BACI: International Trade Database at the Product-level

Guillaume Gaulier and Soledad Zignago

CEPII

May 2009

May 2009

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 is increasingly demanding for accurate and disaggregated trade statistics. This paper document the construction of BACI, our detailed international trade database covering more than 200 countries and 5,000 products, between 1994 and 2006. Original procedures have been developed to reconcile data reported by over 150 countries to the United Nations Statistics Division, which disseminate them via COMTRADE. When both exporting and importing countries do report, we have two different figures for the same flow. In order to have a single consistent figure of a bilateral flow, we reconcile them upon the following procedure.

Firstly, to enable comparisons between import values, generally reported CIF (cost, insurance and freight), and exports values, reported FOB (free on board), we need to estimate the transport cost. In order to go back to FOB-FOB data, an average CIF cost is estimated and removed from the declarations of imports to provide FOB import values. The average CIF cost is estimated using a gravity-type equation by an OLS method on pooled data. Since we observe a strong positive relation between values and quantities ratios, the CIF-FOB observed ratios of unit values are our dependent variable. Hence, we assume that errors or differences in countries ways of reporting are likely to affect in the same way values and quantities reported for a given traded product. For the right-hand-side variables, the gravity-type equation takes into account bilateral distance, dummies for contiguity and landlocked feature of countries, dummies for years, and the world median unit-value for each product category. A non-linear relationship is considered between CIF implicit costs and distances by introducing the squared distance.

Secondly, we need a criteria to average mirror numbers. We evaluate the reliability of countries reporting by computing a "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 (the share of poor/good reporters in its trade partners and the share of products with frequent report errors due to lack of homogeneity in the 6-digit position for instance). These estimated qualities of reporting are finally used as weights in the averaging of mirror flows.

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 methodology is purely statistic and it does not require very much data. BACI database is useful to manage several countries data with multiple dimensions since the main aim of this work was to provide ourself and the scientific community with an international trade database covering the larger number of countries at the highest degree of disaggregation (5,000 products of the Harmonized System). Studies on African trade have take benefit of its geographical coverage. The CEPII 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 (for instance, $TradePrices$ database uses unit-values from BACI to provide coherent international trade price indices between 1995 and 2004).

Since 2007, researchers using Comtrade can freely download our BACI database on several classifications (HS92, HS96, CTCI). One important caveat must to be mentioned regarding the time coverage. A different treatment of the source data, aiming to limit missing quantities, is applied to years have induce a serious breakdown in the evolution of unit-values. However, this correction seems to impact mostly little exporters. This is why we provide with values and quantities for the period 1994-2006.

ABSTRACT

We document BACI, our international trade database covering more than 200 countries and 5,000 products, between 1994 and 2006. Original procedures have been developed to reconcile data reported by almost 150 countries to the United Nations Statistics Division, which disseminate them via COMTRADE. When both exporting and importing countries do report, we have two different figures for the same flow, which is useful to reconcile in a single figure. Firstly, as import values are reported CIF (cost, insurance and freight) and the exports are reported FOB (free on board), CIF costs have to be estimated and removed from imports values to compute FOB import values. We regress the unit-values ratios reported for a given elementary flow by gravity variables and for years, and world median unit-value for each product category . The second step is an evaluation of the reliability of country reporting, based on the reporting distances among partners. We decompose the absolute value of the ratios of mirror flows using a (weighted) variance analysis, and an index is build for each country. These reporting qualities are used as weights in the reconciliation of each bilateral trade flow twice reported. Taking advantage of this double information on each flow, we end up with a large coverage of countries not reporting at a given level of the product classification with a special care in the treatment of unit-values. BACI is freely available to users of COMTRADE database in our webpage: http://www.cepii.fr/anglaisgraph/bdd/baci.htm

JEL Classification: F10, F14, F13, C80.

Keywords: International Trade, Trade Costs, CIF-FOB, Data reconciliation.

BACI: BASE POUR L'ANALYSE DU COMMERCE INTERNATIONAL

RÉSUME NON TECHNIQUE

Ce document de travail...

RÉSUMÉ COURT

Ce document de travail

Classification JEL : F10, F14, F13, C80.

Mots clés : Bases de données, Commerce, coût au commerce, CIF, BACI, Harmonisation des flux du commerce.

BACI: International Trade Database at the Product-level
The 1994-2007 Version[1]

Guillaume Gaulier*
Soledad Zignago†

1. Introduction

Empirical studies in international trade are increasingly demanding 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, the countries can be reporting in a more aggregated classification. At the international level, the finest product classification is the 6-digit Harmonized System (HS), which applies progressively from 1989 and distinguish 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

[1] The development, communication and diffusion 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 those having closely participating, Matthieu Crozet, Yvan Decreux, Alix de Saint-Vaulry, Houssein Boumelassa and, more recently, Charlotte Emlinger and Antoine Berthou, and those having use BACI intensively such as, Lionel Fontagné, Isabelle Méjean, Jacques Gallezot, among others. We are indebted to the United Nations Statistical Division for our collaboration, in particular to Matthias Reister, Ronald Jansen and Markie?? 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.

† CEPII (soledad.zignago@cepii.fr).

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

[3] Underneath this level, there is not a common international classification of commodities. In other words, national or regional customs having adopted the Harmonized System differ underneath the 6-digit level, at their tariff-lines level, in which they report their trade to the UN.

products in its 4-5 digits level.[4]

Countries reporting their imports and their exports, we have actually: 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 providing 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).

Original procedures have been developed to reconcile flows reported by exporters and importers. Firstly, because import values are reported CIF (cost, insurance and freight) and the exports are reported FOB (free on board). To allow the comparison between mirror declarations, CIF costs have to be estimated and removed from imports values to compute FOB import values. We use a gravity-type equation to estimate them. Secondly, an assessment of the reliability of country reports is then used as weights in the average of mirror values and quantities.

This working paper documents these reconciliation methodology applied to public versions of BACI, providing information at the most disaggregated level. The most known and used version of BACI, BACI0 in the following, gives values and quantities for more than 200 countries at the HS 6-digit level in its fisrt version, HS0. Since its first version, the HS was importantly revised in 1996 (HS1), 2002 (HS2), and 2007(HS3). We have applied our reconciliation procedure firstly to the HS0 in order to have the largest time period, between 1995 and 2004. Most of results illustrating our methodology use then also the BACI0 version.

Other versions of BACI, based in the same methodology, BACI1, BACI2, BACISITC, BACITL, are also available (or will be soon) and are based respectively on the HS1, HS2, SITC and the country specific tariff lines classification (TL). Since July 2007, BACI0 is available for COMTRADE users in our webpage: www.cepii.fr/anglaisgraph/bdd/baci.htm. Until recently, BACI0 was updated only for values and the year 2005 after a change in the UN Statistical Division treatment of quantities from the year 2005 onwards (and will probably change also years before 2005 in the future). This methodological change, aiming to reduce missing quantities, tend to reduce variance in the unit-values and can introduce a serious breakdown in the evolutions of unit-values that have to be underlined. Even if users are invited to have this breakdown in mind when they are interested in the long-term evolution of unit-values applied by some developing exporters, this correction seems to impact mostly little exporters. The updated version of BACI0 in values and quantities and covering the period 1994-2007 is named BACI0_2007.

BACI is largely used in the CEPII work 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 par-

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

ticularly convenient to analyse international trade prices. Its exhaustive coverage is useful also to analyse international of generaly non reporting countries such as African countries. BACI is also an input to other CEPII databases like TradeProd, TradePrices and MacMap.

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 ?? 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 ?? concludes and announce future developements.

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 (in its full title, HS in the following) since it is the more detailed classification (over 5,000 products) at the international level. by criteria based on raw materials and the stage of production of commodities. The industrial origin criterion is considered whenever it is compatible with the main criteria set out above.[5] 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. The objectives: to harmonise a) external trade classifications to guarantee direct correspondence; and b) countries' external trade statistics and to guarantee that these are comparable internationally. The HS is organized in four hierarchical levels:

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

[5] 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).

World Customs Organization revise the HS in principle every few years. But since its first version, in 1988, also known as 1992 (HS0 in the following), the HS was importantly revised in 1996 (HS1), 2002 (HS2), and 2007(HS3). We have applied our reconciliation procedure firstly to the HS0 in order to have the largest time period and end up with the most known version of BACI covering more than 200 countries and 5,000 products, between 1995 and 2004 (BACI0 in the following). Most of results illustrating our methodology use then also the BACI0 version. Other versions of BACI, BACI1, BACI2, BACISITC, BACITL, are described further and are based respectively on the HS1, HS2, SITC and the country specific tariff lines classification (TL).

An increasing number of countries (See Figure 1) report to the United Nations Statistics Division their annual international trade statistics detailed by commodity and partner country. The UN disseminates them via COMTRADE (Commodities Trade Statistics database), which provides over one billion commodities trade data, relying to more than 95% of the world trade. Imports, exports, re-imports and re-exports, are reported in the current classification and revision: ?? the Harmonized System (HS) revision 2002 in most cases in the recent years. According to the UNSD (2004), since 2001 there are 102 countries that are *Contracting Parties*, *i.e.* they recognized the Harmonized System as a legal instrument. Another 78 countries are not *Contracting Parties*, but use the HS System. As the more recent product classification can be converted generally all the way down to the earliest classifications, time series can start as far back as 1962 and go up to the most recent registered year.

The COMTRADE database is used as the unique source of information to build BACI. It provides with trade data at various level of aggregation but the most disaggregated is the Harmonized System 6-digit level, available since 1989.[6] For the current version of BACI, we have extracted all the reported data classified in HS from 1992 and 1996.[7] Data do not includes flows below 1,000 dollars.[8].

For a given level of disaggregation, COMTRADE comprises two sets of series when 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).[9] In principle exports from country i to country j should be identical to imports from country i to country j, for any given product, 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 be-

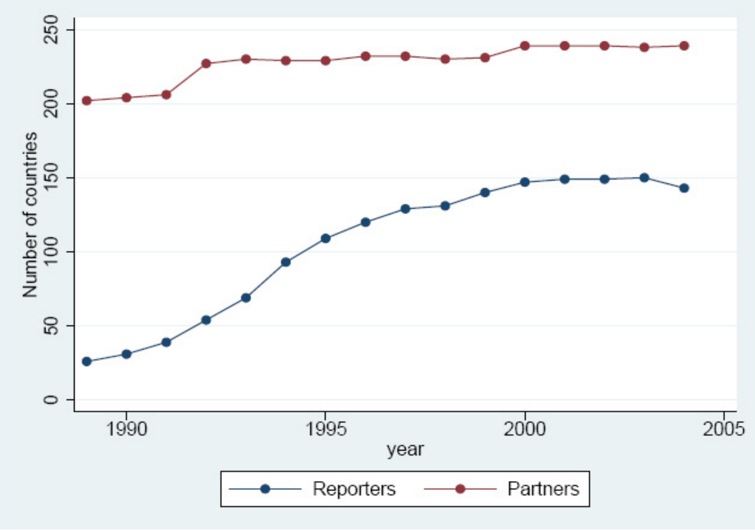
[6] COMTRADE also provides with longer series, starting in 1967, for more aggregated product decompositions and available for a larger number of countries.

[7] More recently we have also applied our reconciliation methodology to more agregated data, in SITC, to have a larger time coverage. We focus however in this document in the HS version since it is the most detailed and then the more accurate to deal with unit-values, which is the first aim of the BACI construction.

[8] Further details on COMTRADE see http://unstats.un.org/unsd/comtrade/.

[9] It should be mentioned that there are many other regimes of delivery. UNSD (2004) identifies 13, according to the costs actually involved in the reported value of the country. See UNSD (2004) Annex B, p. 55.

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



cause this determines the level of tariff that will be applied to it. On the other hand they may be less careful when it comes to the actual destination of exports. If the product is to be reexported to country *j'* after some little modifications, they will still consider it as an export to *j*. 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 class trade declaration). Many other source of misreport can be imagined: product misclassification, different reporting year, fraud... We will see that the difference between the two reported figures may be significant for some flows but the purpose of COMTRADE is to provide information as close as possible to original reports, not to try and reconcile them.

BACI can represent a useful tool for international trade analysis at high degrees of disaggregation, in complement to COMTRADE. Firstly, it provides in a coherent database values and quantities allowing for **international comparison**. Rending comparable import and export reporting, allow us to largely complete missing reportings. Secondly, it provides **comparable quantities** and thus unit values. Whereas values are done 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 .[10] The next subsection describes the first step of our methodology, the estimation of transport cost included in import reports and excluded in export reports.

[10] This implicit rate of conversion is then applied to flows reported in heterogeneous units.

2.2. Conversion in tons

Even if most of quantities are reported in tons, there is 15% reported in other quantity units(Units, meters, watt, etc). The international trade analysis needs reliable data on unit-values (values divided by quantities) of products exchanged to investigate prices, or quality issues. Having the objective of exhaustibility, rates of conversion by product are estimated between different quantity units, using mirror flows reported in tons by a country and in another unit by the other trade partner. Quantities reporteed in unknown units or in Kwh are dropped for simplicity.

When an implicit rate of conversion is obtained for a product, the conversion is only performed if a minimum of 10 data have been used in its construction, and if the standard deviation is inferior to 2.5. Note also that the conversion is applied to quantities of mirror flows (that is, before the harmonization process). Consequently, these rules applied to both quantities, and it is possible that only one of both is converted. About 8,5% of final flows in BACI have been coverted using this method.

2.3. The CIF-FOB ratios estimation

The present subsection presents the estimation of CIF-FOB ratios. Generally importer 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 the comparison with export reports.

2.3.1. Empirics on the evaluation of transport costs

Direct transport costs are rarely available at the product-levels. With six importer countries, Hummels and Skiba (2004) paper is one of the most complete: Argentina, Brazil, Chile, Paraguay, Uruguay and the USA provide very precisely their bilateral freigth costs. Concerning this latter country, the NBER via Robert Feenstra 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 of such limited coverage all the cross country variability of real freigth costs in all pair country combinations. A flourishing literature has then discussed the way to assess these costs.

A first class of empirical papers rely on directly measured trade barriers in terms of detailed freight and tariff rate data for a limited number of countries. For instance Hummels (2001) exploiting data on U.S. imports from U.S. Census Bureau shows the wide dispersion in freight rates over commodities and across countries in 1994. The all-commodities trade weighted average transport cost from national customs data ranges from 3% of FOB price for he U.S. to 13.3% for the Paraguay. [11] Alternatively, Limão & Venables (2001) highlight the dependence

[11] Hummels (2001) starts from a multi-sector model of trade and uses a more sensible trade costs function than

of trade costs on infrastructure. [12]

In 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) have been revisited. 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 imports and exports flows reported in values and in the same time by importing and exporting country may exceed 100% for more than half of the observation in UN COMTRADE database. Indeed, given the fact that 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 the shipments, significant discrepancies in CIF – FOB ratios are recorded. Note that the discrepancies need not to 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 of 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, since the huge discrepancies between mirror flows, they 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 UN 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 if 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 error but are still usable. Hummels & Luguvskyy (2006) state that IMF CIF – FOB ratios only seem to reveal meaningful cross-exporter

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.

[12] 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.

variation that might be usefully exploitable 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. For instance, 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

Our estimation of CIF costs use indirectly the implicit CIF-FOB ratios: the predicted mirrors flows are used then as *estimates* of CIF between partner countries i and j ($\widehat{CIF}_{ij}^{kt}$). The implicit ratios are regressed within a gravity-type equation on a set of explanatory variables.

Indubitably, 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. In this perspective, we consider a non linear relationship between distance and CIF rate and the distance (denoted $dist_{ij}$) and its square (denoted $Dist_{ij}^2$) are introduced in the gravity equation. 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 geographical variables are taken from CEPII's distances database.[13]

Besides distance, the gravity equation includes as explanatory variables the world median unit value for product k (value/quantity or UV_k) which aims at to capture the transportability of the commodities. In other words, it controls for the higher costs of trading heavy commodities. UV^k is unit value, which is a world-median for each 6-digit product (there is no country dimension).

The gravity equation also introduces time dummies in order to capture any potential time evolution of the CIF rate.[14] Thus, the gravity equation, estimated by OLS on pooled data over the period 1989-2007, is basically as follows:

[13] Available at http://www.cepii.fr/anglaisgraph/bdd/distances.htm. There are two kinds of distance measures: Simple distances, for which only one city is necessary to calculate international distances, and weighted distances, for which we need population, latitude and longitude data on principal cities in each country. We use weighted distances when available (148 countries out of 225 partner countries).

[14] A suitable 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.

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

We consider four different gravity equations in which the dependent variable is alternatively the CIF-FOB ratios in values or in unit-values, weighted 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 Q denotes quantities reported by the exporter, Qx, or by the importer Qm) or not. 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, still partly cleaned from noise and denoted UVm_{ij}^{kt} and UVx_{ij}^{kt} for importer and exporter reports respectively.[15] The weighting confers a higher importance to trade flows equally reported by partners, differences between reported import and export values are then more likely to correspond to freight costs.

Table 1 presents the estimation results of the gravity equation on for the 1989-2004 period. Note that, except for exporter landlocked, there is no reversion of signs in the coefficients when the dependent variable changes, and the magnitudes are similar, resulting in similar estimations for the mean CIF, 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 distance influence in the estimated CIF costs. The sign for contiguity supports the idea that the CIF rate decreases when two partner share the same land border. In contrast, the sign of the coefficient of the variable capturing the landlocked status of a given country 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 less easy. This is confirmed in all models for the importer country, but in the case of exporters that are landlocked, the results using unit-values ratios as dependent variable are slighlty negative.

For the estimation, the database contains more than 9.3 millions of observations. However, in order to assure consistent and robust parameter estimates, a statistical mopping-up operation is used to remove atypical and influential observations.[16] Weighted regressions (specially Model II) suffers less of this procedure, allowing for an estimation with more observations. Despite the negative value for the exporter landlocked, we consider the Model II in the following as our preferred estimation.

Employing the coefficients to generate a mean for the CIF results in a rate of 3,3%. This value is

[15] 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 since they incorporate both information.

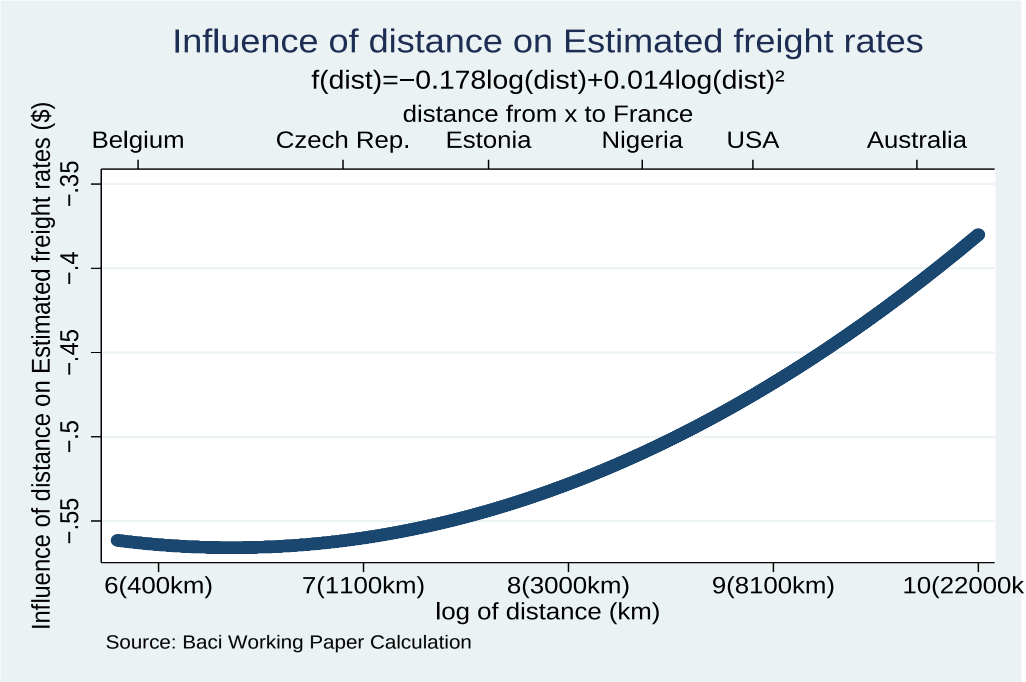
[16] To identify those observations, we compute the D distance of Cook (1977).

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***	0.32***	0.442***	0.3***
	(0.016)	(0.011)	(0.01)	(0.007)
$\ln Dist_{ij}$	-0.178***	-0.11***	-0.122***	-0.086***
	(0.004)	(0.003)	(0.002)	(0.002)
$\ln Dist_{ij}^2$	0.014***	0.01***	0.009***	0.007***
	(0.000)	(0.000)	(0.000)	(0.000)
$\ln UV^k$	-0.032***	-0.032***	-0.042***	-0.038***
	(0.000)	(0.000)	(0.000)	(0.000)
$Contiguity_{ij}$	-0.066***	-0.044***	-0.025***	-0.024***
	(0.001)	(0.001)	0.001)	(0.000)
$Landlocked_j$	0.066***	0.049***	0.024***	0.02***
	(0.001)	(0.001)	(0.001)	(0.000)
$Landlocked_i$	-0.021***	-0.009***	0.012***	0.01***
	(0.001)	(0.001)	(0.001)	(0.000)
Time FE	Yes	Yes	Yes	Yes
N. obs.	8856312	9053610	8897367	8936618
R^2	0.008	0.012	0.014	0.02
Outlier values	482840	285542	441785	402534
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 and in the two last ones there is the ratio of mirror values (UVm_{ij}^{kt} and UVx_{ij}^{kt} are respectively importer and exporter reported unit-values for the same flow from i to j. Standard errors in parentheses. ***, ** and * denote a significant coefficient at 1%, 5% and 10% respectively. Models II and IV are weighted by $Min\left(Qx_{ij}, Qm_{ji}\right)/Max\left(Qx_{ij}, Qm_{ji}\right)$.

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



Source: Baci Working Paper Calculation

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 among 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 with regard of distance (denoted π). We take from Anderson & van Wincoop (2004) the following equation linking the elasticities in the both alternative specification(the Hummels one and ours) : $\gamma = \pi\ cif/(1 + cif)$, where γ is the *implicit* CIF rate with regard of distance in our specification. Using a $cif = 12\%$ (taken from Anderson & van Wincoop, 2004), the π reported by Hummels (0.27) is consistent with our result.

We remove its equivalent in value from the import trade flows. We have implement several criteria to assure that this procedure will truly improve trade data (and never worse it).[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).

[17] There some additional criteria to cope with particular cases: (1) the procedure is not implemented to countries which do not report their flows in CIF (such Algeria, Georgia, South Africa and other SACU countries); (2) In countries that declare 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 CIF rate is set to zero.

2.4. Evaluation of the quality of country declarations

In this subsection we explain the evaluation of the quality of contry reports, which will serves as weights in the averaging procedure between reported mirror flows, now cleaned from CIF costs. This harmonisation concerns 35% of observations (those for which both mirror flows exist). The procedure of reconciliation consists in computing weighted averages of mirror figures on the basis of an estimated quality indicator of import and export reportings for each country. The second step in the our methodology aims thus at determining a ranking for the qualities of the country reports. This evaluation is obtained using a (weighted) variance analysis via a decomposition of the absolute value of the mirror flows ratios, considered as "reporting distances" ($RD_{ij}^{kt} = \left|\ln\left(\frac{VM_{ij}^{kt}}{VX_{ij}^{kt}}\right)\right|$).

It is assumed in the following model that the logarithm of absolute difference between mirror figures can be decomposed as a sum of four parts: a part due to the exporting country i, a part due to the importing country j, a part due to the year t, and a part due to the product k.

$$RD_{ij}^{kt} = \alpha_i + \beta_j + \lambda_t + \gamma_k + \varepsilon_{ij}^{kt} \quad with \ \sum_i \alpha_i = \sum_j \beta_j = \sum_t \lambda_t = \sum_k \gamma_k = 0 \qquad (2)$$

Based on this OLS estimation, three groups of variables are obtained: the adjusted means of the dependent variable (denoted by $LS\_\overline{RD_i}$ and $LS\_\overline{RD_j}$), the standard errors (denoted by $stderr_i$) and the number of flows for each importer, each exporter and each year. The estimation of fixed effects gives the marginal impact – that is adjusted for the influence from the other factors - on discrepancies between flows that can be attributed to all country or sector. Each observation is weighted with the natural log of the sum of the two reports. 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 report errors because of lack of homogeneity in the 6-digit position for instance). The product dimension is taken into account through the within transformation, to avoid employing the more than 5,000 fixed effects needed.

The quality indicator (RQ_i) is generated considering the adjusted mean (*e.g.* $LS\_\overline{RD_i}$ for the country i) and the standard deviation of the coefficient, in order to capture also the precision of the estimation (details about the specific formulae employed are in the Appendix). Hence, the rank of the country is obtained by ranging in ascending order the estimated qualities. The same principle is used for the evaluation of the quality of reporting quantities. Figures 3 and 4 exhibit the quality indicators of country declarations as exporter for values and quantities (similar results are obtained for the countries as reporting importers, although the places in the ranking are not necessarily the same). We see that both measures are correlated, as expected. The worst reporters are usually affected for a relatively high standard deviation. 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.

Figure 3 – Quality of exporter reports on quantities and values

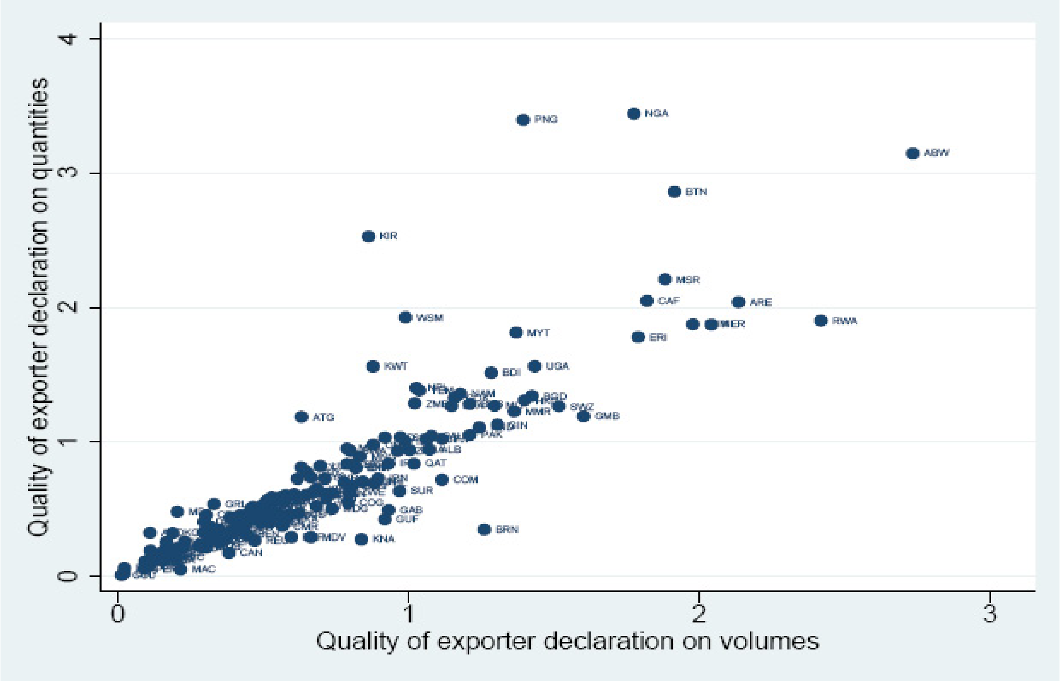
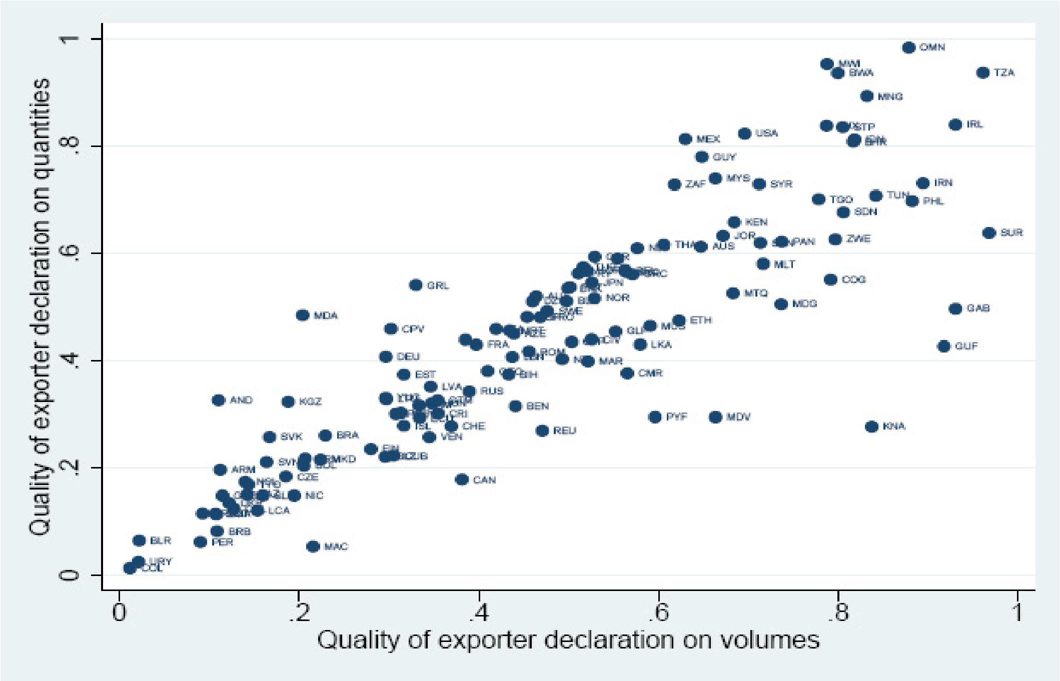


Figure 4 – Quality of exporter reports on quantities and values for better reporters



2.4.1. Reconciliation of mirror values and quantities

The last step relies on the averaging of two figures to be reconciled. The reporting qualities (RD_i, RD_j) are converted into weights (denoted as Wm and Wx for the importer and exporter, respectively) to be used in the weighted averages of raw declarations. Specifically, the weights for importer (Wm) and exporter (Wx) must fulfill this condition: $Wm+Wx = 1$, with $Wm = f\left(RQ_i, RQ_j\right)$ and $Wx = f\left(RQ_i, RQ_j\right)$. We chose these functions that minimize the variance of the errors (we assume log-normal multiplicative errors in the values reported by countries):

$$Wm = \frac{e^{(RQ_i)^2}(e^{(RQ_i)^2}-1)}{e^{(RQ_i)^2}(e^{(RQ_i)^2}-1)+e^{(RQ_j)^2}(e^{(RQ_j)^2}-1)},$$

$$Wx = \frac{e^{(RQ_j)^2}(e^{(RQ_j)^2}-1)}{e^{(RQ_i)^2}(e^{(RQ_i)^2}-1)+e^{(RQ_j)^2}(e^{(RQ_j)^2}-1)}$$

The harmonization will affect values as well as quantities when both mirror flows exist.[18] We perform a series of rules and thresholds. The more general case (when the absolute and relative error are small), considers both weights:

- for quantities : $Q_{ij}^{kt} = WmQm_{ij}^{kt} + WxQx_{ij}^{kt}$
- for values : $V_{ij}^{kt} = WmVm_{ij}^{kt} + WxVx_{ij}^{kt}$

In sum, this harmonization method 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 ordered by their quality of reportings.

3. Comments on the resulting datasets

3.1. Different versions of BACI

The methodology described above was firstly applied to the HS0 to maximise the number of available years. Since July 2007, **BACI0** is available for COMTRADE users in our webpage, but until recently, it was updated only for values in 2005 after a change in the UN Statistical Division treatment of quantities from the year 2005 onwards (and will probably change also years before 2005 in the future). This methodological change, aiming to reduce missing quantities, tend to reduce variance in the unit-values and can introduce a serious breakdown in the evolutions of unit-values that have to be underlined. Even if users are invited to have this breakdown in mind when they are interested in the long-term evolution of unit-values applied by some developing exporters, this correction seems to impact mostly little exporters. The updated version of BACI0 in values and quantities and covering the period 1994-2007 is named **BACI0_2007**.

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

Other versions are also available online. In order to match correctly with protection data of MAcMap-HS6 (Bouët et al., 2008), which is generally in the current version of the HS, **BACI1** is based on HS1 data (from 1996 revision) and covers the period 2000-2004??.

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 gives trade and production industrial data in a consistent classification (ISIC) for a long time period: **BACISITC** covers the period 1980-2006.

Our recent collaboration agreement with the US Statistics Division allow us to deal with their source data to compile COMTRADE: country-specific tariff lines (TL) datasets. Since the coherence of unit-values is our main concern in the developement of BACI, more disagregated data permit to better proxy prices reducing the agregation bias. **BACI2TL** is an agregation in HS2 6-digit level of the tariff-lines unit values.

Users of COMTRADE can register themselves in our webpage (http://www.cepii.fr/anglaisgraph/bdd/baci.htm) and freely download our datasets, available by year in the csv format. They will find also complementary information such as country and product codes. BACI users are kindly 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. In particular, we consider the NBER database from Feenstra et al. (2005), the CHELEM database from CEPII[19], the GTAP Project[20] and COMTRADE itself. A general comparison is presented in Table 2.[21] Overall, the highest disaggregation level is reached with the BACI and COMTRADE datasets.

The NBER-UN database has not been built on reconciled trade flows in a harmonization perspective, as BACI is. 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

[19] There is a tradition at the CEPII of compiling exhaustive trade data at the world level, using harmonized and stable trade classifications going back to the 1970s. As a result, the CEPII's CHELEM (French acronym of "Comptes Harmonisés sur les Echanges et L'Economie Mondiale) trade database is informative regarding the big shifts of countries' specialisation and the emergence of new competitors. The long period covered (since 1967) allows the current changes to be placed in a medium term perspective, helpful for the analysis. See De Saint-Vaulry (2008) for detailed documentation on the CHELEM international trade database.

[20] "Global Trade Analysis Project", mainly centered in the applied general equilibrium analysis of global economic issues

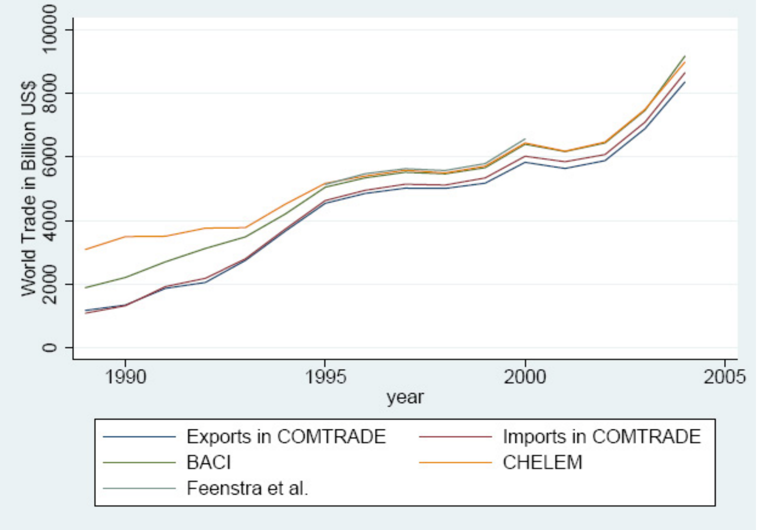
[21] COMTRADE contains information as far as 1962, but the 6-digit disaggregation starts in 1989. Similarly, available countries are increasing in time.

Table 2 – Comparison between International Trade Databases

	BACI0	COMTRADE	NBER-UN	CHELEM	GTAP
Period	1995-2004	idem[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.[2]
N Commodities	5041	5,041	1,276[3]	71	43

N.A.: Not Applicable. [1] The public BACI0 version is released for the 1995-2004 period but the HS 6-digit classification starts in 1989. COMTRADE provides with more datasets and years.[2] Codes are in letters.[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.

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



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 desegregation (72 exporting countries receiving imports from any country in the world at the 4-digit level of the SITC).

GTAP database combines, for a reference year (2001), detailed bilateral trade (also obtained from COMTRADE) with transport and protection data characterising economic linkages among regions, together with individual country input-output databases which account for inter-sector linkages within regions. In order to operationalize this large database, a standard framework has been developed. The components of this multi-region, applied general equilibrium model are relatively standard. GTAP is useful for the needs of computable general equilibrium models. However, trade flows are not reconciled: only one flow is selected to build the world trade matrix. The selection of this flow is done on the basis of a comparison of the reliability indices of the exporter and the importer for every individual flow.

CHELEM provides, at a world level, commodity trade flows, balances of payment, populations and incomes. In that line, CHELEM fills in the statistics provided by international institutions such as United Nations (UN), World Bank (WB) or International Monetary Fund (IMF), in terms of completeness and consistency. The bilateral flows series provided by CHELEM are provided on an annual basis expressed in millions U.S. dollars. Although CHELEM covers a longer time span (1967-2005) than BACI does, it displays less products (71 categories of commodities) and less countries (71 Countries/Regions corresponding to 99% of the world trade).[22] The CHELEM reconciliation of mirror flows proceeds also to a fobisation of import reports and to an average of taking into account the accuracy and the regularity of the declarations of the countries (de Saint-Vaulry, 2008).

The Figure 5 displays the evolution of the total world trade according to the mentioned databases. The evolutions are rather convergent. Note that BACI reaches in 1995 a total level of trade very close to CHELEM and the NBER database. The figure 6 in the Appendix on results provides a closer look at the 1995-2000 period, where NBER dataset is available. Although very similar, CHELEM, NBER and BACI datasets exhibit some differences. 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.* without removing CIF costs. In fact, the difference with BACI is around 2%, similar to the mean CIF estimated by BACI. The evolution of the recent years is depicted in figure 7 in the appendix. During this period, CHELEM and BACI converge even more, except in the last year, where BACI exhibit more trade data. Concerning the COMTRADE exports and imports, we see a stable gap of 10% in terms of value of trade for exports, and of 5% for imports.

[22] In CHELEM, only countries which represent an important weight in terms of international trade or population are individualized. The other countries are grouped in regional categories, mainly, on the basis of their proximity or their belongings to an economic zone of integration. Furthermore, the major interest of CHELEM is to provide a consistent view of the world economy via a specific harmonization process of the international trade data.

3.3. Some applications

BACI was largely used in the CEPII work to analyse trade patterns at the product-level, countries specialization, competitiveness, trade policy, exchange-rate pass-through, etc. Since its availability in 2007 to 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. We distinguish three types of cases for which BACI is particularly well suited: its product detail, its geographical exhaustivity and their unit-values.[23]

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. Disdier, Fontagné and Mimouni (2007) for example use BACI to analyse the impact of SPS and TBT agreements on agricultural trade; Fontagné, Laborde and Mitaritonna (2004) to study the impact of the EU-ACP Economic Partnership Agreements; Matthews and Gallezot (2006) regarding 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 using product-level data allow some analysis that more agregated data not??, since one can precisely identify some characteristics of product such as their main use in production (finals, intermediates or capital goods, Curran and Zignago, 2009, for instance), their technological content (World Bank, 2008; Mulder, Paillacar and Zignago, 2009; Cheptea, Fontagné and Zignago, 2009, among others) or any other particularity as for example the cultural dimension of goods (Disdier, Tai, Fontagné and Mayer, 2009). Finally, analysis of intra-industry trade at the world level could be done (Ecochard, Fontagné, Gaulier and Zignago, 2006; Fontagné, Freudenberg and Gaulier, 2006).

Secondly, BACI geographical exhaustivity allows to 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, 2009). But also the reorganisation of trade flows in Asia with the China's emergence (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 database (Fontagné et al. 2004).

Thirdly, BACI was 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. Several papers use BACI to confirm this trend 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 (2009). More generally, BACI is particularly useful when one want to analyse trade prices. Gaulier, Martin, Méjean and Zignago (2008) use it to propose *TradePrices*, an international trade prices indices. Gaulier and

[23] BACI is also an input to other CEPII databases like *TradeProd*, *TradePrices* and *MacMap-HS6*.

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 providing not only values but also consistent quantities, allowing to the analysis of international trade prices via the unit-values.

We describe in this working paper the methodology developed to build BACI. We estimate the CIF rate and subtracting it from the import reports. We turn then to the comparison between mirrors declarations and the computation of quality indicators of country reports. Under reasonable assumptions, our rather simple statistic methods – requiring no other input than raw trade statistics and "treated" traded quantities in the same way as values - can be applied to product data to provide consistent measures of international trade flows.

We have applied our methodology in first place to the HS0 datasets of COMTRADE. The resulting database (in its different versions, BACI0, the most known, but also BACI1 and BACISITC) is freely available since 2007 in our webpage to researchers having access to COMTRADE. BACI0 is now updated and covers the period 1994-2007. However, an important caveat must to be recall to users of BACI unit-values: A change in the United Nations treatment of quantities car affect the evolution of unit-values for some countries. Thanks to the UN collaboration, ongoing research is then 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). This increased disagregation is likely to reduce the agregation bias in the interpretation of unit values as prices, but it is not compatible at the international level. Unit values based on this tariff lines data and agregated at the 6-digit level will be available soon.

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*, forthcoming.

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 (2009), "Evolution of EU and its Member States Competitiveness", *CEPII Working Paper* forthcoming.

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*, forthcoming.

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.

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

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

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.

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.

WORLDBANK (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. METHODOLOGICAL APPENDIX

6.1. Allocation of “Areas Not Elsewhere Specified” (NES)

COMTRADE has some trade data without specification of destination or origin, classified as “Areas Not Elsewhere Specified” or “Areas NES”. BACI deals with these cases by conferring to this flows a new allocation when possible. The justification is that “Areas NES” flows may generate problems when flows are harmonized in BACI. In fact, BACI provides either weighted average of mirror declarations (from importing and exporting countries) when they exist, or weighted average of export declarations or of import declarations. Thus, a flow, which is declared as “Area NES”, could hide a flow towards a partner country itself declaring such that a double counting in the total trade may arise. In other words, the same flow is recorded as an export declaration and as an import declaration.

The allocation is made according to the weight of the partner countries that have declared the import of the commodity under consideration. Specifically, in the case of an exporting country i that declares “Area NES” flows for a given commodity in a given year, the BACI treatment is the following: if the sum of the flows towards partners countries (denoted by Vx) which have themselves declared is less than the sum of the mirrors declarations (denoted by Vm), then it is guessed that all (or a part of) the flows declared as “Area NES” (denoted by $Vnes$) are in fact devoted to these identified partners:

$$\sum_i Vx_{ij}^{kt} < \sum_i Vm_{ij}^{kt} \tag{3}$$

Therefore, BACI reallocates over the exporter declarations the minimum between on the one hand the difference between Vm and Vx and on the other hand, $Vnes$. Afterwards 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 (denoted by Vm').

$$Vnes_i'^{kt} < \sum_j Vm_{ij}'^{kt} \tag{4}$$

If $Vnes'$ is less than the sum of Vm'; 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 $Vnes'$.

Let us remark that in BACI such an incremental procedure of the country declaration – 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. In BACI, the implementation of the latter procedure, for each exporter and importer declaring "Area NES" flows results in rather limited modifications. About 11.5% of final flows are concerned by this treatement.

Besides of "Area NES" declaration from a given country, UN COMTRADE provides "Area NES" declaration towards a given group of countries such as "Asia NES". No treatment is done by BACI for such declarations. A double counting is then avoided in the sum of the harmonized values per countries. Note that the noise generates by this class of "NES" is of a weak extent since such "NES" declarations generally correspond to flows towards no declaring countries There also exists in UN COMTRADE a category "Commodities NES" in the commodities – "Commodities Not Elsewhere Specified" -. This category is dropped in BACI in order to avoid double counting due to the fact that partners may classify commodities in other category than "NES". In the case where there exist no mirror flows, even if the trade is underestimated, the extent of this underestimation is abstemious since the flows concerned mostly specific commodities such as military equipment or commodities for which no adequate category would has been found in the nomenclature.

6.2. Zero values and missing flows

Distinguish between missing values and zero trade flows is an important issue in all trade datasets. In COMTRADE, missing values may appear for many reasons. Globally, one reason that can explain the presence of missing values is the fact that a country do not necessarily have all others countries as partners and / or its trade flows may not cover all the product of the commodity classification used.[24] For an accurate accounting of the zero flows, it is important

[24] In addition, there are many countries which do not report their trade statistics (for a product really exchanged). For example, if we consider our data source COMTRADE, 54 countries, up today still not report their annual trade statistics (exports and imports). However, a noticeable improvement is in progress: while in 1989 only 26 countries do report their annual trade statistics (exports and imports), in 2003 138 countries do.

to distinguish whether a flow reported as missing is really a missing flow (that is we get no information) or if we are in the case of an absence of trade flow between partners. To tackle this issue, BACI uses the mirrors flows to determine whether a flow is really missing or if we are in presence of non declaration of one of the partner.

To discriminate between missing values and zeros, a reasonable assumption is consider that if a country report something for a given year, then this nation will declare all its bilateral flows: all partners, all products. Consequently, its flows may be either positive values or zero. BACI missing values corresponds to cases in where, for a given year, none of the partners report. This principle is applied for instance at the final stage of the BACI construction, to take into account the corrections generated by the methodology of allocation of "Area NES" flows, as explained in the previous section. A yearly matrix of country partners is provided in our webpage, indicating if we estimate the potential flows as missing or zeros.

6.3. More precisions about the harmonization procedure

Hypotheses on the error term. The *true* values V of trade are unobservables. For instance the value reported by the exporter country contains an error E_i, that we assume multiplicative and log-normal:

$$V_i = V * E_i, \qquad with \quad \epsilon_i = \ln E_i \sim N(0, \sigma_i^2) \tag{5}$$

The objective is to find the weights for the importer (w_j) and exporter(w_i). The minimization of σ_i^2 and σ_j^2 gives an optimal weight, as indicated in the section 2.4.1, and the measures of σ_i^2 and σ_j^2 are obtained by the Analysis of Variance as explained in section 2.4, using the fact that under our hypotheses, the mean of the Declaration Distances is:

$$\overline{DD_{ij}} = \sqrt{\frac{2}{\pi}}\sqrt{\sigma_i^2 + \sigma_j^2} \tag{6}$$

Quality of declaration Formula. The quality indicators of country declaration is obtained as follows:

The minimum is set to zero and only the precision of the estimation is taking into account (through the standard deviation):

$$if \quad LS\_\overline{DD_i} = \min(LS\_\overline{DD_i}), \;\; then \;\; dec_i = 2 \times stderr_i \times \frac{\pi}{2} \tag{7}$$

For the rest of cases:

$$if \quad LS\_\overline{DD_i} > \min(LS\_\overline{DD_i}), \quad then \;\; dec_i = (LS\_\overline{DD}_i + 2 \times stderr_i - \min(LS\_\overline{DD}_i)) \times \frac{\pi}{2} \quad (8)$$

The same procedure is done for the importer j.

7. Appendix on Results

Figure 6 – Evolution of Total World Trade 1995-2000.

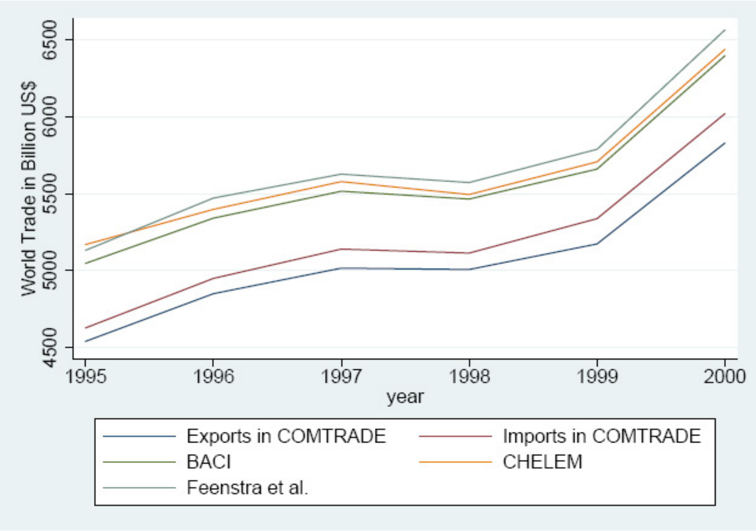
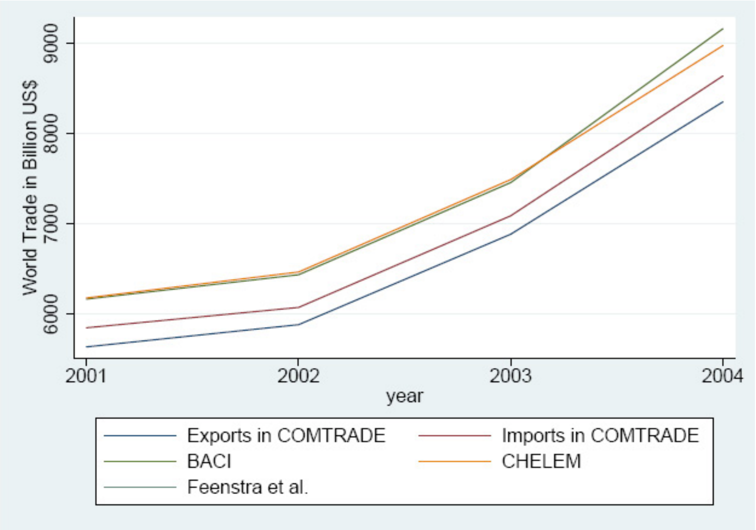


Figure 7 – Evolution of Total World Trade 2001-2004.



LIST OF WORKING PAPERS RELEASED BY CEPII