BACI: International trade database at the product-level. The 1994-2007 version

Gaulier, Guillaume and Zignago, Soledad

CEPII, Banque de France

October 2010

CEPII
CENTRE
D'ÉTUDES PROSPECTIVES
ET D'INFORMATIONS
INTERNATIONALES

No 2010 – 23
October

DOCUMENT DE TRAVAIL

BACI: International Trade Database at the Product-level
The 1994-2007 Version

Guillaume Gaulier
Soledad Zignago

TABLE OF CONTENTS

BACI: INTERNATIONAL TRADE DATABASE AT THE PRODUCT-LEVEL THE 1994-2007 VERSION

NON-TECHNICAL SUMMARY

Empirical international trade analysis increasingly calls for accurate and disaggregated trade statistics. This working paper documents the construction of BACI, a detailed international trade database, which covers more than 200 countries and 5,000 products, between 1994 and 2007. New approaches have been developed to reconcile data reported by over 150 countries to the United Nations Statistics Division, which disseminates them via their COMTRADE database. When both exporting and importing countries report flows, we have two different figures for the same flow. In order to have a single consistent figure of a bilateral flow, we reconcile them using the procedure detailed below.

Firstly, to enable comparisons between import values, which are generally reported CIF (cost, insurance and freight), and export values, reported FOB (free on board), we need to remove transport costs from the reported imports. Detailed information on the product and transport components of CIF rates is not available. The CIF rate is therefore estimated using a gravity-type equation taking into account bilateral distance (in a non-linear manner), dummies for both contiguity and landlockness, year fixed-effects and the world median unit-value for each product category.

Secondly, we need criteria to average the FOB-FOB mirror numbers. We evaluate the reliability of each country's reporting by computing an indicator of the reporting distance among partners (the absolute value of the natural log of the ratio of mirror flows) and decompose it using a (weighted) variance analysis. The relative reliability of country reporting is then cleaned from the effects of its geographical and sectoral specialization. These adjusted qualities of reporting are finally used as weights in the averaging of mirror flows.

The three main advantages of BACI data, in comparison to other similar databases, are its product-level (more than 5,000), its geographical coverage (more than 200 countries) and its unit values, which are more reliable than the raw data, since the reconciliation of mirror figures tend to correct discrepancies. Since our methodology is purely statistical and does not require extensive additional data, the procedure can be applied exhaustively, even to cases in which knowledge on each country and product is very limited. Thus, the main aim of this work is to provide with an international trade database covering the largest number of countries at the highest degree of product-detail, for the longest period.

Users of Comtrade can freely download our BACI database in different available classifications (HS92, HS96, SITC) from http://www.cepii.fr/anglaisgraph/bdd/baci.htm.

ABSTRACT

This paper documents the construction of BACI, our international trade database, which covers more than 200 countries and 5,000 products, between 1994 and 2007. New approaches have been developed to reconcile data reported by almost 150 countries to the United Nations Statistics Division, collated via COMTRADE. When both exporting and importing countries report to Comtrade, we have two different figures for the same flow, so it is useful to reconcile these into a single figure. To do this, firstly, as import values are reported CIF (cost, insurance and freight) while exports are reported FOB (free on board), transport and insurance rates have to be estimated and removed from import values. We regress the observed CIF/FOB ratios for a given flow on gravity variables and a product-specific world median unit value. In a second step we evaluate the reliability of countries reporting. We decompose the absolute value of the ratios of mirror flows using a (weighted) variance analysis. These measures of the reliability of reported data are used as weights in the reconciliation of each bilateral trade flow which is reported twice. Taking advantage of this bilateral information on each flow, we end up with a large coverage of countries and more reliable data, especially in terms of unit-values. BACI is freely available online to users of COMTRADE database, in different product classifications.

JEL Classification: F10, F14, F13, C80.

Keywords: International Trade, Trade Costs, CIF/FOB, Trade data reconciliation.

BACI: BASE POUR L'ANALYSE DU COMMERCE INTERNATIONAL VERSION 1994-2007

RÉSUME NON TECHNIQUE

L'analyse empirique du commerce international réclame de plus en plus de données désagrégées et fiables. Nous présentons ici la méthode de construction de BACI, notre base de données du commerce international qui couvre plus de 5 000 produits et pratiquement tous les pays du monde (plus de 200), annuellement de 1994 à 2007. Pour construire BACI, nous avons développé des méthodes originales d'harmonisation des données-sources de la base COMTRADE des Nations Unies qui compile les déclarations de quelque 150 pays. Il peut y avoir en effet, pour un même flux, deux sources d'information : le pays exportateur et le pays importateur, et les divergences entre ces deux flux miroir peuvent être importantes. Notre procédure d'harmonisation consiste à réconcilier ces deux sources, afin de disposer de données plus exhaustives et plus fiables, notamment en termes de valeurs unitaires (rapport des valeurs aux quantités).

Tout d'abord, les importations, généralement déclarées CAF (y compris les coûts, assurances et fret), sont calculées hors fret pour pouvoir être comparées aux déclarations FAB (franco à bord) des exportateurs. N'ayant pas d'information suffisamment détaillée sur les taux de ces coûts de transport, nous les estimons à partir d'une équation de type gravitationnel retenant comme variables explicatives la distance entre les partenaires (en tenant compte de l'existence d'une éventuelle frontière commune ou d'une situation d'enclavement) et la valeur unitaire médiane mondiale de chaque produit retenue comme indicateur de la plus ou moins grande facilité de son transport.

L'harmonisation consiste ensuite à calculer une moyenne des deux flux miroir, en accordant plus de poids à la déclaration considérée la plus fiable. La qualité de déclaration est estimée en considérant que la distance observée entre les deux déclarations d'un même flux comporte quatre composantes, attribuables respectivement à l'exportateur, à l'importateur, au produit et à l'année considérés ; de cette façon, il est possible d'isoler la qualité propre d'un déclarant, indépendamment de sa spécialisation géographique ou sectorielle.

Notre objectif est de permettre avec BACI des analyses très détaillées du commerce international pour le plus large échantillon de pays et la plus longue période possibles. De fait, BACI est utilisée dans des travaux de recherche pour trois raisons principales : sa dimension produit, son exhaustivité géographique et la fiabilité de ses valeurs unitaires. Notre méthodologie, purement statistique, permet de corriger de nombreuses sources d'erreur. Pour les utilisateurs de COMTRADE, BACI est librement disponible sur le site du CEPII (http ://www.cepii.fr/anglaisgraph/bdd/baci.htm), en différentes nomenclatures produit.

RÉSUMÉ COURT

Ce travail décrit BACI, notre base de données de commerce international au niveau des produits (plus de 5 000) et couvrant pratiquement tous les pays du monde de 1994 à 2007. Nous avons développé pour la construire des méthodes originales d'harmonisation des données de COMTRADE, Nations Unies, qui est notre source des données et qui compile les déclarations de quelque 150 pays. Il peut y avoir deux sources d'information pour un même flux, celle de l'importateur et celle de l'exportateur, et les divergences entre elles peuvent être importantes. Notre procédure d'harmonisation a pour but de proposer des données plus exhaustives et plus fiables, surtout en termes de valeurs unitaires. Tout d'abord, les importations, souvent déclarées incluant les coûts de transport, sont nettoyées de ceux-ci pour pouvoir être comparées aux déclarations des exportateurs. Ensuite, nous estimons une qualité de déclaration pour chaque pays et l'utilisons pour pondérer la moyenne des deux déclarations. BACI est librement disponible en ligne pour les utilisateurs de COMTRADE, en différentes nomenclatures produit.

Classification JEL : F10, F14, F13, C80.

Mots clés : Commerce international, Bases de données, Coûts au commerce, CAF/FAB, Harmonisation des flux de commerce international.

BACI: INTERNATIONAL TRADE DATABASE AT THE PRODUCT-LEVEL THE 1994-2007 VERSION[1]

Guillaume Gaulier*
Soledad Zignago†

1. INTRODUCTION

Empirical studies in international trade increasingly call for accurate and disaggregated trade statistics. However, researchers using trade datasets may be discouraged by missing information or inconsistencies between sources. Drawn on United Nations COMTRADE data, BACI[2] aims to provide comprehensive and disaggregated reconciled values and quantities of international trade for the larger set of countries, products and years.

Countries report yearly their disaggregated bilateral trade flows to the United Nations Statistical Division, which disseminates them via COMTRADE (Commodities Trade Statistics database), the most comprehensive database on world trade. Despite the wealth of this excellent tool, there are still too many missing flows if one wants to have all countries of the world (for the largest period and the most disaggregated product level). Firstly, simply because many countries do not report their detailed external trade to the United Nations, even if the number of reporting countries is rapidly increasing over time. Secondly, some countries report their trade flows in a more aggregated classification. At the international level, the finest product classification is the 6-digit Harmonized System (HS), which has applied progressively from 1989 and distinguishes about 5,000 items.[3] At the beginning of the 2000's, many countries were still reporting in the previous classification, the Standard International Trade Classification (SITC), which covers around 1,200 products in its 4-5 digits level.[4]

[1]The development, communication and dissemination of the BACI project was possible thanks to the excellent research assistance of Dieudonné Sondjo, Adja Sissoko, Rodrigo Paillacar and Julien Martin. We are grateful to our CEPII colleagues for their rich comments and suggestions, in particular Lionel Fontagné, Matthieu Crozet, Yvan Decreux, Isabelle Méjean, Charlotte Emlinger, Antoine Berthou, Alix de Saint-Vaulry, Jacques Gallezot, Houssein Boumelassa. We are indebted to the United Nations Statistical Division for our collaboration, in particular to Matthias Reister and Ronald Jansen for their expert advice. Usual disclaimers apply. Users of BACI are kindly asked to send their questions and comments to baci@cepii.fr.

*Banque de France and CEPII (guillaume.gaulier@banque-france.fr).
†Banque de France (soledad.zignago@banque-france.fr).

[2]French acronym of "Base pour l'Analyse du Commerce International": Database for International Trade Analysis.

[3]Underneath this level, there is no common international classification of commodities. In other words, national or regional customs having adopted the Harmonized System report their trade to the UN in their own tariff-lines classification, which is internationally the same only until the 6-digit level.

[4]In 2008 however, only one country, Palestine, still reported in the SITC classification.

Becauses countries report both their imports and their exports, we can have: i) two figures for the same flow reported by the importer *j* and the exporter *i*, if *i* and *j* are both reporting countries in the 6-digit HS; or ii) only one figure for a flow reported only by the importer (or only by the exporter) and a missing value in the export (import) side; or iii) missing values on both sides. BACI takes advantage of the double information on each trade flow to fill out the matrix of bilateral world trade and provide a unique "reconciled" value (or quantity) for each flow reported at least by one of the partners. Therefore, the sole missing values in BACI are those concerning trade between two non-reporting countries (iii).[5]

We have developed an original procedure to reconcile flows reported by exporters and importers. In general, import values are reported CIF (cost, insurance and freight) and exports are reported FOB (free on board). To allow the comparison between mirror data, CIF rates have to be estimated and removed from imports values. We use a gravity-type equation to estimate them. We evaluate then the reliability of country reports and use it as weights in the average of mirror values and quantities.

This working paper documents these reconciliation methodology, which can be applied to different trade datasets. The most exhaustive version of BACI provides values and quantities at the 6-digit level of the first HS classification, launched in 1988. Since this first version (HS0), the HS was importantly revised in 1996 (HS1), 2002 (HS2), and 2007 (HS3).[6] Other versions of BACI, built with in the same methodology but based on the 1996 HS classification (HS1) and on the SITC, are also available (or will be soon). Since July 2007, BACI has been available for COMTRADE users in our webpage: www.cepii.fr/anglaisgraph/bdd/baci.htm.

A methodological change in the UN Statistical Division treatment of quantities has to be underlined. Aiming to reduce missing quantities, this new treatment of quantities applied from year 2005 onwards tends to reduce variance in the unit-values and can introduce a serious break in their evolution before and after 2005. Users are invited to have this break in mind, in particular when they are interested in the long-term evolution of unit-values of some developing countries. However, this correction seems to impact mostly minor exporters. This caveat being noted, we provide values and quantities for the entire period 1994-2007.

BACI is widely used to analyse trade patterns at the product-level, countries specialization, competitiveness, trade policy, exchange-rate pass-through, etc. Since it is the unique database providing consistent unit-values at the world and product level, BACI is particularly convenient to analyse international trade prices. Its exhaustive coverage is useful also to analyse international trade of non reporting countries such as African countries. BACI is also an input to other CEPII databases like TradeProd, TradePrices and MacMap.

[5]We provide the matrix of reporting countries in our webpage (http://www.cepii.fr/anglaisgraph/bdd/baci.htm, files named zeros since they allow to distinguish between zero or missing flows).

[6]We have applied our reconciliation procedure firstly to the HS0 in order to have the longest time period, between 1995 and 2004. This is why results presented here use this first BACI version, instead of the more recent one covering the period 1994-2007.

The remaining of the paper is as follows. The next section presents the methodology developed to reconcile mirror flows: the data source, the evaluation of CIF rates, the assessment of the quality of country reports. Section 3 comments the resulting datasets: different versions of BACI, a brief comparison with other trade databases and some main applications of BACI in the literature. Section 4 concludes and announces future developments.

2. THE METHODOLOGY OF RECONCILIATION OF BILATERAL TRADE FLOWS

2.1. Data used for BACI: UN COMTRADE

The methodology described in this section was firstly applied to the Harmonized Commodity Description and Coding System (HS in the following) since it is the most detailed classification (over 5,000 products) at the international level. The HS is at the heart of the whole process of harmonisation of international economic classifications being jointly conducted by the United Nations Statistics Division and Eurostat. Its items and sub-items are the fundamental terms on which industrial goods are identified in product classifications.[7] The World Customs Organization revises the HS every few years. Since its first version (HS0 in the following), the HS has been importantly revised in 1996 (HS1), 2002 (HS2), and 2007 (HS3). The HS is organized in four hierarchical levels:

- Level 1: sections coded by Roman numerals (I to XXI);
- Level 2: chapters identified by 2-digit numerical codes;
- Level 3: headings identified by 4-digit numerical codes;
- Level 4: sub-headings identified by 6-digit numerical codes (we name them *products*).

Countries report to the United Nations their international trade statistics detailed by commodity and partner country. The UN Statistics Division disseminates the annual data reported via COMTRADE (Commodities Trade Statistics database), which provides very detailed trade data, accounting for more than 95% of the world trade.[8] COMTRADE provides data on imports,

[7] According to Ramon-Eurostat (ec.europa.eu/eurostat/ramon/), linked classification(s) are:

1) Central Product Classification (CPC);

2) International Standard Industrial Classification of All Economic Activities, Third Revision (ISIC Rev.3);

3) Standard International Trade Classification, Third Revision (SITC Rev.3);

4) Statistical Classification of Products by Activity in the European Economic Community (CPA);

5) Statistical Classification of Economic Activities in the European Community (NACE Rev.1);

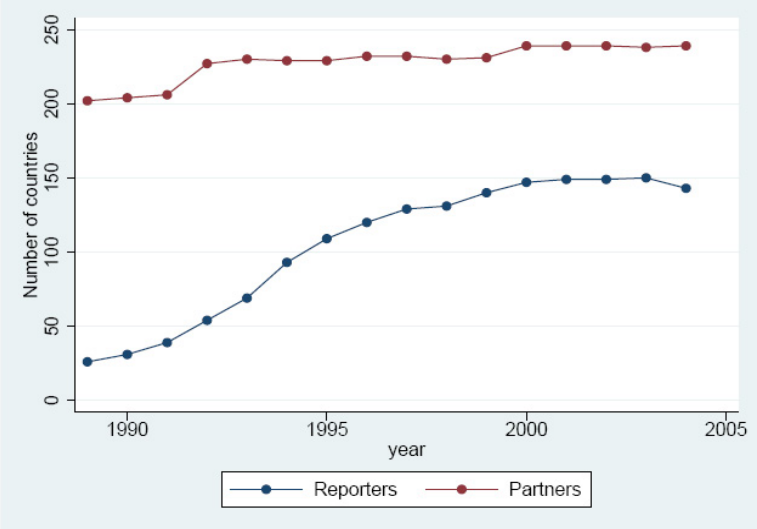
6) Combined Nomenclature (CN) : Full agreement at six-digit level.

Free downloads of classifications and tables of correspondence are also avaiable in the UN Classifications website (http://unstats.un.org/unsd/cr/registry/regdnld.asp?Lg=1).

[8] The International Trade Center (a joint agency of the World Trade Organization and the United Nations) provide also these reported trade data, as well as very detailed tarif data and market access indicators. One im-

exports, re-imports and re-exports (in values and quantities) in different international product classifications but the most disaggregated is the Harmonized System 6-digit level. Since 1989, an increasing number of countries (See Figure 1) has reported in the HS current classification (the HS2 version in most cases in recent years).[9] Data do not include flows below 1,000 dollars.

Figure 1 – Number of reporting and partner countries in COMTRADE.



Source: Comtrade.

The COMTRADE database is used as the single source of information to build BACI. We have applied our reconciliation procedure firstly to the HS0 in order to have the largest time period version of BACI at the HS 6-digit level: 1994-2007 (the number of countries reporting in the HS classification is satisfactory since 1994). We provide also BACI datasets in the HS1 classification covering the period 1998-2007 and in the SITC classification (around 1200 items) covering the period 1980-2007.

For a given trade flow, COMTRADE provides two sets of series if both commercial partners report their data to the UN. In general exports are reported Free On Board (FOB), while imports are reported inclusive of the Cost for Insurance and Freight (CIF).[10] In principle exports from

portant value-added of ITC international trade data in comparison to COMTRADE is that their *Trade Map* database provide monthly and quaterly series collected from national custom offices or regional organisations (http://www.intracen.org/marketanalysis/TradeMap.aspx).

[9] According to the UNSD (2004), since 2001 there has been 102 countries that are *Contracting Parties*, *i.e.* they recognize the Harmonized System as a legal instrument. Another 78 countries are not *Contracting Parties*, but use the HS System. COMTRADE also provides with longer series, starting in 1967, for more aggregated product decompositions. For further details on COMTRADE see http://unstats.un.org/unsd/comtrade/.

[10] It should be mentioned that there are many other regimes of delivery. UNSD (2004) identifies 13 of them, according to the costs actually involved in the reported value of the country.

country *i* to country *j* should be identical to imports from country *i* to country *j*, for any given product and year, except for the CIF additional cost. In practice this may be untrue for several reasons. Firstly, the identification of the actual trading partner may be difficult. Generally customs officials pay more attention to the actual origin of an imported product because this determines the level of tariff that will be applied to it. They may be less careful when it comes to the actual destination of exports. Secondly, the reported values detailed by commodities do not necessarily sum up to the total trade value for a given country. Due to confidentiality for instance, countries may not report some of its detailed trade. However, this trade will be included at the higher commodity level and in the total trade value (and sometimes via the use of a specific item of the product classification). Many other sources of misreport can be imagined: product misclassification, different reporting year if goods are shipped at the end of the year... We will see that the difference between the two reported figures may be significant for some flows.

BACI is a useful tool for international trade analysis at high degrees of disaggregation, in complement to COMTRADE. Firstly, it provides in a coherent database for a single harmonized value, allowing for international comparison. The use of mirror flows allow us to largely complete missing reportings. Secondly, it provides comparable quantities and thus unit values. Whereas values are reported in thousands of US dollars, quantities can be registered in different units of measure (meters, square meters, etc.). Since most of exchanged quantities are reported in tons, we convert the remaining quantities by estimating implicit rates of conversion of other units into ton units. Quantities are then harmonized in the same way that values using mirror data, ending up with a very complete database (more than 200 countries in BACI whereas in COMTRADE there are 130 on average in the period 1994-2007).[11]

2.2. Conversion in tons

Even though most quantities are reported in tons, there is 15% reported in other quantity units (units, meters, watt, etc). International trade analysis needs reliable data on unit-values (values divided by quantities) of products exchanged to investigate prices, or quality issues. For each product concerned, we estimate the rates of conversion into tons of the different units in which it is reported, using mirror flows reported in tons by a country and in another unit by the other trading partner. Quantities reported in unknown units or in Kwh are dropped for simplicity.

[11] Obviously BACI displays at an aggregated level the same trends than other trade databases, in particular Comtrade. BACI was mostly designed for high disaggregated studies, as a tool to describe medium term changes in the international division of labor (variety of exported products, vertical differentiation, technological content, stage of production). Since the main objective of BACI is to provide with very detailed data, BACI does not take into account some country aggregations provided by COMTRADE. BACI provides trade data between individual countries (or custom unions reporting as a single country) which are present in the entire period 1994-2007. Thus, flows within some groups of countries (Belgium-Luxembourg for instance) are dropped in orther to have a consistent geography over the period. Re-exportations for Hong Kong and USA are also dropped since there is no way to know the final destination of the flow.

These implicit rates of conversion are then applied to quantities reported in heterogeneous units. However, the conversion is only performed if a minimum of 10 mirror flows have been used in its computation, and if the standard deviation is inferior to 2.5. About 8,5% of final quantities in BACI have been converted using this method.

2.3. The CIF/FOB ratios estimation

The present subsection details the estimation of CIF/FOB ratios. Generally importers report CIF values while exporters report FOB values. Because of the scarcity of the transport cost data at a suitable level of detail, we choose a *fobization* technique of CIF import values. We estimate the CIF rates, which will then be removed from import reports to allow for the comparison with export reports.

2.3.1. Empirics on the evaluation of transport costs

Direct transport costs are rarely available at the product-level. Hummels and Skiba's (2004) paper is one of the most complete review of this costs, with only six importer countries: Argentina, Brazil, Chile, Paraguay, Uruguay and the USA provide very precise bilateral freigth costs. Concerning the latter country, the NBER via Robert Feenstra's webpage provides time series since 1972. Australia and New Zealand give also detailed information (see Hummels and Lugovsky, 2006, for instance). It seems difficult to infer from such limited coverage all the cross-country variability of real freigth costs in all possible country-pairs. A flourishing literature has then discussed the way to estimate these costs.

A first class of empirical papers rely on directly measured trade barriers in terms of detailed freight data for a limited number of countries. For instance Hummels (2001), exploiting imports data from U.S. Census Bureau, shows the wide dispersion in freight rates accross commodities and countries in 1994. The all-commodities trade weighted average transport cost from national customs data ranges from 3% of FOB price for the U.S. to 13.3% for Paraguay.[12] Alternatively, Limão & Venables (2001) highlight the dependence of trade costs on infrastructure. [13]

[12]Hummels (2001) starts from a multi-sector model of trade and uses a more sensible trade costs function than commonly done in the literature. Such a technique permits a complete featuring of the trade costs: elasticities of substitution between goods are identified and meaningful interpretation of common proxy variables in terms of their *ad-valorem* trade barrier equivalent is provided. According to Hummels (2001), for a given elasticity of substitution, production migrates to minimize costs such that nearby country produce complementary sets of goods (this explanation is consistent with the large estimates derived from the border effect literature). Unfortunately, Hummels (2001)'s promising approach requires the use of explicit data on freight and tariff rates that are unavailable for most of countries in the world at a high degree of disaggregation.

[13]Using shipping company quotes for the cost of transporting a standard container from Baltimore to selected destinations, they found that a deterioration of infrastructure from the median to the 75th percentile of destination raises transport costs by 12%. The inconvenient with these approach is that they are generally characterized by a wide variation over countries, and charges are affected by the particular routes, frequencies and opportunities for back-hauling and for exploiting monopoly power that are present.

In the absence of direct measures, a second class of papers turns to alternative techniques to derive estimates of trade costs, indirect measures of freight costs drawing on ratios of mirror trade reports (CIF/FOB ratios). In principle, comparing the valuation of the same flow reported by both the importer (in CIF) and exporter (in FOB) would yield a difference equal to freight costs. However, in practice, we have to deal with important measurement problems: at the 6-digit level, the discrepancies displayed by the importer and the exporter values reported exceeds 100% for more than half of the observations in the COMTRADE database. Statistical offices in exporting and importing countries may value commodities differently for many reasons ranging from the exchange rate variation to differences between partners in the way they track shipments. Note that the discrepancies need not be large to have a sizable impact on the measured CIF – FOB ratios. As highlighted by Hummels & Lugovskyy (2006) if we consider a CIF –FOB ratio of 1.06 (which implies a transportation costs of 6% *ad-valorem*), an increase of the importer's CIF value of trade by 1.5% combined with a decrease of the exporter's FOB value by 1.5% yields a CIF – FOB ratio of 1.09 which changes implied transport costs by 50%.

Hence, the huge discrepancies observed between mirror flows cannot be used directly as measures of freight costs. Yeats (1978) provides an evaluation of the shipping costs data collected from US imports in 1974 to the quality of matched partner data by comparing CIF/FOB ratios computed from the COMTRADE database. He decomposes the observed variation in matched partner CIF/FOB ratios into two parts: one corresponding to the shipping costs and the remaining being unexplained (noise). Even though Yeats finds that for some exporters and commodities very little error is reported, he underlines that matched partner CIF/FOB data contain a non-negligible part of noise. More recently, using IMF data, Hummels & Lugovskyy (2006) state that CIF/FOB ratios are badly error-ridden in levels, and contain no useful information for time-series and cross-commodities variation. Nevertheless, they also conclude that an indirect use of the CIF/FOB ratios can be made. Data do contain errors but are still usable. Hummels & Lugovskyy (2006) state that IMF CIF/FOB ratios only seem to reveal meaningful cross-exporter variation that might be usefully exploited by researchers. In BACI, we exploit this fact in postulating that even if matched partner CIF/FOB data are systematically wrong in levels, they might be strongly correlated with direct measures of shipping costs such that matched partner technique may provide an interesting source of data. As Hummels and Lugovskyy (2006) show, IMF freight data are positively correlated with distance between partner countries and weight of commodities shipped between them. Such findings provide insights to make use of the matched partner CIF/FOB data.

2.3.2. A gravity-type equation to evaluate CIF rates

We explain the implicit CIF/FOB ratios by a set of gravity-type explanatory variables. The predicted mirrors flows ($\widehat{CIF}_{ij}^{kt}$) are used then as *estimates* of CIF rates of j imports from the exporter i of product k in year t.

Of course, the distance between partner countries play an important role in the transportation

costs. But it remains to define the shape of the relation which ties the distance with the CIF rate. Probably, on short distances the CIF rate has a different evolution than it could have on longer distances. We consider thus a non-linear relationship between bilateral distance and CIF rate by introducing both the distance and the squared distance as determinants of CIF rates. Dummies for landlockness and contiguity are also included. Those variables control respectively for the fact that CIF rate should decrease if the exporter and the importer countries are contiguous and increase if one of them is landlocked. This geographic variables are taken from CEPII's distances database (Mayer and Zignago, 2011).[14]

Besides geographic characteristics, our equation includes as explanatory variables the world median unit value for each (6-digit) product k (value/quantity or UV_k) which aims to capture the transportability of the commodities. In other words, it controls for the higher costs of trading heavy commodities.

We introduce also time dummies t in order to capture any potential time evolution of the CIF rate.[15] Thus, the gravity equation, estimated by OLS on pooled data over the period 1989-2007, is basically as follows:

$$\ln(CIFrate_{ij}^{kt}) = \alpha + \beta \ln Dist_{ij} + \chi \ln Dist_{ij}^2 + \delta Contiguity_{ij} + \phi Landlocked_i + \gamma Landlocked_j + \eta \ln UV^k + \sum_{l=1989}^{2004} \varphi_l t_l + \varepsilon_{ij}^{kt} \quad (1)$$

We consider four different specifications of this equation. The dependent variable is alternatively the CIF/FOB ratios in values ($Vm_{ij}^{kt}/Vx_{ij}^{kt}$) or in unit-values ($UVm_{ij}^{kt}/UVx_{ij}^{kt}$), where V and UV denotes respectively values and unit values reported by the exporter (Vx), or by the importer (Vm). Since errors on values and quantities are correlated for a given reporter-product pair, we prefer the estimation of the CIF rate using the unit values ratios.[16] The fact that the transportation costs of commodities depend both on quantities and values can also support the preference for ratios in terms of unit-values.

[14] Available at http://www.cepii.fr/anglaisgraph/bdd/distances.htm, this database provides geodesic distances between all countries in the world, which take into account the geographical dispersion of the economic activity within each country, by considering in the computation the latitude and longitude of its main cities and weighting them by their population. We use weighted distances when available (148 countries out of 225 partner countries).

[15] A proper specification of the gravity equation could also include country fixed effects. However, country-specific dimensions are considered in the second stage of our reconciliation, where we establish a ranking of quality of country reporting based on the gaps between partners reports.

[16] For instance, an overvaluation of a trade flow implies a higher value reported and can also imply a higher quantity (for example if the exporter reports a minor total than the importer of annual bilateral flows because it ignores some type of individual firm-level reportings). By dividing this value by an also overvaluated quantities will reduce the overvaluation in the estimation of the CIF rate.

These two kinds of CIF/FOB ratios can be weighted, or not, by the inverse of the gap between reported mirror quantities ($Min\left(Qx_{ij}, Qm_{ji}\right)/Max\left(Qx_{ij}, Qm_{ji}\right)$, where Qdenotes quantities reported by the exporter, Qx, or by the importer Qm). The weighting confers a higher importance to trade flows similarly reported by partners: differences between reported import and export values are then more likely to correspond to freight costs.

Table 1 – Results of the estimation of freight costs (1989-2004)

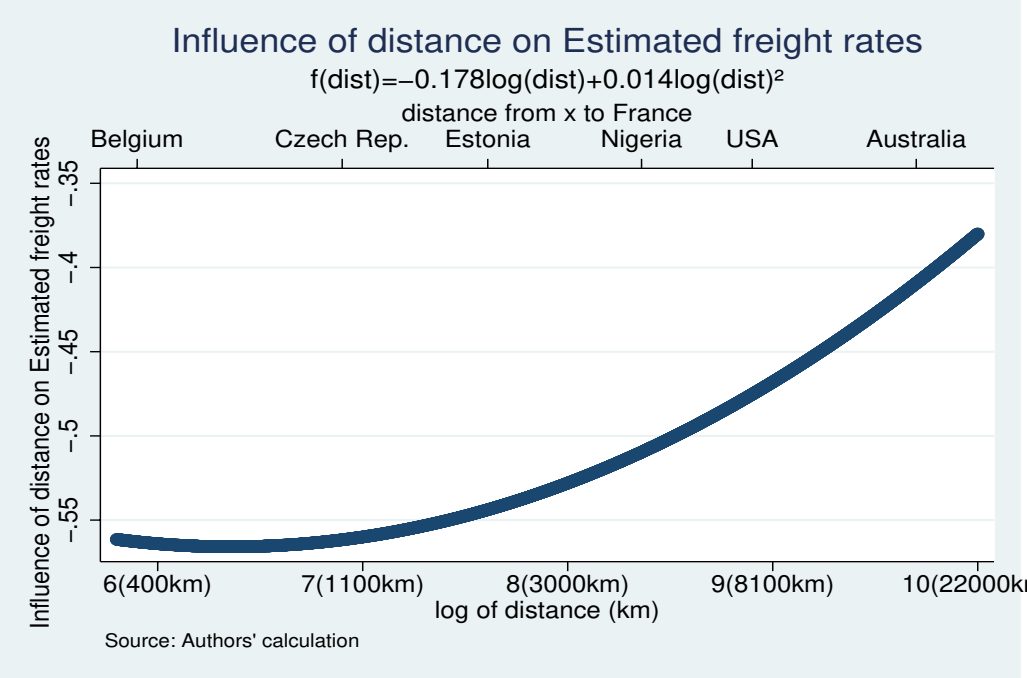
Dep. Variable	$\ln(UVm_{ij}^{kt}/UVx_{ij}^{kt})$		$\ln(Vm_{ij}^{kt}/Vx_{ij}^{kt})$	
	I	II	III	IV
	(no weighting)	(weighting)	(no weighting)	(weighting)
$Intercept$	0.534^a	0.32^a	0.442^a	0.3^a
	(0.016)	(0.011)	(0.01)	(0.007)
$\ln Dist_{ij}$	-0.178^a	-0.11^a	-0.122^a	-0.086^a
	(0.004)	(0.003)	(0.002)	(0.002)
$\ln Dist_{ij}^2$	0.014^a	0.01^a	0.009^a	0.007^a
	(0.000)	(0.000)	(0.000)	(0.000)
$\ln UV^k$	-0.032^a	-0.032^a	-0.042^a	-0.038^a
	(0.000)	(0.000)	(0.000)	(0.000)
$Contiguity_{ij}$	-0.066^a	-0.044^a	-0.025^a	-0.024^a
	(0.001)	(0.001)	0.001)	(0.000)
$Landlocked_j$	0.066^a	0.049^a	0.024^a	0.02^a
	(0.001)	(0.001)	(0.001)	(0.000)
$Landlocked_i$	-0.021^a	-0.009^a	0.012^a	0.01^a
	(0.001)	(0.001)	(0.001)	(0.000)
Time FE	Yes	Yes	Yes	Yes
N. obs.	8,856,312	9,053,610	8,897,367	8,936,618
R^2	0.008	0.012	0.014	0.02
Outlier values	482,840	285,542	441,785	402,534
Mean CIF	0.03	0.033	0.027	0.034

Note: In the first two columns the dependent variable is the ratio of mirror unit-values (UVm_{ij}^{kt} and UVx_{ij}^{kt} are respectively importer and exporter reported unit-values for the same flow from i to j). It is the ratio of mirror values in the two last columns. Standard errors are in parentheses. ^{a}, ^{b} and ^{c} denote a significant coefficient at 1%, 5% and 10% respectively. Models II and IV are weighted by $Min\left(Qx_{ij}, Qm_{ji}\right)/Max\left(Qx_{ij}, Qm_{ji}\right)$.

Table 1 presents the estimation results of the gravity equation over the period 1989-2004. Note that, except for exporter landlockness, there is no reversion of signs in the coefficients when the dependent variable varies, and the magnitudes are similar, resulting in similar estimations for the mean CIF rate, ranging between 2.7% and 3.4%. All coefficients are significant at 1%. The estimated impacts of time dummies show a uniform evolution with a positive sign each year (2004 is the year of reference). The estimated coefficients imply that CIF rates increase with distance and decrease with the world median unit value of the product k. Figure 2 gives an

example of the influence of distance on the estimated CIF rates. The coeficient for contiguity supports the idea that the CIF rates are lower when the two partners share the same land border. In contrast, the sign of the coefficient on the variable capturing landlockness depends on the model under examination. Theoretically, the sign should be positive in order to corroborate the fact that the access to a landlocked country is more costly. This is confirmed in all models for the importer country, but in the case of landlocked exporters, the results using unit-values ratios as dependent variable are slighlty negative.

Figure 2 – Example of the influence of distance in the estimated CIF costs: Distance to France using coefficients of the first column of Table 1



The database contains more than 9.3 millions of observations. However, in order to obtain consistent and robust parameter estimates, we compute the distance of Cook (1977) to remove atypical and influential observations. Weighted regressions, specially model II, suffers less from this procedure, allowing for an estimation with more observations. Accordingly to this model II, the average estimated CIF rate is 3.3%. This value is weaker than what is generally assumed. For instance, according to Anderson & Eric Wincoop (2004) a world possible mean would be 8%. Nevertheless, it is consistent with the result of Hummels (2001), once the differences across specifications are taken into account. Hummels uses shipping cost data (for USA, New Zealand and some South-American countries) and the coefficient obtained in that case is the *explicit* CIF rate (based on observed data on freights) denoted α. We take from Anderson & van Wincoop (2004) the following equation linking the CIF rates in both specifications, the Hummels one and ours : $\beta = \alpha\, CIF/(1 + CIF)$, where β is the *implicit* CIF rate (based in observed CIF/FOB ratios) in our specification. Using a $CIF = 8\%$ (taken from Anderson

& van Wincoop, 2004), the α reported by Hummels (0.27) is consistent with our result (i.e; $\beta = 0.27 \times 0.08/1.08 = 0.02$).

Now, we use our estimator of CIF rates to convert CIF values into FOB values. We take into account the specific bilateral and product characteristics of each flow ant remove the resulting value of freight from the import reported values. To ensure that this procedure will truly improve trade data (and never worsen it) we apply it under some conditions.[17] About 17 millions of trade flows are actually treated by this procedure, representing 21% of the total number of flows (or about 40% of import flows).

2.4. Evaluation of the accuracy of reports

In this subsection we describe our evaluation of the quality of country reports, which serves as weight in the averaging procedure between reported mirror flows, now cleaned from CIF rates. This reconciliation concerns 35% of observations (those for which both mirror flows exist). This second step of our reconciliation methodology consists in computing weighted averages of mirror figures on the basis of an estimated indicator of the accuracy of reports of each country. This evaluation is obtained using a (weighted) variance analysis via a decomposition of the absolute value of the mirror flows ratios (in log).[18]

The *true* trade value V is unobservable, since the value reported contains an error E_i, which is specific to reporting counry i. We assume a multiplicative and log-normal form for this error:

$$V_i = V * E_i \qquad with \qquad \epsilon_i = \ln E_i \sim N(0, \sigma_i^2) \tag{2}$$

where σ_i^2 is the variance of $\ln E_i$.[19] Our objective is to find the weights w to use when averaging figures reported by both the exporter V_i and the importer V_j. The reconciled value (RV) is defined as $RV = w * V_i + (1 - w) * V_j$, which in terms of true flows gives: $RV/V = w * E_i + (1 - w) * E_j$. The minimization of the variance of RV/V gives the optimal weighting:

$$w = \frac{e^{\sigma_j^2}(e^{\sigma_j^2} - 1)}{e^{\sigma_i^2}(e^{\sigma_i^2} - 1) + (e^{\sigma_i^2} - 1)} \tag{3}$$

[17]For instance, there are some criteria to cope with particular cases: (1) the procedure is not implemented to countries which do not report their imports in CIF (such Algeria, Georgia, South Africa and other SACU countries); (2) in countries that report their imports in FAS (such as Canada), we do implement the correction but only if it minimizes the gap between the mirror flows; and (3) a negative FOB-import value is set to zero.

[18]Ten Cate (2007) also uses mirror data to estimate the accuracy of the reporters and to compute optimal combinations of mirror data.

[19]If m and σ^2 are the mean and the variance of a normal distribution, then the mean and the variance of the log-normal distribution are: $m + e^{\sigma_i^2/2}$ and $e^{2m+\sigma_i^2}(e^{\sigma_i^2} - 1))$. Since we suppose a mean equals to zero, $E_i \sim logN(e^{\sigma_i^2/2}, e^{\sigma_i^2}(e^{\sigma_i^2} - 1))$.

We need then an expression of variances σ_i^2 and σ_j^2. We define the absolute value of the log of the mirror figures ratios as the "reporting distance": $RD_{ij} = \left|\ln(\frac{V_i}{V_j})\right| = |\ln E_i - \ln E_j|$. Given our assumptions on the error term, $(\ln E_i - \ln E_j) \sim N(0, \sqrt{\sigma_i^2 + \sigma_j^2})$ the mean of the reporting distance for a couple of countries i and j is:

$$\overline{RD_{ij}} = \sqrt{\frac{2}{\pi}}\sqrt{\sigma_i^2 + \sigma_j^2} \tag{4}$$

We assume that reporting distances can be decomposed into four terms: a term due to the exporting country i, a term due to the importing country j, a term due to the year t, and a term due to the product k. The two last types of fixed effects allow us to isolate the source of discrepancies which are independent of the quality of trading partners reportings. Therefore, the (relative) quality of declaration of a country i would be *cleaned* from the effects of its specialization (the share of poor/good reporters in its trade partners and the share of products with frequent reporting errors because of lack of homogeneity in the 6-digit position for instance).[20] The estimation of i and j fixed effects allows us to compute the marginal (weighted) mean of RD for each exporter and importer. These least square means are adjusted for the influence from the other factors and are noted $LS\_\overline{RD_i}$ and $LS\_\overline{RD_j}$. They are obtained running the following OLS estimation, as well as the standard errors, denoted by $stderr_i$.

$$RD_{ij}^{kt} = \alpha_i + \beta_j + \lambda_t + \gamma_k + \varepsilon_{ij}^{kt} \quad with \quad \sum_i \alpha_i = \sum_j \beta_j = \sum_t \lambda_t = \sum_k \gamma_k = 0 \tag{5}$$

Given our assumptions on discrepancies, the weighted mean of errors in reports of an exporter i, for instance, can be proxied as $\overline{RD_i} = 2\sigma_i/\pi + K_i$, where K_i is an i-specific constant.[21] Replacing $\overline{RD}$ by the estimated least square means of country-specific discrepancies we obtain an expression of $\widehat{\sigma_i} = \frac{\pi}{2}(LS\_\overline{RD_i} - K_i)$ (and similarly for $\widehat{\sigma_i}$). The constant is set in order for the best to display the smallest value of σ: $K_i = \min_i LS\_\overline{RD_i} + 2stderr_i$.[22]

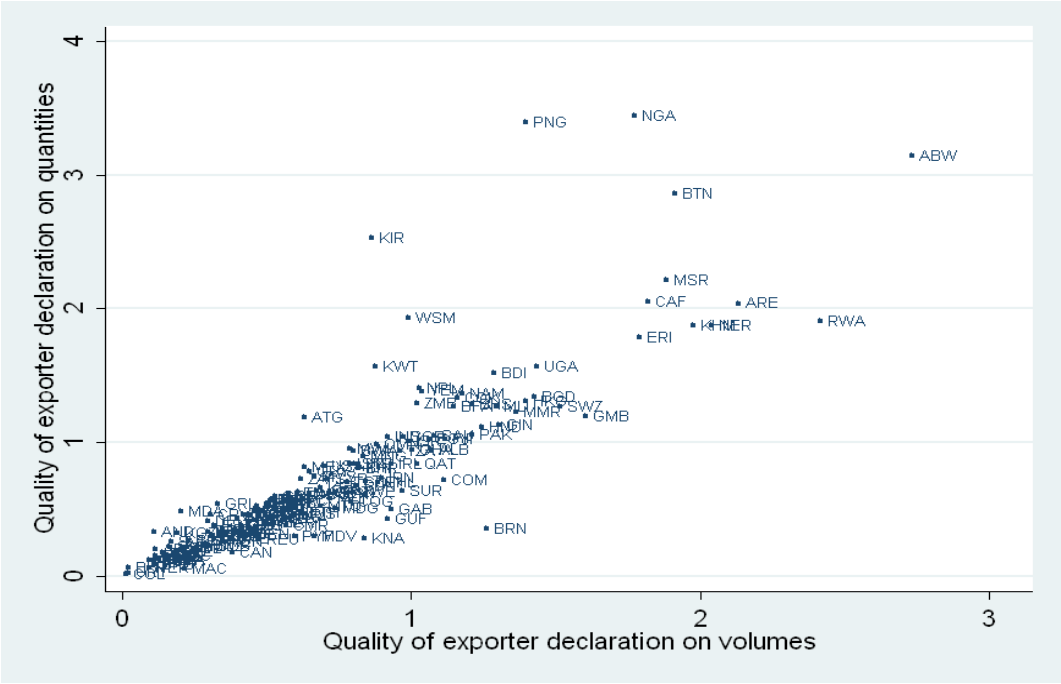
A ranking of the qualities in data reportings can be obtained by ranging in ascending order the estimated qualities. Figures 3 and 4 exhibit the quality indicators of exporter reports for values (horizontal axis) and quantities (vertical axis). Similar results are obtained for the countries as

[20]The product dimension is taken into account through the within transformation, to avoid employing the 5,000 product fixed effects.

[21]Each observation is weighted with the natural log of the sum of the two reports, denoted s. Equation 4 implies that the weighted mean error in the reports of the given exporter i is: $\overline{RD_{ij}} = \sqrt{\frac{2}{\pi}}\sum_j s_j\sqrt{\sigma_i^2 + \sigma_j^2} \approx \sqrt{\frac{2}{\pi}}\sum_j s_j(\sigma_i + \sigma_j)\sqrt{\frac{2}{\pi}} = \frac{2}{\pi}\sigma_i + K_i$.

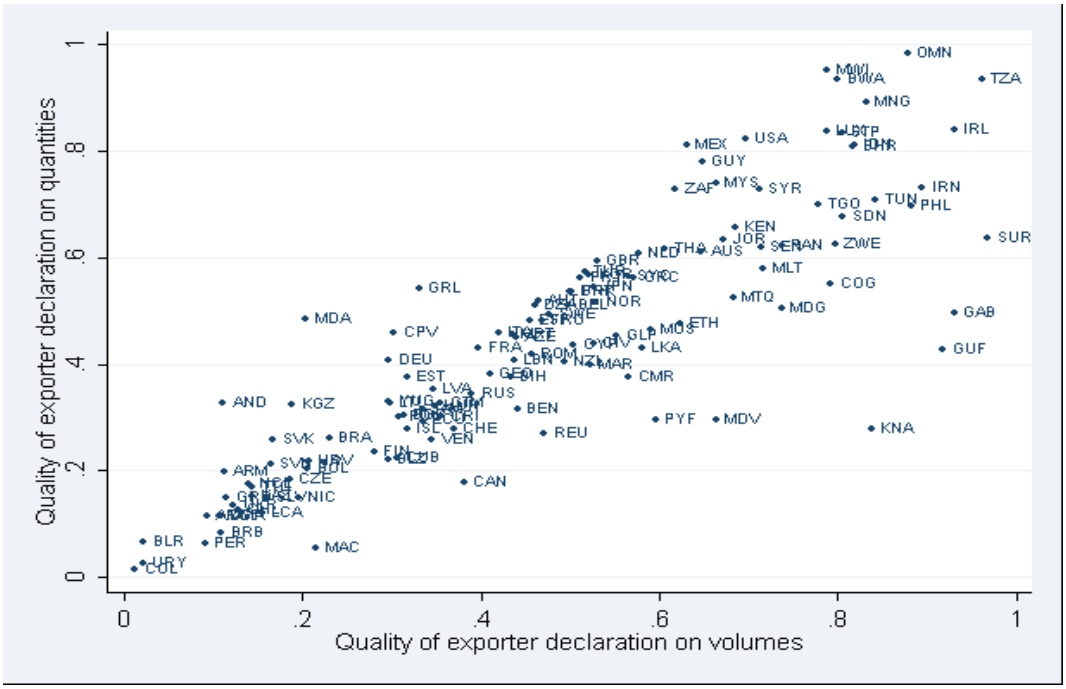
[22]This ad-hoc transformation can be ignored but gives more differentiated weights than the direct use of least square means and the taking into account of the standard deviation of the i fixed effect coefficient $stderr_i$ allow to capture also the precision of the estimation. Thus, $\widehat{\sigma_i} = \frac{\pi}{2}(LS\_\overline{RD_i} - \min(LS\_\overline{RD_i}) - 2stderr_i)$.

Figure 3 – Quality of exporter reports on quantities and values, all reporters



Source: Authors' calculations.

Figure 4 – Quality of exporter reports on quantities and values for better reporters



Source: Authors' calculations.

reporting importers, although the rankings are not systematically the same. We see that both measures are correlated, as expected. Looking at the best reporters (Fig 4), we find most of industrialized countries, but also some emerging and developing countries, in particular several from Latin America and Eastern Europe.

The last step relies on the averaging of two figures to be reconciled. This reconciliation will affect values as well as quantities when both mirror flows exist.[23]

The advantage, but also the limit, of our reconciliation method is its application to very exhaustive data covering, to cases in which the expertise on each country and product is impossible. Our procedure is purely statistic and it does not require as input other data than raw trade statistics, allowing for an improvement of the quantity and quality of the trade data with an arguably reasonable ranking of countries in terms their data quality.

3. COMMENTS ON THE RESULTING DATASETS

3.1. Different available versions of BACI

We have applied our reconciliation procedure to the HS0, HS1 and SITC, covering respectively the periods 1994-2007[24], 1998-2007 and 1980-2006. We end up with very disagregated databases: over 5,000 products for the HS0 and HS1 and 1,200 for the SITC data.

The first version of BACI, using the first version of the HS (named HS0 or 1992), has the longest time coverage. But the version based on HS1 data (from 1996 revision) could be prefered if one wants to match BACI with tariff data, which is generally provided in the current version of the HS. For instance, the unit values from HS1 version of BACI are used in the MAcMap-HS6digit database (Bouët et al., 2008) to compute specific tariffs.

Before the implementation of the HS, countries reported their international trade in the SITC classification. We have also run our reconciliation procedure on SITC with the purpose to update the *TradeProd* database, which provides trade and production industrial data in a consistent classification (ISIC) for a long time period.

Finally, the country coverage is largely improved due to our use of mirror flows: we provide trade data for more than 200 countries, at the most detailed international level. Users of COMTRADE can register themselves in our webpage (www.cepii.fr/anglaisgraph/bdd/baci.htm) and freely download our datasets, available by year in the csv format, in thousands of current dollars. They will find also complementary information such as country and product codes as well as useful datasets needed to distinguish zeros from missing flows.[25] BACI users are kindly

[23]When only one of the reports is missing, the non missing declaration is used (cleaned from CIF costs). See the appendix for more details about the special cases of reconciliation, where only the exporter or the importer declaration is employed, despite the existence of both flows.

[24]With a potential break in 2005 for some reporters, see Section 1.

[25]BACI datasets, as COMTRADE, do not report zero values because the size of datasets would exponentially

asked to contact baci@cepii.fr for any question or to let us know the references of their work using BACI.

3.2. Comparison between BACI and other databases

In this subsection, we present a brief comparison between BACI and some other similar trade databases (annual data). In particular, we consider the NBER database from Feenstra et al. (2005), the CHELEM database from CEPII, the Global Trade Analysis Project (GTAP) and COMTRADE itself. A general comparison is presented in Table 2. Overall, the highest disaggregation level is reached with the BACI and COMTRADE datasets.

Table 2 – Comparison between International Trade Databases

	BACI0	COMTRADE	NBER-UN	CHELEM	GTAP
Period	1995-2004	1995-2004[1]	1962-2000	1967-2005	2001
N. of Countries / Regions	239	150	72	71	96
Classification	HS0	HS0	SITC	CHELEM	GTAP
Disagregation Level	6-digit	6-digit	4-digit	3-digit	N.A.
N Commodities	5,041	5,041	1,276[3]	71	57

N.A.: Not Applicable. [1] The first public version of BACI was released for the 1995-2004 period but the HS 6-digit classification starts in 1989. COMTRADE provides with more datasets and years, but the coverage in terms of reporting countries is large enough since 1994. [3] This total number of products contains several items used to represent "residual categories", *i.e.*, trade within 3-digit code that could not be accurately assigned to a 4-digit code.

Unlike BACI, the NBER-UN database has not been built in a reconciliation perspective. In the NBER-UN database, the primacy is given to importer's reports, whenever they are available. If the importer report is not available for a country pair, the corresponding exporter report is used instead. Only some corrections and additions are made to the UN data for trade flows to and from the USA, exports from Hong Kong and China and imports into many other countries. Furthermore, since the new NBER-UN database spans on a long period (1962-2000), it covers a rather limited number of countries at a lower level of sector disagregation (72 exporting countries receiving imports from any country in the world at the 4-digit level of the SITC).

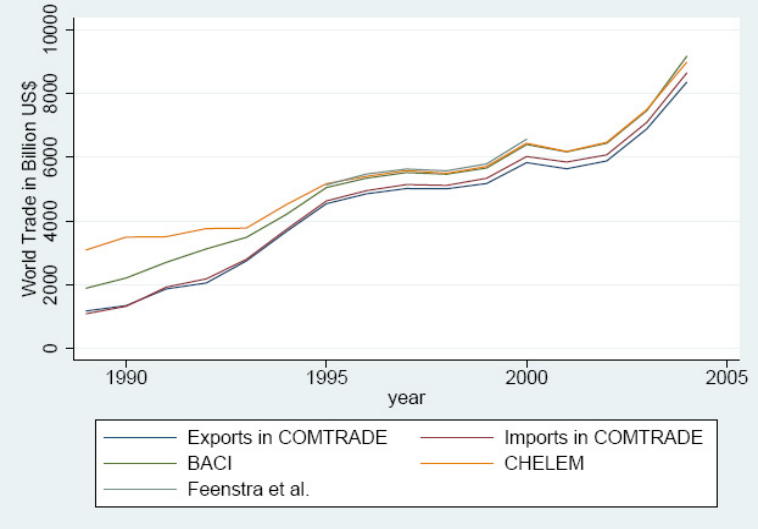
GTAP database is mainly devoted to applied general equilibrium analysis of global economic issues. The GTAP database combines, for a reference year (2001), detailed bilateral trade (also obtained from COMTRADE) with transport and protection data characterising economic linkages between regions, together with individual country input-output databases which account for inter-sector linkages within regions. Trade flows are not reconciled: only one flow is selected to build the world trade matrix. The choice of this flow is done on the basis of a comparison of

increase. We consider a missing observation as a zero when at least one of the trading partners does report its trade to the UN. If both partners are not reporting countries to the UN, then the missing observation is considered a true missing value.

reliability indices of the exporter and the importer. Finally, the level of disagregation is much lower than in BACI.

CHELEM provides, at a world level, commodity trade values in different sectoral classifications. Although CHELEM covers a longer time span (1967-2005) than BACI does, it is much more aggregated in the product and country dimensions and does not inform about quantities. The CHELEM reconciliation of mirror flows proceeds also to a *fobization* of import reports, taking into account the accuracy and the regularity of the declarations of the countries (de Saint-Vaulry, 2008). [26]

Figure 5 – Evolution of Total World Trade 1989-2004.

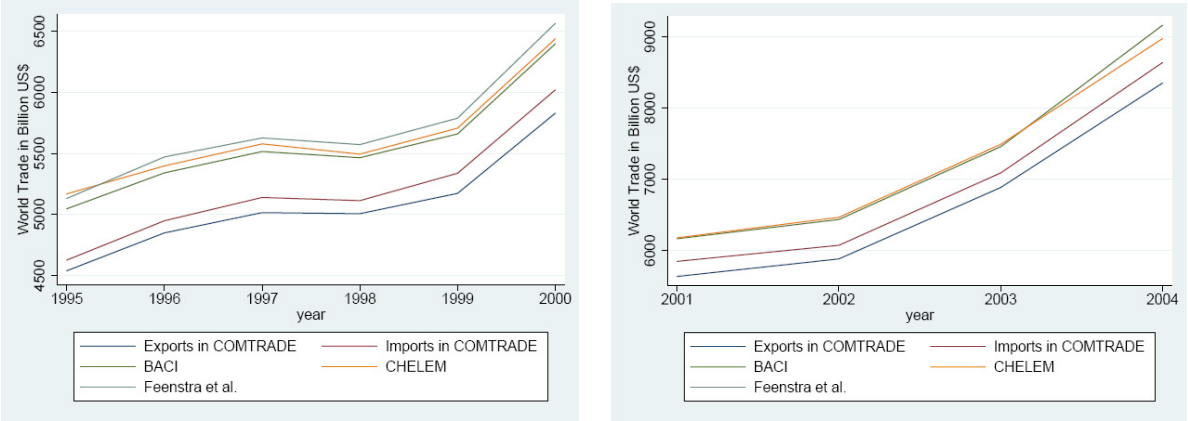


Source: Authors' calculations.

Figure 5 displays the evolution of the total world trade according to the above-mentioned databases. The evolutions are rather convergent. Note that BACI reaches in 1995 a total level of trade very close to CHELEM and to the NBER database. Figure 3.2 provides a closer look at two subperiods, 1995-2000 and 2001-2004, where the NBER dataset is available. Although very similar, CHELEM, NBER and BACI datasets exhibit some differences. The NBER database has higher values of trade for all years except 1995. This could be explained by the absence of harmonization of flows, *i.e.* the fact that CIF costs are not removed from NBER data. Actually, the difference with BACI is around 2%, close to the mean CIF estimated by BACI. The evolution of the recent years is depicted in the right side of Figure 3.2. During the period 2000-2004, CHELEM and BACI converge even more, except in the last year, where BACI exhibit more trade. Concerning the COMTRADE exports and imports, we see a stable gap with

[26]There is a tradition at the CEPII of compiling exhaustive trade data at the world level, using reconciled and stable trade classifications going back to the 1970s. The major interest of CHELEM is to provide a consistent view of the world economy in the long period. CHELEM provides also balances of payment, populations and incomes data.

Figure 6 – Evolution of Total World Trade



(a) 1995-2000 (b) 2001-2004

our database of 10% in terms of value of trade for exports, and of 5% for imports. This is likely because we use mirror information to include non-reported trade in our database.

3.3. Some applications

BACI has been largely used in CEPII's research to analyse trade patterns at the product-level, countries specialization, competitiveness, trade policy, exchange-rate pass-through, etc. Since its availability in 2007 to all users of COMTRADE, more than 200 international trade specialists have registered in the BACI webpage and one could imagine many other topics for which BACI can be useful. These economic analyses can benefit of the three main advantages of BACI, in comparison with other similar databases: its product detail, its geographical exhaustivity and its unit values.

Firstly, BACI allows international trade analysis at the most detailed product level. This can be needed for instance to assess the impacts of trade policy. For instance, Disdier, Fontagné and Mimouni (2007) use BACI to analyse the impact of SPS and TBT agreements on agricultural trade; Fontagné, Laborde and Mitaritonna (2004) study the impact of the EU-ACP Economic Partnership Agreements; Matthews and Gallezot (2006) assess the role of EBA in the political economy of CAP reform. Similarly, Gaulier and Zignago (2002) use an embryonary version of BACI to reveal market access difficulties at the product level. The analysis of international specialisation takes benefit of the product-level data to precisely identify some characteristics of products such as their main use in production (finals, intermediates or capital goods, Curran and Zignago, 2010, for instance), their technological content (World Bank, 2008; Mulder, Paillacar and Zignago, 2009; Cheptea, Fontagné and Zignago, 2009, etc.), their intra-industry flows at the world level (Ecochard, Fontagné, Gaulier and Zignago, 2006; Fontagné, Freudenberg and Gaulier, 2006), their cultural dimension (Disdier, Tai, Fontagné and Mayer, 2009).

Secondly, BACI geographical exhaustivity allows to draw a very complete view of the world trade. The European industry's positioning in the international division of labour has been oftenly analysed using BACI (Fontagné, Gaulier and Zignago, 2008; Cheptea et al., 2009, Curran and Zignago, 2010). But also the reorganisation of trade flows in Asia with the emergence of China (Gaulier, Lemoine and Ünal-Kesenci, 2006), or the market positioning of Latin America compared to Asia (Mulder et al., 2009). The most obvious gain in terms of geographical coverage is the African trade since several countries of the continent are not usually reporters in international trade databases (Fontagné et al. 2004).

Thirdly, BACI is especially designed to allow comparison of unit values of international trade. There is increasing empirical evidence that trade specialisation and competition takes place in varieties rather than in products or industries. This is confirmed in several studies using BACI to assessing the specialisation of countries or regions in terms of quality or market segments: Fontagné et al. (2008), Mulder et al. (2009), Curran and Zignago (2010). More generally, BACI is particularly useful when one want to analyse trade prices. Gaulier, Martin, Méjean and Zignago (2008) use it to provide *TradePrices*, a consistentdatabase of trade price indices. Gaulier and Méjean (2006) studies the aggregate price effect of newly imported varieties. Imbs and Méjean (2009) use BACI to structurally identify elasticities of substitution. Johnson (2009) estimate an heterogeneous firms trade model taking into account prices and use BACI to control for world prices.

4. CONCLUSION

International trade analysis is increasingly demanding for very detailed data. The aim of BACI is to provide researchers with the most disagregated database in terms of products, above all, but also covering the largest set of countries and years. The particularity of BACI is to provide not only values but also consistent quantities, allowing to the analysis of international trade prices via unit values.

In this working paper we describe the methodology developed to build BACI. We estimate the CIF rates and subtract them from the import values reported. We turn then to the comparison between mirrors declarations and the computation of quality indicators of country reports to average them. Under reasonable assumptions, we propose a rather simple statistic procedure – requiring no other input than raw trade values and quantities converted in tons - to provide consistent measures of international trade flows, more reliable since the possible errors in reported information are partly cleaned in the process.

The resulting database, BACI in its different classifications, is freely available since 2007 in our webpage to researchers having access to COMTRADE. The three main advantages of BACI in comparison to other trade databases are: its product detail, its geographical exhaustivity and its unit values. CEPII's research on international trade has often used BACI to study medium term changes in the international division of labor (quality of exported products, vertical differentiation, technological content, etc). BACI is particularly well suited to analyse international trade

prices since it provides unit values at a satisfactory product level, and more reliable than the raw data.

An important caveat must be recalled to users of BACI unit values: a change in the United Nations treatment of quantities affects the evolution of unit-values for some countries before and after 2005. Thanks to the UN collaboration, ongoing research is now focused on the raw data reported by countries to the UN, which has the advantage to be even more disagregated since countries report at their specific tariff-line level (6, 8, 10 or more digits). Even though this increased disagregation is not compatible at the international level, it is likely to reduce the agregation bias in the interpretation of unit values.

5. REFERENCES

ANDERSON J.E. (1979), "A Theoretical Foundation for the Gravity Equation", *American Economic Review* 69, 106-116.

ANDERSON J.E. AND E. VAN WINCOOP (2004), "Trade Costs", *Journal of Economic Literature* 42(3), 691-751.

BERGSTRAND J.H. (1985), "The Gravity Equation in International Trade: some Microeconomic Foundations and Empirical Evidence", *Review of Economics and Statistics* 67, 474-481.

BOUËT A., Y. DECREUX, L. FONTAGNÉ, S. JEAN AND D.LABORDE (2008), "Assessing Applied Protection across the World", *Review of International Economics* 16(5), pages 850-863.

CHENG I-HUI AND H.J. WALL (1999), "Controlling for Heterogeneity in Gravity Models of Trade", *Federal Reserve Bank of St Louis working paper* N° 99-010.

CHEPTEA A., L. FONTAGNÉ AND S. ZIGNAGO (2009), "European export performance", *CEPII Working Paper* 12.

CHEPTEA A., G. GAULIER AND S. ZIGNAGO (2005), "World Trade Competitiveness: a Disaggregated View by Shift-Share Analysis", *CEPII Working Paper* 23.

COOK R.D. (1977), "Detection of Influential Observation in Linear Regression" *Technometrics* 19(1), 15-18.

CURRAN L. AND S. ZIGNAGO (2010), "How regional is the supply chain in the new EU? An analysis of the effect of enlargement on EU trade in intermediate products", *Multinational Business Review* Vol 18:1.

DE SAINT-VAULRY A. (2008), "Base de données CHELEM – commerce international du CEPII", *CEPII Working Paper* 09.

DEARDORFF A. V. (1998), "Determinants of Bilateral Trade: Does Gravity Work in a Neoclassical World?" in J.A. Frankel ed., *The Regionalization of the World Economy*, University of

Chicago Press.

DISDIER A-C., L. FONTAGNÉ AND M. MIMOUNI (2007), "The Impact of Regulations on Agricultural Trade: Evidence from SPS and TBT Agreements", *CEPII Working Paper* 04.

DISDIER A-C., L. FONTAGNÉ, T. MAYER AND S.H.T. TAI (2009), "Bilateral Trade of Cultural Goods", *Review of World Economics*, 145(4): 575-595.

DISDIER A-C. AND K. HEAD (2007), "The Puzzling Persistence of the Distance Effect on Bilateral Trade", *Review of Economics and Statistics* 90(1): 37-41.

ECOCHARD P., FONTAGNÉ L., GAULIER G. AND ZIGNAGO S. (2006), "Intra-Industry Trade and Economic Integration, in D. Hiratsuka, *East Asia's De Facto Economic Integration*, Macmillan.

EVENETT S.J. AND W. KELLER (2002), "On Theories explaining the Success of the Gravity Equation", *Journal of Political Economy* 110(2), 281-316.

FEENSTRA R.C. (1996), "U.S. Imports,1972-1994: Data Concordances", *NBER working paper* 5515.

FEENSTRA R.C. (2002),"Border Effect and the Gravity Equation: Consistent method of Estimation", *Scottisch Journal of Political Economy*, 49, 491-506.

FEENSTRA R.C., R. E. LIPSEY AND H.P. BOWEN (1997), "World Trade Flows, 1970-1992, with Production and Tariff Data", *NBER working paper* 5910.

FEENSTRA R.C., R. E. LIPSEY, H. DENG, A. C. MA AND H. MO (2005), "World Trade Flows: 1962-2000", *NBER working paper* 11040.

FEENSTRA R.C, J. ROMALIS AND P.K. SCHOTT (2002), "US Imports,Exports and Tariff data, 1989-2001", *NBER working paper* 9387.

FONTAGNÉ L., G. GAULIER AND S. ZIGNAGO (2008), "Specialisation across Varieties within Products and North-South Competition", *Economic Policy* 23.

FONTAGNÉ L., M. FREUDENBERG AND G. GAULIER (2006), "A Systematic Decomposition of World Trade into Horizontal and Vertical IIT", *Review of World Economics* 142 (3) : 459-475.

FONTAGNÉ L., D. LABORDE AND C. MITARITONNA (2008), "An Impact Study of the EU-ACP Economic Partnership Agreements (EPAs) in the Six ACP Regions", *CEPII working paper* 04.

MATTHEWS A. AND J. GALLEZOT (2006), "The role of EBA in the political economy of CAP reform", in *Everything But Arm*, Routledge ed.,June, Ghent editor.

GAULIER G., F. LEMOINE AND D. ÜNAL-KESENCI (2006), "China's Emergence and the Reorganisation of Trade Flows in Asia", *CEPII Working Paper* 05.

GAULIER G. AND I. MÉJEAN (2006), "Import Prices, Variety and the Extensive Margin of Trade"", *CEPII Working Paper* 16.

GAULIER G. AND S. ZIGNAGO (2002), "La discrimination commerciale révélée comme mesure désagrégée de l'accès au marché", *Economie Internationale* 89-90.

GROSSMAN G. (1998), "Comments on Deardorff", in J.A. Frankel ed;, "The Regionalization of the World Economy", University of Chicago Press.

IMBS J. AND I. MÉJEAN (2009), "Elasticity Optimism", *CEPR Discussion Paper* 7177 and *Working Paper Ecole Polytechnique* 2009-05.

JOHNSON R.C. (2009), "Trade and Prices with Heterogeneous Firms", *mimeo*.

HUMMELS D. (2001), "Toward a Geography of Trade Costs", Global Trade Analysis Project Working Paper 17, Purdue University.

HUMMELS D. AND V. LUGOVSKYY (2006), "Are Matched Partner Statistics a Usable Measure of Transportation Costs?", *Review of International Economics* 14(1), 69-86.

HUMMELS D. AND A. SKIBA (2004), "Shipping the Good Apples Out? An Empirical Confirmation of the Alchian-Allen Conjecture", *Journal of Political Economy* vol. 112(6), 1384-1402.

LIMÃO N. AND A.J. VENABLES (2001), "Infrastructure, Geographical Disadvantage, and Transport Costs", *The World Bank Economic Review* 15(3), 451-479.

MATYAS L. (1997), "Proper Econometric Specification of the Gravity Model", *The World Economy* 20, 363-368

MAYER T. AND S. ZIGNAGO (2011, "Notes on CEPII's Distances Measures: The GeoDist Database", *CEPII Working Paper* 23.

MULDER N., R. PAILLACAR AND S. ZIGNAGO (2009), "Market Positioning of Varieties in World Trade: is Latin America Losing Out on Asia?", *CEPII Working Paper* 09.

TEN CATE, A. (2007), "Modelling the reporting discrepancies in bilateral data", www.cpb.nl

UNITED NATIONS (2004), "International Merchandise Trade Statistics: Compilers Manual", *UN Statistics Division (UNSD), Department of Economic and Social Affairs*, Series F, No.87. 114 p.

WORLD BANK (2008), "Determinants of Technological Progress: Recent Trends and Prospects", in *Global Economic Prospect 2008, Technology Diffusion in the Developing World*, Chapter 3, pages 105-164.

6. APPENDIX: ALLOCATION OF "AREAS NOT ELSEWHERE SPECIFIED"

COMTRADE has some trade data without specification of destination or origin, classified as Areas Not Elsewhere Specified (NES). BACI deals with these cases by conferring to these flows a new allocation when possible, in order to correct one of the sources of discrepancies between mirror flows: when a country is reporting a flow towards a trading partner which reports instead a non specified areas. The reconciliation procedure tends thus to underestimate the real flow, since this last reported value is inferior to the true bilateral value.

The reallocation of these non specified flows is made according to the weight of the partner countries that have reported flows of the commodity under consideration. Suppose an exporting country i reports "Area NES" flows for a given commodity in a given year. If the sum of the flows towards partner countries reported by i ($\sum_i Vx_{ij}^{kt}$) is less than the sum of the mirror reported values ($\sum_i Vm_{ij}^{kt}$), then it is guessed that all (or a part of) the flows declared as "Area NES" ($Vnes$) are in fact devoted to these identified partners. We suppose the same distribution of partners in the non specified flows as in "missing" flows ($\sum_i Vm_{ij}^{kt} - \sum_i Vx_{ij}^{kt}$) and reallocate the minimum between them.

After having subtracted the total reallocated value from the $Vnes$, the residual value of the $Vnes$ (denoted by $Vnes'$) is compared with the sum of the declarations from partner importing countries which have no mirror in the declarations of the exporter (Vm'). If $Vnes_i'^{kt}$ is less than the sum of $\sum_j Vm_{ij}'^{kt}$; then $Vnes'$ is assumed to be included in Vm' and in order to avoid double counting $Vnes'$ is set to zero. Otherwise, if $Vnes'$ is more than Vm', then the value Vm' is substracted from $Vnes'$.

Note that in BACI such an incremental procedure of the country reports – which is the choice between on the one hand Vx and Vx' and on the other hand Vm and Vm' – is only done to the extent that the outcome is a reduction of the gap between mirror flows. About 11.5% of final flows are concerned by this treatement.[27]

[27]Besides of "Area NES" reported by a given country, there is also reported destinations such as "Asia NES". No treatment is done in these cases to avoid a double counting in the sum of the harmonized values per countries. Note that the noise generated by this class of "NES" is of a weak extent since such "NES" reportings generally correspond to flows towards non reporting countries. There also exists in COMTRADE a category "Commodities NES", which is dropped in BACI also to avoid double counting due to the fact that partners may classify commodities in other category than "NES". The extent of this underestimation is abstemious since this type of flows concerns mostly specific commodities such as military equipment or commodities for which no adequate category would has been found in the HS.

LIST OF WORKING PAPERS RELEASED BY CEPII