FDI flow as literature tends to adopt stock values as proxy of foreign affiliate activities due to the volatile nature of FDI flow. Second, we adopt a novel computation method developed by Casella (2019) that seeks to identify "phantom FDI" and reallocate them according to their ultimate investor country. Third, we made a couple extensions in our econometric framework: we include three country dummies as separate regressors which each takes the value of 1 if either the host or source country is Luxembourg, the Netherlands or Hong Kong, China. These dummies are included as an additional control for transit-FDI in order to limit potential distortions on the estimated coefficients. After running a pooled estimation to compare our results with Gouel et al. (2012), we further deviate from the pooled estimation approach and instead regress FDI separately for each one of the 34 sectors. In this case, we are able to get sector-specific coefficients which allows us to introduce sector variation between country-pairs when performing data extrapolation.

We subsequently adopt a quadratic optimisation procedure to balance the dataset and reconcile original FDI data from different data sources. After adjusting for phantom FDI, we construct a three-dimensional FDI stock database with two baseline years: 2014 and 2017, which is compatible with the GTAP version 10 and version 11 database.

References

Alviarez, V. (2019). Multinational production and comparative advantage. *Journal of International Economics*, 119:1–54.

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

Arkolakis, C., Ramondo, N., Rodríguez-Clare, A., and Yeaple, S. (2018). Innovation and Production in the Global Economy. *American Economic Review*, 108(8):2128–2173.

Bergstrand, J. H. and Egger, P. (2007). A knowledge-and-physical-capital model of international trade flows, foreign direct investment, and multinational enterprises. *Journal of International Economics*, 73(2):278–308.

Burger, M., Oort, F. v., and Linders, G.-J. (2009). On the Specification of the Gravity Model of Trade: Zeros, Excess Zeros and Zero-inflated Estimation. *Spatial Economic Analysis*, 4(2):167–190.

Casella, B. (2019). Looking through conduit FDI in search of ultimate investors – a probabilistic approach. *Transnational Corporations*, 26(1):109–149.

Dee, P. and Hanslow, K. (2000). Multilateral liberalisation of services trade. Productivity Commission Staff Research Paper.

Egger, P. H. and Staub, K. E. (2016). GLM estimation of trade gravity models with fixed effects. *Empirical Economics*, 50(1):137–175.

Gouel, C., Guimbard, H., and Laborde, D. (2012). A Foreign Direct Investment database for Global CGE models. *CEPII Working Paper*, No. 2012-08:35.

Haftel, Y. Z. (2010). Ratification counts: US investment treaties and FDI flows into developing countries. *Review of International Political Economy*, 17(2).

Head, K. and Ries, J. (2008). FDI as an outcome of the market for corporate control: Theory and evidence. *Journal of International Economics*, 74(1):2–20.

Kox, H. L. M. and Rojas-Romagosa, H. (2020). How trade and investment agreements affect bilateral foreign direct investment: Results from a structural gravity model. *The World Economy*, 43(12):3203–3242.

Lakatos, C. and Fukui, T. (2014). The Liberalization of Retail Services in India. *World Development*, 59:327–340.

Lakatos, C. and Walmsley, T. (2011). A Global Multi-sector Multi-region Foreign Direct Investment Database for GTAP. *GTAP Research Memoranda*, No. 18.

OECD (2014). Measuring international investment by multinational enterprises - implementation of the oecd's benchmark definition of foreign direct investment, 4th edition.

Olivero, M. P. and Yotov, Y. V. (2012). Dynamic gravity: endogenous country size and asset accumulation. *Canadian Journal of Economics/Revue canadienne d'économique*, 45(1):64–92.

Petri, P. A. (1997). Foreign Direct Investment in a Computable General Equilibrium Framework. (Paper prepared for the conference: Making APEC Work Economic Challenges and Policy Alternatives).

Ramondo, N. and Rodríguez-Clare, A. (2013). Trade, Multinational Production, and the Gains from Openness. *Journal of Political Economy*, 121(2):273–322.

Santos Silva, J. and Tenreyro, S. (2006). The Log of Gravity. *The Review of Economics and Statistics*, 88(4):641–658.

van der Marel, E. and Shepherd, B. (2020). Trade facilitation in services : Concepts and empirical importance. *World Bank Policy Research Working Paper*, No. 9234.

World Investment Report (2019). *World Investment Report 2019: Special Economic Zones.* United Nations Conference on Trade and Development, Geneva.

Annex A

1 Sectoral Coverage

Table A1: Full Sector List

Sector Description	Eurostat/OECD Code	ISIC Rev. 4	GTAP 10/11
Agriculture, forestry and fishing	A	01-03	1-14
Mining and quarrying	B	05-09	15-18
Manufacturing			
Manufacture of food products; beverages and tobacco products	C10_12	10-12	19-26
Manufacture of textiles and wearing apparel	C13_14	13-14	27-28
Manufacture of wood, paper, printing and reproduction	C16_18	16-18	30-31
Manufacture of coke and refined petroleum products	C19	19	32
Manufacture of chemicals and chemical products	C20	20	33
Manufacture of basic pharmaceutical products and pharmaceutical preparations	C21	21	34
Manufacture of rubber and plastics products	C22	22	35
Manufacture of basic metals and fabricated metal products, except machinery and equipment	C24_25	24-25	37-39
Manufacture of computer, electronic and optical products	C26	26	40
Manufacture of machinery and equipment n.e.c.	C28	28	42
Manufacture of motor vehicles, trailers and semi-trailers	C29	29	43
Manufacture of other transport equipment	C30	30	44
Other manufacturing	C_OTH	15, 23, 27, 31, 32, 33	29,36,41,45
Electricity, gas, steam and air conditioning supply	D35	35	46-47
Water supply; sewerage, waste management and remediation activities	E	36-39	48
Construction	F	41-43	49
Wholesale and retail trade; repair of motor vehicles and motorcycles	G	45-47	50
Accommodation and food service activities	I	55-56	51
Transportation and storage			
Land transport and transport via pipelines	H49	49	52
Water transport	H50	50	53
Air transport	H51	51	54
Warehousing and support activities for transportation	H52	52	55
Postal and courier activities	H53	53	56
Information and communication	J	58-63	56
Financial and insurance activities			
Financial service activities, except insurance and pension funding	K64	64	57
Insurance, reinsurance and pension funding, except compulsory social security	K65	65	58
Real estate activities	L	68	59
Professional, scientific and technical activities	M	69-75	60
Administrative and support service activities	N	77-82	60
Arts, entertainment and recreation	R	90-93	61
Other service activities	S	94-96	61
Public administration; activities of households and of extraterritorial organisations	O_T_U	84, 97-99	62
Education	P85	85	63
Human health and social work activities	Q	86-88	64
Private real estate activities	PRV_RE	n.a.	65

n.a. - not available

Notes on sector grouping

The aggregation of sectors are largely determined by the grouping and data availability of the OECD and Eurostat since almost all of the sectoral information comes from the two databases. Note that both databases report FDI on 81 sectors at different levels, most of these are not shown in the table above, they are discarded for either being too aggregated (e.g. Sector Manufacturing - C) or too detailed for the GTAP Sector Classification (GSP).

Many OECD/Eurostat sectors listed above represents multiple sectors in the ISIC Rev. 4 (two-digit division) and GTAP Version 10/11. Exceptionally, 'Postal and courier activities - H53' and 'Information and communication - J' are grouped to represent GTAP Sector 56; 'Professional, scientific and technical activities - M' and 'Administrative and support service activities - N' are grouped to represent GTAP Sector 60; finally 'Arts, entertainment and recreation - R' and 'Other services activities - S' are grouped to represent GTAP Sector 61.

Sector 'Private real estate activities - PRV_RE' is available in the OECD and Eurostat database and it corresponds to GTAP Sector 65, interestingly this is the only sector outside of the ISIC framework.

Notes on minor mapping discrepancies

Technically speaking, GTAP Sector 61 incorporates 'Arts, entertainment and recreation - R' and 'Othservices activities - S' as well as **'Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use - T'**. However, Sector T is bundled together with Sector O and Sector U in the OECD and Eurostat database as 'Public administration; activities of households and of extraterritorial organisations - O_T_U' and these three sectors are not available separately.

The tables further indicates that GTAP Sector 57 and 58 match one to one with 'Financial service activities, except insurance and pension funding - K64' and 'Other services activities - K65', but strictly speaking GTAP Sector 57 should further include ISIC Sector 661 and 663 and likewise GTAP Sector 58 should include ISIC Sector 662. However, these three-digit ISIC sectors are not available in the OECD or Eurostat database.

2 Country Coverage

Table A2: Full Country List

Pending

Annex B

Table B1: Sector-Specific Estimation (PPML Estimator & GTAP Sector 1-35)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES	ag (1-14)	coaoilgas (15-18)	foodbt (19-26)	texwap (27-28)	lumppp (30-31)	p_c (32)	chm (33)	bph (34)	rpp (35)
ln_DIST	-0.266*	-0.0718	-0.339***	-0.543***	-0.381***	-0.741***	-0.259***	-0.166	-0.950***
	(0.142)	(0.0718)	(0.0894)	(0.127)	(0.0676)	(0.148)	(0.0759)	(0.120)	(0.0770)
ln_GDP_host	0.271**	0.593***	0.689***	1.091***	0.478***	0.832***	0.926***	0.518***	0.767***
	(0.138)	(0.0626)	(0.0534)	(0.0781)	(0.0494)	(0.125)	(0.0572)	(0.103)	(0.0412)
ln_GDP_source	0.911***	0.760***	0.703***	0.807***	0.285***	0.609***	0.674***	0.597***	0.871***
	(0.0600)	(0.0492)	(0.0581)	(0.0540)	(0.0427)	(0.0688)	(0.0457)	(0.0591)	(0.0497)
ln_GDPpc_host	1.862***	1.033***	1.297***	-0.461***	1.670***	1.954***	2.836***	1.760***	0.124
	(0.406)	(0.223)	(0.249)	(0.136)	(0.412)	(0.465)	(0.450)	(0.402)	(0.251)
ln_GDPpc_source	0.397**	1.086***	0.760***	-0.210	1.550***	1.594***	1.367***	1.727***	0.660***
	(0.166)	(0.160)	(0.177)	(0.145)	(0.142)	(0.459)	(0.223)	(0.273)	(0.158)
Developing Country_host	4.421***	2.613***	2.191***	-0.261	2.921***	3.299***	3.871***	1.509***	1.271***
	(0.659)	(0.334)	(0.370)	(0.322)	(0.569)	(0.729)	(0.695)	(0.434)	(0.405)
Developing Country_source	0.138	1.783***	-0.476	-1.736***	-0.426*	1.010	-0.703**	-1.700***	0.190
	(0.414)	(0.282)	(0.324)	(0.315)	(0.235)	(0.652)	(0.351)	(0.359)	(0.350)
Common Language	0.692**	1.414***	0.283	0.730**	0.525**	1.620***	0.445**	-0.208	0.659**
	(0.294)	(0.181)	(0.175)	(0.291)	(0.229)	(0.316)	(0.210)	(0.316)	(0.284)
Colonial Relationship	-0.639***	1.039***	1.159***	0.284	0.671**	0.0168	0.137	1.901***	0.447**
	(0.224)	(0.170)	(0.137)	(0.420)	(0.318)	(0.284)	(0.164)	(0.239)	(0.203)
LUX Dummy	0.184	1.111***	1.229***	3.522***	-0.330	1.053**	1.765***	1.115**	2.080***
	(0.811)	(0.394)	(0.303)	(0.330)	(0.512)	(0.468)	(0.480)	(0.513)	(0.330)
NLD Dummy	0.399*	1.804***	2.775***	1.262***	1.885***	3.373***	3.175***	2.654***	0.380**
	(0.222)	(0.226)	(0.127)	(0.344)	(0.179)	(0.251)	(0.121)	(0.423)	(0.167)
HKG Dummy	1.706***	2.317***	0.758**	2.728***	-0.710	-3.111***	0.946**	-2.294***	2.242***
	(0.353)	(0.308)	(0.369)	(0.391)	(0.563)	(0.444)	(0.372)	(0.479)	(0.615)
Observations	7,060	12,693	7,907	7,936	7,447	8,100	7,917	7,006	7,192
Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sector F.E.	-	-	-	-	-	-	-	-	-
Pseudo R-squared	0.386	0.389	0.531	0.446	0.458	0.529	0.565	0.566	0.508

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table B2: Sector-Specific Estimation (PPML Estimator & GTAP Sector 37-49)

VARIABLES	(1) metals (37-39)	(2) ele (40)	(3) ome (42)	(4) mvh (43)	(5) otn (44)	(6) otherman (29/36/41/45)	(7) elygas (46-47)	(8) wtr (48)	(9) cns (49)
ln_DIST	-0.440***	-0.130**	-0.404***	-0.251***	-0.973***	-0.239***	-1.028***	-1.155***	-0.795***
	(0.0548)	(0.0637)	(0.0456)	(0.0851)	(0.109)	(0.0690)	(0.0895)	(0.0854)	(0.0586)
ln_GDP_host	0.667***	1.052***	0.748***	0.874***	1.296***	0.884***	0.669***	0.734***	0.663***
	(0.0294)	(0.0709)	(0.0366)	(0.0506)	(0.0860)	(0.0708)	(0.0592)	(0.0701)	(0.0352)
ln_GDP_source	0.617***	0.667***	0.619***	0.876***	1.039***	0.700***	0.897***	0.915***	0.608***
	(0.0362)	(0.0532)	(0.0413)	(0.0472)	(0.0621)	(0.0449)	(0.0505)	(0.0556)	(0.0467)
ln_GDPpc_host	-0.182*	0.773***	0.529*	-0.403***	-0.497**	0.805***	-0.275**	-0.408***	-0.169**
	(0.103)	(0.210)	(0.286)	(0.131)	(0.237)	(0.282)	(0.134)	(0.142)	(0.0770)
ln_GDPpc_source	0.778***	0.958***	1.320***	0.280*	-0.479	0.167	-0.434***	0.125	0.313*
	(0.137)	(0.214)	(0.166)	(0.155)	(0.333)	(0.187)	(0.158)	(0.167)	(0.180)
Developing Country_host	0.310	0.902***	1.119***	-0.161	-0.530	1.255***	0.108	0.0881	0.712***
	(0.253)	(0.327)	(0.424)	(0.381)	(0.481)	(0.457)	(0.301)	(0.282)	(0.162)
Developing Country_source	-0.336	-0.558	-1.066***	-2.085***	-3.441***	-1.473***	-2.258***	-3.338***	-0.490*
	(0.400)	(0.441)	(0.280)	(0.365)	(0.772)	(0.399)	(0.376)	(0.392)	(0.297)
Common Language	0.183	-0.348	-0.288**	-0.255	0.483**	-0.0341	1.096***	0.321*	0.809***
	(0.157)	(0.266)	(0.122)	(0.342)	(0.211)	(0.180)	(0.207)	(0.190)	(0.168)
Colonial Relationship	0.172	0.564**	0.822***	-0.824***	0.776***	0.806***	0.438	0.829***	0.436**
	(0.171)	(0.241)	(0.121)	(0.228)	(0.225)	(0.214)	(0.273)	(0.215)	(0.191)
LUX Dummy	2.608***	1.683***	1.320***	3.054***	5.372***	3.284***	2.546***	2.288***	1.612***
	(0.277)	(0.444)	(0.362)	(0.367)	(0.362)	(0.318)	(0.292)	(0.321)	(0.297)
NLD Dummy	0.949***	2.141***	1.909***	1.507***	-0.654**	1.726***	0.886***	0.527***	0.654***
	(0.126)	(0.199)	(0.109)	(0.182)	(0.295)	(0.202)	(0.231)	(0.192)	(0.167)
HKG Dummy	-0.823***	1.192***	0.198	0.324	-2.108***	0.0321	1.705***	3.447***	1.602***
	(0.229)	(0.282)	(0.202)	(0.378)	(0.587)	(0.258)	(0.312)	(0.516)	(0.361)
Observations	7,425	11,903	7,809	7,508	7,547	7,392	12,733	12,461	12,547
Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sector F.E.	-	-	-	-	-	-	-	-	-
Pseudo R-squared	0.539	0.675	0.676	0.562	0.710	0.677	0.540	0.521	0.466

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table B3: Sector-Specific Estimation (PPML Estimator & GTAP Sector 50-59)

VARIABLES	(1) trd (50)	(2) afs (51)	(3) otp (52)	(4) wtp (53)	(5) atp (54)	(6) whs (55)	(7) cmn (56)	(8) ofi (57)	(9) ins (58)	(10) rsa (59)
ln_DIST	-0.404***	-0.649***	-1.032***	-0.593***	-0.937***	-0.492***	-0.446***	-0.237***	-0.442***	-0.971***
	(0.0562)	(0.0643)	(0.116)	(0.0890)	(0.150)	(0.0612)	(0.0711)	(0.0629)	(0.0687)	(0.0656)
ln_GDP_host	0.647***	0.804***	0.752***	0.0899	0.777***	0.696***	0.974***	0.508***	0.845***	0.804***
	(0.0537)	(0.0552)	(0.0799)	(0.189)	(0.138)	(0.0353)	(0.0687)	(0.0467)	(0.0475)	(0.0388)
ln_GDP_source	0.822***	0.759***	0.725***	0.569***	1.090***	0.648***	0.678***	0.784***	0.790***	0.789***
	(0.0390)	(0.0447)	(0.0677)	(0.0939)	(0.164)	(0.0407)	(0.0736)	(0.0544)	(0.0402)	(0.0403)
ln_GDPpc_host	0.510***	-0.00858	-0.0106	0.439*	-0.408	-0.571***	0.403	0.445***	0.587***	-0.393***
	(0.0961)	(0.103)	(0.244)	(0.233)	(0.378)	(0.134)	(0.464)	(0.113)	(0.205)	(0.0719)
ln_GDPpc_source	0.864***	0.330**	0.885***	1.521***	-0.387	0.619***	1.227***	1.814***	0.481***	0.556***
	(0.125)	(0.155)	(0.213)	(0.214)	(0.620)	(0.130)	(0.412)	(0.179)	(0.139)	(0.0859)
Developing Country_host	0.598***	0.452**	1.989**	-1.790***	-0.644	-1.376***	1.446**	-0.108	-0.0667	-0.00641
	(0.138)	(0.215)	(0.803)	(0.363)	(0.966)	(0.376)	(0.662)	(0.236)	(0.337)	(0.170)
Developing Country_source	0.553**	-0.381	-0.252	-2.382***	-3.243**	-1.687***	1.021*	1.729***	-3.430***	0.255
	(0.280)	(0.314)	(0.491)	(0.585)	(1.298)	(0.340)	(0.611)	(0.306)	(0.317)	(0.194)
Common Language	0.915***	0.959***	-0.782**	-1.891***	0.197	0.478***	0.532	0.656***	0.459***	0.626***
	(0.138)	(0.188)	(0.356)	(0.572)	(0.408)	(0.145)	(0.427)	(0.116)	(0.165)	(0.158)
Colonial Relationship	-0.327*	1.342***	-2.031***	-1.210***	1.808**	-0.0588	-0.506	0.193	0.918***	-0.127
	(0.183)	(0.169)	(0.544)	(0.378)	(0.735)	(0.198)	(0.331)	(0.146)	(0.192)	(0.125)
LUX Dummy	2.054***	2.377***	2.051***	-0.614	0.831	2.993***	3.660***	2.149***	0.805**	2.235***
	(0.203)	(0.321)	(0.444)	(0.436)	(1.007)	(0.221)	(0.295)	(0.189)	(0.367)	(0.146)
NLD Dummy	1.444***	-0.154	2.588***	0.688**	1.892***	1.681***	-0.0498	2.836***	1.475***	0.549***
	(0.0938)	(0.186)	(0.546)	(0.323)	(0.512)	(0.173)	(0.406)	(0.115)	(0.165)	(0.126)
HKG Dummy	2.460***	1.898***	0.983*	0.526	-4.384***	1.470***	3.501***	0.945***	0.313	2.392***
	(0.190)	(0.225)	(0.531)	(0.425)	(0.906)	(0.376)	(0.251)	(0.356)	(0.337)	(0.196)
Observations	12,167	12,723	7,698	7,627	7,701	7,026	13,008	7,396	7,533	8,314
Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sector F.E.	-	-	-	-	-	-	-	-	-	-
Pseudo R-squared	0.655	0.566	0.553	0.315	0.471	0.614	0.622	0.760	0.699	0.620

Robust standard errors in parentheses

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table B4: Sector-Specific Estimation (PPML Estimator & GTAP Sector 60-65)

VARIABLES	(1) obs (60)	(2) ros (61)	(3) osg (62)	(4) edu (63)	(5) hht (64)	(6) dwe (65)
ln_DIST	-0.677***	-0.308**	-1.033***	-0.960***	-0.474**	-1.269***
	(0.105)	(0.142)	(0.284)	(0.125)	(0.184)	(0.109)
ln_GDP_host	0.610***	0.480***	0.602***	1.374***	0.762***	1.232***
	(0.0885)	(0.136)	(0.211)	(0.150)	(0.0838)	(0.0641)
ln_GDP_source	0.900***	0.653***	0.105	1.030***	0.669***	1.340***
	(0.0608)	(0.0741)	(0.122)	(0.128)	(0.0878)	(0.0738)
ln_GDPpc_host	2.449***	0.210	-0.658	-1.216***	-0.0534	-1.201***
	(0.483)	(0.372)	(0.663)	(0.383)	(0.210)	(0.256)
ln_GDPpc_source	2.371***	0.000298	1.302**	1.213***	0.992***	0.200
	(0.407)	(0.0936)	(0.513)	(0.409)	(0.174)	(0.174)
Developing Country_host	3.608***	2.189***	-1.926	-2.190*	0.880	-3.827***
	(0.765)	(0.598)	(1.415)	(1.316)	(0.628)	(0.619)
Developing Country_source	3.860***	1.527***	-0.951	0.538	-0.0293	-1.385***
	(0.623)	(0.179)	(0.661)	(0.599)	(0.610)	(0.374)
Common Language	0.364	-0.240	3.345***	0.538	1.641***	0.602***
	(0.368)	(0.381)	(0.601)	(0.796)	(0.320)	(0.183)
Colonial Relationship	-1.000***	-0.203	-0.0682	0.179	0.564*	-0.894*
	(0.380)	(0.427)	(0.937)	(0.406)	(0.304)	(0.459)
LUX Dummy	2.649***	2.348***	-3.320**	3.909***	2.404***	2.979***
	(0.321)	(0.655)	(1.413)	(0.694)	(0.386)	(0.247)
NLD Dummy	0.285	0.308	-2.050	1.059*	1.282***	-0.456**
	(0.403)	(0.342)	(1.248)	(0.578)	(0.496)	(0.199)
HKG Dummy	4.205***	2.996***	-0.946	1.045*	-5.459***	1.968***
	(0.255)	(0.314)	(0.983)	(0.595)	(0.675)	(0.351)
Observations	12,658	13,266	6,286	7,929	7,978	7,728
Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Sector F.E.	-	-	-	-	-	-
Pseudo R-squared	0.595	0.302	0.605	0.577	0.506	0.698

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Annex C

1 Robustness Checks

As discussed in the main results, although our extrapolation is based on the sector-specific coefficients, in almost all sectors they match fairly well with those of the pooled estimation. Therefore for ease of comparison, in our robustness checks the reference/benchmark model specification is the pooled estimation with the extra country dummies which corresponds to column (3) of Table 6. The sector by sector results of the various alternative specifications below are not shown in this paper but are available upon request.

1.1 Additional Explanatory Variables

Our relatively parsimonious gravity equation no doubt gives rise to greater risk of omitted variable bias which is further compounded by the fact that unlike country fixed effects, the adoption of sector and year fixed effects does not absorb country-specific, time-invariant variables. Hence for the first robustness check, we take insights from literature on the determinants of FDI and expand our gravity equation with aditional explanatory variables. First, we have selected three bilateral dummies variables that each take the value of 1 if the country-pair shares a common legal origin, or are both Members of the European Economic Area (EEA), or has ratified bilateral investment treaty (BIT). The information on common legal origins are available from the CEPII database and the list of BITs are maintained by the Electronic Database of Investment Treaties (EDIT). We follow Haftel (2010) and only consider ratified BITs which are legally binding while discarding signed BITs that are yet to be ratified or those that were terminated.

In addition, we include an index from the Design of Trade Agreements (DESTA) Database that measures the comprehensiveness or depth of free trade agreements (FTAs). The depth of a FTA ranges from 0 to 7 which corresponds to the 7 key provisions identified in the DESTA Project[16], and the scoring is simply based on the number of policy areas covered in each individual FTA. At the time of writing, the database is updated to 2019 and covers all FTAs notified to the WTO, therefore it is arguably a better alternative to the simple binary FTA dummy from the CEPII database which is only updated to 2015. In the application, we further adopt a few more adjustments as recommended in Kox and Rojas-Romagosa (2020) which set the depth score for all country-pairs within the EU to 7 while retaining the EEA dummy as a separate regressor, this is to account for the uniquely deep integration within

[16] (1) Full tariff reductions; (2) provisions on common standards; (3) services trade provisions; (4) competition policy provisions; (5) provisions on public procurement; (6) investment; and (7) intellectual property rights

the Single Market which is not reflected by the depth scoring[17].

Finally, we employ the Worldwide Governance Index (WGI) to control for the institutional quality of the **host** country. The index is updated annually for around 200 countries and it provides scoring ranges from -2.5 to 2.5 (best outcome) for each of the six dimensions of governance[18]. Once again taking insights from past literature, we follow van der Marel and Shepherd (2020) and adopt the simply average for two components of the index: the 'Government Effectiveness' and 'Regulatory Quality' which are deemed to be the most relevant to potential foreign investors. Note that the two sub-indices are not included as two separate regressors due to their high correlation.

The results of the extended specification is shown in column (2) of Table C1. Comparing with column (1), there are barely any changes in the overlapping variables aside from a slight drop in the distance coefficient. Meanwhile, the coefficients of these additional variables are poor: the common legal origin dummy and the WGI variable are insignificant while the EEA and BIT dummy are counter-intuitive (negative and statistically significant). In addition, they are also not providing any additional explanatory power given the fact that the before-and-after r-squared are virtually the same. Nevertheless, the results are hardly surprising, beside the potential bias from the omission of multilateral resistance terms, the inclusion of these treaty and institutional variables brings in another source of bias - endogeneity & reverse causality. Indeed, as demonstrated in Kox and Rojas-Romagosa (2020), the effect of treaty variables on FDI are only correctly estimated with the combination of country-pair, host-year and source-year fixed effects, however their estimation strategy is not applicable here since only time-varying bilateral variables can be identified.

As a final note, our effort in incorporating a comprehensive list of explanatory variables is not restricted to the ones discussed above, for instance we have also explored the possibility of including the Ease of Doing Business Score, the OECD FDI Regulatory Restrictiveness Index or the Global Indicators of Regulatory Governance. Ultimately, they are deemed not suitable for reasons such as structure break in the data, a lack of country coverage or a lack of time dimensions.

1.2 Host- & Source-Year Fixed Effects

To Be Completed

[17] The EEA Agreement only receives a depth score of 5, this is not adequate since many other EU-wide treaties that give broad rights and protection to intra-EU investors are not considered

[18] (1) Voice and Accountability; (2) Regulatory Quality; (3) Political Stability and Absence of Violence; (4) Rule of Law; (5) Government Effectiveness; and (6) Control of Corruption

1.3 3-Year averages & SPEs

To Be Completed

Table C1: Robustness Checks - Pooled Estimation (PPML Estimator)

VARIABLES	(1) Benchmark	(2) Extra	(3) Alt. F.E.	(4) 3Y Avg	(5) Incl. SPE
ln_DIST	-0.428***	-0.338***	-0.988***	-0.446***	-0.367***
	(0.0468)	(0.0511)	(0.0589)	(0.0754)	(0.0406)
ln_GDP_host	0.673***	0.693***	-	0.689***	0.692***
	(0.0275)	(0.0282)		(0.0424)	(0.0265)
ln_GDP_source	0.827***	0.856***	-	0.839***	0.780***
	(0.0435)	(0.0412)		(0.0695)	(0.0377)
ln_GDPpc_host	0.611***	0.461***	-	0.637***	0.754***
	(0.0877)	(0.122)		(0.135)	(0.0835)
ln_GDPpc_source	1.211***	1.107***	-	1.211***	1.113***
	(0.109)	(0.108)		(0.175)	(0.0917)
Developing Country_host	0.887***	1.040***	-	0.941***	1.117***
	(0.139)	(0.136)		(0.222)	(0.133)
Developing Country_source	1.255***	1.311***	-	1.332***	0.899***
	(0.239)	(0.209)		(0.397)	(0.218)
Common Language	0.694***	0.653***	0.240**	0.710***	0.774***
	(0.0715)	(0.0734)	(0.102)	(0.109)	(0.0634)
Colonial Relationship	0.0793	0.105	-0.261*	0.125	0.0523
	(0.109)	(0.0964)	(0.150)	(0.167)	(0.0924)
LUX Dummy	2.467***	2.692***	-	2.462***	2.541***
	(0.121)	(0.134)		(0.183)	(0.106)
NLD Dummy	2.074***	2.137***	-	2.080***	2.261***
	(0.0872)	(0.0824)		(0.136)	(0.0746)
HKG Dummy	2.272***	2.590***	-	2.368***	2.025***
	(0.259)	(0.225)		(0.414)	(0.248)
Common Legal Origins	-	-0.00977	-	-	-
		(0.0556)			
EEA	-	-0.710***	-	-	-
		(0.132)			
BIT	-	-0.586***	-	-	-
		(0.0856)			
DESTA Depth	-	0.141***	-	-	-
		(0.0153)			
Worldwide Governance Indicators	-	0.101	-	-	-
		(0.0847)			
Observations	307,621	306,922	290,685	123,170	307,621
Year F.E.	Yes	Yes	No	Yes†	Yes
Sector F.E.	Yes	Yes	No	Yes	Yes
Host-Year F.E.	No	No	Yes	No	No
Source-Year F.E.	No	No	Yes	No	No
Pseudo R-squared	0.688	0.695	0.525	0.686	0.703

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

† The Year Fixed Effects in this instance refers to the periods of 2013-2015 and 2016-2018

2 Out-of-sample Validation

To Be Completed