Notes on BACI (analytical database of international trade). 1989-2002 version

Gaulier, Guillaume and Zignago, Soledad

CEPII

2 October 2004

Notes on BACI

(Analytical Database of International Trade)

1989-2002 Version

Guillaume Gaulier & Soledad Zignago[•]

CEPII

02/10/2004

Working draft

Abstract

BACI draws on United Nations COMTRADE data and covers more than 200 countries and 5,000 products, between 1994 and 2004[1]). Imports and exports flows are reported annually by 130 countries to United Nations in values and quantities. When both exporting and importing countries report, we have two figures for the same flow, which have to be harmonised given the huge discrepancies between them: at the 6-digit level of the Harmonised System, the gap between mirror declarations exceeds 100 % for half of the observations in COMTRADE. Original procedures are developed aiming at providing the more disaggregated and rigorous trade database for the largest possible number of countries and years, with a special care in the treatment of unit values.

1. Introduction

International trade statistics can only be used to carry out detailed studies related to recent developments in economic theory, at the cost of extensive, fastidious work on treating data from

[•] We thank the assistance from Sessi Tokpavi.

[1] The declarations in HS appear in 1989, but the current version of BACI reaches a very broad world coverage in 1994 (or 1995 if one is interested in the declarations of some important countries like France or Belgium. More exactly BACI in HS from 1992 covers the period 1994-2004 and BACI in the HS from 1996 the period 1996-2004.}.

numerous, heterogeneous sources. To meet these difficulties, we have constructed a database on international trade which brings together and renders consistent various levels of analyses and classifications, drawing on the most detailed information available. In doing so, we continue the work done in Gaulier and Zignago (2002), with the aim to provide researchers with a new database allowing a more detailed description of trends in world trade than is presently the case.[2] A particular goal is to put forward a characterisation of trade flows in terms of trade types (one-way trade, cross-trade in similar products, cross-trade in vertically differentiated products), product ranges, technological levels and stages of production, etc. BACI permits also the calculation of price (unit values) indices. We present here the treatment of a first version of this database that covers the period 1989-2002 for all countries declaring their annual international trade statistics to the United Nations (COMTRADE Database) and made available to CEPII researchers.

Trade flows are reported in value and quantity by both exporting and importing country (mirror flows, when available). We have developed original procedures to harmonise COMTRADE data. **Two major steps are used for the treatment :** The first consists in taking a look at the source data and prepare the database for harmonisation. Along this step, we pull out useful information like list of reporting countries, list of partner countries. In the second step - the most important - we harmonise COMTRADE data. For doing so, we make an evaluation of CIF rates to remove freight costs from the import declarations which are declared C.I.F (Cost Insurance and freight) and an evaluation of the quality of country declarations to average mirror flows.

2. Methodology

2.1. Source: COMTRADE Database

Every year over 130 countries provide the United Nations Statistics Division with their annual international trade statistics, detailed by commodity and partner country. These data are processed into a standard format with consistent coding and valuation. All values are converted into US dollars using exchange rates supplied by the countries, or derived from monthly market rates and volume of trade. Quantities are, if provided by the country and if possible, converted into metric units. For many countries the data coverage starts as far back as 1962 and goes up to the most recent completed year. Commodities are classified according to SITC (Rev.1

[2] Similar work, but in at a more aggregated level, is done for the CHELEM international trade data by De Vaulry (2008).

from1962,Rev.2 from 1976 and Rev.3 from 1988) and the Harmonised System (HS) (from 1988 with revisions in 1996 and 2002). Currently most data are reported according to HS, version 2002[3].

Preparing the COMTRADE data for harmonisation, many transformations are made: We first make conversion in tonnes of the other units of quantities exchanged. In fact, 86% of quantities are declared in tonnes. The other quantities are converted into weights by estimating for each 6-digit product a rate of conversion between each unit (units, watt, meter, etc.) and tonne, using flows reported in heterogeneous units. We also suppressed quantities declared in unknown units. In order to have a database with only one commodity classification (HS 1988), we use Correspondence Table between the Harmonised System, version 1988 and the Harmonised System, version 2002. Thus, after complementary matrix operations (like transposition) we generate a new database for harmonisation. The following table presents all variables available in this database and retained for BACI' s construction.

Table 1. Example of COMTRADE data.

hs6	**i**	**j**	**t**	**vx**	**qx**	**vm**	**qm**	**ux**	**um**
711100	United Kingdom	Japan	1997	11	.	.	.		
760110	Indonesia	Hong Kong	1998	89	77.00	98	76	W	W
961590	Ireland	Australia	1994	.	.	35	.		

Where i and j are respectively exporter and importer countries, hs6 a level of commodity classification (HS0, named oftenly HS1988 or HS2002), t a year (between 1989 and 2002), vx value reported by i (qx and ux respectively quantity and unity corresponding), vm value reported by j (qm and um respectively quantity and unity corresponding). 42 countries, up today do not report their annual trade statistics (export and import)[4]. However, there is a real progress: in 1989 only 24 countries reported their annual trade statistics (export and import) against 108 in 2002 as one can see in the Table 2 [5].

Table 2. List of non reporting countries (import & export / 1989-2002)

Afghanistan
Angola
Anguilla
Antilles néerlandaises

Irak
Jamahiriya arabe libyenne
Koweït
Lao, Rép. Dém. Pop.

[3] The original data are also converted and stored in all the other classifications. For the current version of BACI, the source data is classified in HS from 1988 and 1996 and not includes flows below 1,000 dollars.
For more details on COMTRADE see http://unstats.un.org/unsd/comtrade/.

[4] For the list, see table 2

[5] For more information (data for all years) report to appendix

Aruba
Cambodge
Congo, Rép. dém. Du
Corée, Rép. pop. Dém. De
Djibouti
Gibraltar
Guinée équatoriale
Guinée-Bissau
Île Christmas
Île Norfolk
Îles Caïmans
Îles Cocos (Keeling)
Îles Falkland
Îles mineures éloignées des Etats-Unis
Îles Salomon
Îles Turks et Caïques
Îles Vierges britanniques
Libéria
Mauritanie
Mozambique
Nauru
Nioué
Pakistan
Pitcairn
Sainte-Hélène
Saint-Pierre-et-Miquelon
Samoa
Ship Stores and Bunkers
Sierra Leone
Somalie
Taïwan
Territoire britannique de l`Océan Indien
Timor Oriental
Tokelaou

Table 3. List of reporting countries (1989 & 2002)

1989

Australie
Bangladesh
Brésil
Canada
Chypre
Corée
Danemark
Espagne
Finlande
Grèce
Inde
Indonésie
Islande
Japon
Malaisie
Nouvelle-Zélande
Grenadines
Oman
Paraguay
Portugal
Roumanie
Singapour
Suisse
Thaïlande
Turquie

2002

Afique du Sud seule
Albanie
Allemagne
Andorre
Arabie saoudite
Argentine
Arménie
Australie
Autriche
Azerbaïdjan
Barbade
Bélarus
Belgique
Belize
Bolivie
Brésil
Brunéi Darussalam
Canada
Chili
Chine
Chypre
Colombie
Corée
Costa Rica
Croatie
Danemark
Dominique
El Salvador
Equateur
Espagne
Estonie
Etats-Unis
Ethiopie seul
Fédération de Russie
Fidji
Finlande
France
Grèce
Grenade
Guatemala
Guinée
Guyana
Honduras
Hong-Kong
Hongrie
Inde
Indonésie
Iran
Irlande
Islande
Israël
Italie
Jamaïque
Japon
Jordanie
Kenya
Kirghizistan
Lettonie
Lituanie
Luxembourg
Macao
Malaisie
Maldives
Maroc
Maurice
Mexique
Moldova, Rép. de
Nicaragua
Norvège
Nouvelle-Zélande
Oman
Ouganda
Panama
Paraguay
Pays-Bas
Pérou
Philippines
Pologne
Polynésie française
Portugal
Qatar
République arabe syrienne
République tchèque
Roumanie
Royaume-Uni
Rwanda
Sainte-Lucie
Saint-Vincent-et-les
Sao Tomé-et-Principe
Sénégal
Seychelles
Singapour
Slovaquie
Slovénie
Soudan
Sri Lanka
Suède
Suisse
Swaziland
Togo
Trinité-et-Tobago
Tunisie
Turquie
Ukraine
Uruguay
Venezuela
Zambie
Zimbabwe

Given the huge discrepancies between reported mirror flows, trade data have to be harmonised[6]. For doing so, we successively make an evaluation of CIF rates to remove freight costs from import declarations, evaluation of the quality of country declarations to average mirror flows.

2.2. Evaluation of CIF rates to remove freight costs from import declarations

In COMTRADE, import values are reported C.I.F. (cost, insurance and freight) and the exports are reported F.O.B. (free on board). In order to remove C.I.F., we have to estimate freight costs. Being plagued with large measurement errors, mirror flows ratios can not be directly identified with freight costs. However, we use predicted mirror flows ratios from the following gravity-type equation as estimates of C.I.F.:

$$\begin{aligned}
&\ln(\hat{CIF}_{ijk}) = \ln(VM / VX) = a + b.\ln dist_{ij} + c.\ln dist_{ij}^{\,2} + d.\ln UV_k + e.contiguity_{ij} + f.landlocked_j \\
&+ g\cdot t1989 + h\cdot t1990 + i\cdot t1991 + + j\cdot t1992 + k\cdot t1993 + l\cdot t1994 + m\cdot t1995 + n\cdot t1996 + o\cdot t1997 \\
&+ p\cdot t1998 + q\cdot t1999 + r\cdot t2000 + s\cdot t2001 + \varepsilon
\end{aligned}$$

Where i and j countries dimensions, respectively for exporter and importer, and k is product dimension. Each observation used for the estimation, combined these three dimensions. distij is geographical distance. This geographical variable is taken from CEPII's distances measures database (Mayer and Zignago, 2006)[7]. UV_k is unit value (value/quantity), which is a world-median for each 6-digit product (no country dimension). Contiguityij and landlockedj are dummies variables; they are used to capture the fact that the C.I.F should decrease if the exporter and the importer countries are contiguous (for the first) and increase (for the second) if the importer country is landlocked. We also introduce temporal dummies variables; the idea is to consider an eventual temporal evolution of CIF[8]. The equation is estimated by OLS on pooled data. As we observe a strong positive relation between ratio of mirror flows for reported values and those for reported quantities (discrepancies are likely to be observed simultaneously for values and quantities) we weight observations in the equation for implicit C.I.F. by the inverse of the gap between reported mirror quantities (Min(QXij,QMji)/ Max(QXij,QMji)). This gives the higher

[6] Let's remind that harmonisation concern 38% of observations (those for which both mirror flows exist).

[7] 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).

[8] We don't keep t2002 for the estimation; the principal reason is to avoid an evident problem of multicollinearity. Thus, t2002 is the reference and the estimates coefficients of other temporal dummies can be interpreted as gap between each of them and the reference.

weight to trade flows equally reported by partners, differences between reported import and export values are then more likely to be freight costs.

There are 9,944,957 observations available for the estimation. In order to obtain consistent and robust parameter estimates, we used a statistical mopping-up operation that help us to remove 345,879 atypical and influential observations[9]. After this operation, changes of estimates coefficients are insignificant (*ie* coefficients are enough stable). The results of the estimation are shown in the Table 5.

Table 5. Results of the estimation of freight costs.

Variables	Parameter estimates
Intercept	0.16417***
Ln $dist_{ij}$	-0.07500***
Ln $dist_{ij}^2$	0.00781***
Ln UV_k	-0.02615***
$Contiguity_{ij}$	-0.03508***
$Landlocked_j$	0.04588***
T1989	0.01471***
T1990	0.01054***
T1991	0.02516***
T1992	0.02230***
T1993	0.00414***
T1994	0.00156*
T1995	0.00235**
T1996	-0.00482***
T1997	-0.00273***
T1998	0.00182*
T1999	0.00255***
T2000	-0.00577***
T2001	0.00789***
*** Significance level > 99%	
** Significance level > 95%	
* Significance level > 85%	

All variables are pertinent, with a significance level above than 95%, except *t1994*. The estimated coefficients for respectively distance and unit value imply that CIF increases with distance and decrease with unit value. We obtain the expected sign for respectively *contiguityij* and *landlockedj*. The estimated values of temporal dummies do not show a uniform evolution, some appearing with a positive sign and others with a negative one. But the most important tendency is the

[9] To identify those observations, we compute the D distance of Cook (1977) and the measurement of Student Residuals.

decrease of C.I.F. considering the period of study, as shown by the sign of variable *t1989*; between 1989 and 2002, the logarithmic value of C.I.F. decrease on average by 1.5%, which imply a drop in freight costs in the course of time. If we remove the unessential quadratic term for distance we get an elasticity of (implicit) C.I.F. with regard to distance of 4.9%.

However, the mean of the estimated value of our endogenous variable is too low (0.01) comparing with what is generally admitted (a world possible mean would be 0.12[10]). This result is not amazing and does not raise doubts about the relevance of our model. In fact, Hummels and Lugovsky (2006), further to Yeats (1978) investigation[11] found that the matched partner cif/fob data strongly co-vary with direct measures of shipping costs despite being systematically wrong in levels. Accepting those explanations and in order to reach a more consistent level, we apply the following transformation:

$$\ln(C\hat{I}F_{ijk}) = 0.12 + \ln(C\hat{I}F_{ijk}) - \overline{\ln(C\hat{I}F_{ijk})}$$

Where: $\overline{\ln(C\hat{I}F_{ijk})}$ is the mean of freight costs estimated through our gravity-type equation. And these new values are those finally retained to estimate freight costs. As the Table 5 shows, our estimation of freight costs are very similar to those of Hummels (2001). The results presented in the following table are the variability of C.I.F with regard to distance, when the importer country is more and more distant from the exporter.

[10] See Anderson & van Wincoop (2004).

[11] Yeats provides an early examination of the quality of matched partner data by comparing cif/fob ratios constructed from UN COMTRADE data to shipping cost data collected from US imports in 1974. His analysis consists of decomposing observed variation in matched partner cif/fob ratios into a part due to shipping cost (signal) and a remaining unexplained part (noise). The main result is that matched partner cif/fob data contains significant amount of noise which make its level very different from the direct measures. And the difference increases with aggregation.

Table 6. BACI and Hummels Estimates of freight costs

Dist (km)	Hummels Estimates	Estimated CAF
	Low UV	
100	19%	13%
300	25%	13%
1000	34%	16%
2500	44%	20%
5000	53%	25%
10000	64%	32%
14000	70%	36%
20000	77%	40%
	High UV	
100	5%	1%
300	7%	1%
1000	10%	3%
2500	13%	8%
5000	15%	12%
10000	19%	18%
14000	20%	22%
20000	22%	26%
	Average UV	
100	9%	6%
300	12%	5%
1000	16%	8%
2500	21%	13%
5000	25%	17%
10000	30%	24%
14000	33%	27%
20000	37%	31%

2.3. Evaluation of the quality of country declarations to average mirror flows

In this stage, we calculate indicators of quality of import and export declaration for each country, which are used, in the last stage, as weights when averaging the mirror flows to get the harmonised flow. It exists in the concerned literature several techniques to evaluate gaps between mirror flows, or more exactly quality of declaration of a given country. For example, the one used by the International trade Centre (ITC) consists for a given country in measuring the quality of declaration by the mean of the gaps of mirror flows (export or import) and this for all partners, products and years. The gaps are pooled by a factor, which reflect their respective importance at a world-wide level.[12] This technique is debatable, because, a given country which would record high levels of the gaps of mirror flows can also be a good reporting country, the gap in that case, attributable to its partners. However, this reasoning remains, the procedure proposed here to evaluate the quality of country reports.

[12] Of course, the relative relevance of this procedure depends on the idea that the country has a reasonable number of partners, gaps in that case, cannot be automatically attributable to its partners, seeing their diversity, a part of the responsibility certainly resting with it..

The main idea is that, the more a country is a bad reporting country, the more its distribution of the gap of mirror flows is distant to a theoretical reference distribution, supposed "ideal"**.** Nevertheless**,** we take a precaution, considering the previous critic. The matter is to be sure, that the gap between mirror flows is entirely attributable to the country being evaluated, ensuring that the concerned flows, are shipped mainly towards good reporting countries (*ie* countries which have in mean, low level of gaps, and - this restriction is very important- have a sufficient number of partners).

We use a measure of distance, inspired from Kullback, between the distribution of the ratios of mirror flows (log of reported export from *i* to *j* on reported import of *j* from *i)* of this country with the reference distribution of these ratios for all exporters (and symmetrically for the quality of import declarations). For a given countries we consider the distribution of all declarations to any partner, for any product or year. For the reference (world) distribution, all observations available are pooled (export, import, product and year). Figure 1 illustrates this procedure. The Kullback-Leibner Distance formula used is the following:

$$KLD = \int_{-\infty}^{+\infty} p(x) \ln\left(\frac{p(x)}{\overline{p}(x)}\right) dx$$

$$KLD \approx \sum_{h} P(h) \ln\left(\frac{P(h)}{\overline{P}(h)}\right)$$

where $P(h) = freq\left[h < \ln\left(\frac{X_{i,j}}{M_{j,i}}\right) < h + 0.1\right]$

X_{ij} is the value (respectively quantity) reported by the exporter M_{ij} , the value (respectively quantity) reported by the importer, and *p(.)* is density function or mirror flow ratio for the country being evaluated and ?? is the density for the reference, a measure of the world mean of discrepancies between values (respectively quantity) reported by exporter and importer. Thus, the more the distribution *p(.)* is atypical for a country, the more the difference between *p(.)* and the corresponding world distribution is important and the higher is *KLD*. For example, in the figure 1, the quality of export declaration might be higher for Nigeria than Australia, because ratio of mirror flows distribution for Nigeria is more different from the reference than the Australia one's.

The database is first resampled in order to remove geographical bias: if a country export only to good reporters, it will appear itself as a good reporter. The resampling consists, for each exporter, in modifying the frequencies of each partner in order to have a distribution of partner the closest as possible as the world distribution of trade flows. If 1% of the export flows from Albania are oriented toward the US and 80% to Germany, flows to the US will be duplicated, on the contrary only a subset of the flows to Germany will be (randomly) selected in the final sample. Figure 2 illustrates this procedure for the Tunisian total imports. The geographical distribution of the number of flows is corrected to match to the world distribution of trade flows (in frequency).

Figure 1. Ratios of Mirror Flows Distribution for 3 Exporters & Reference Distribution

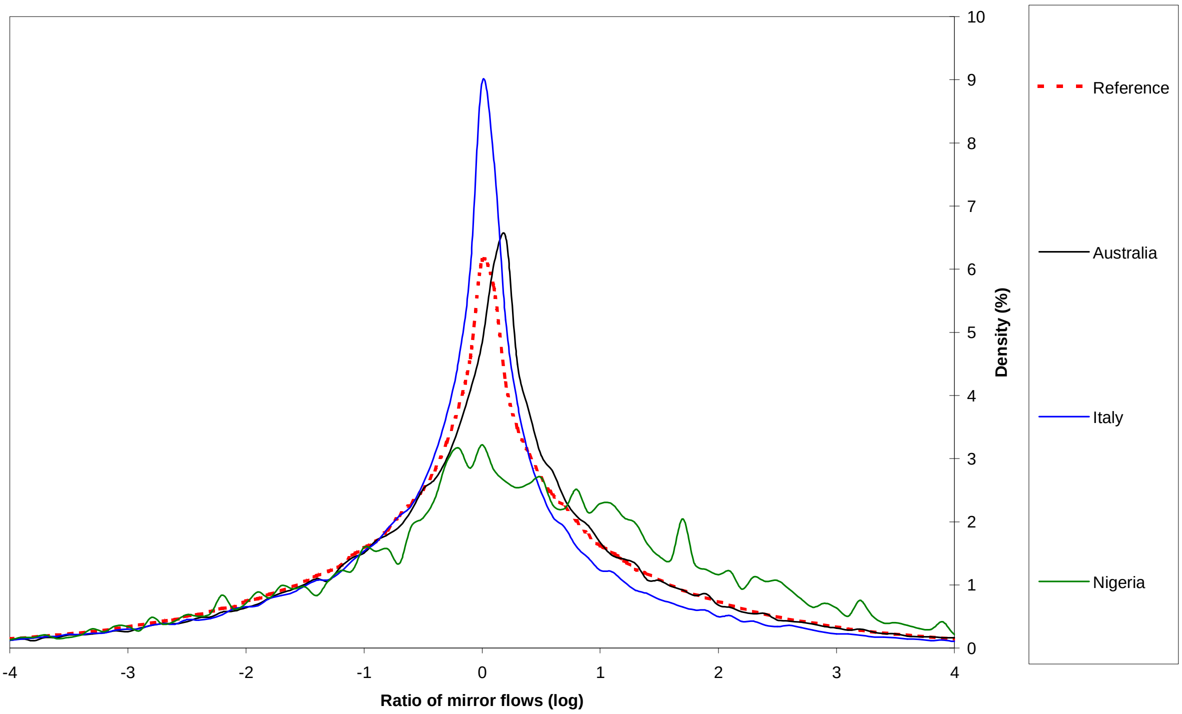

Figure 2 (a). Geographical Frequency of Tunisian Imports (1995-2001).

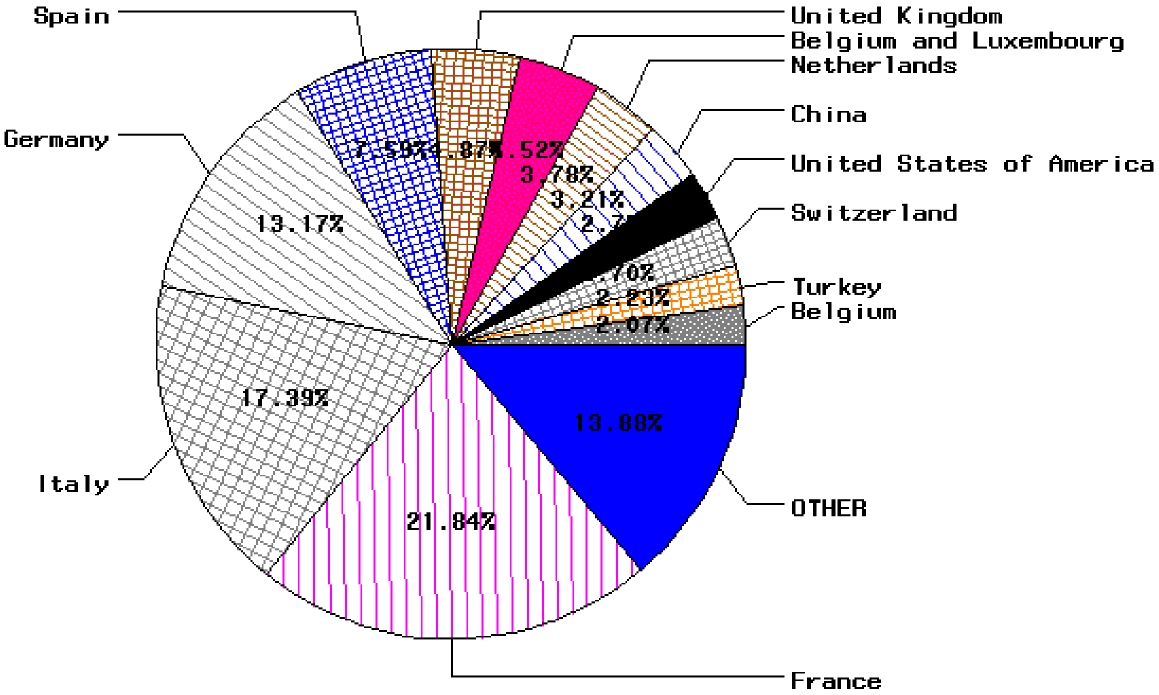

Figure 2 (b). World Distribution of Trade Flows Frequencies (1995-2001).

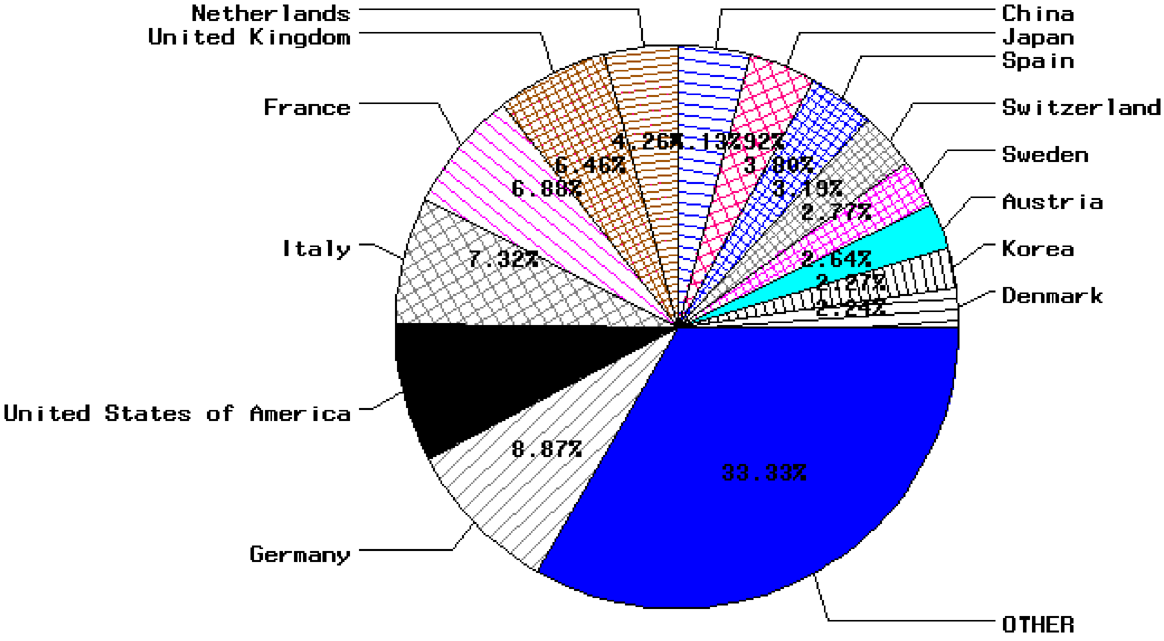

Table 5 shows the correlation matrix of the 4 different types of quality presented in Table 5. All correlations are important and the higher (correlation between quality for export declaration for value and quality for export declaration for quantity) imply that an exporter, good reporting for values declaration is also good for quantity declaration. And the countries concerned are in the majority Latin American one's, some PECOS and emerging countries (*report to Table 6*) and one can see visibly the high rank correlation between rank1 and rank2.

Table 7. Matrix of correlations

	Qual_exp_value	Qual_imp_value	Qual_exp_quant	Qual_imp_quant
Qual_exp_value	1			
Qual_imp_value	0.766	1		
Qual_exp_quant	0.922	0.777	1	
Qual_imp_quant	0.762	0.884	0.818	1

Note: Qual_exp_value = quality for export declaration for value
Qual_imp_value = quality for import declaration for value
Qual_exp_quant = quality for export declaration for quantity
Qual_imp_quant = quality for import declaration for quantity

Table 8. Best and worse reporting countries (for export declaration, value & quantity)

Good reporting countries	qual_exp_value	Rank1	qual_exp_quant	Rank2	Bad reporting countries	qual_exp_value	Rank1	qual_exp_quant	Rank2
Argentine	-8.5399	4	-9.6961	3	Bahamas	70.13	143	84.128	139
Brésil	-6.6095	12	-7.4677	9	Bhoutan	246.08	164	83.369	138
Bulgarie	-5.9251	14	-7.4019	10	Botswana	143.638	154	93.897	142
Chili	-10.5298	1	-10.2634	1	Burkina Faso	241.585	163	248.832	161
Colombie	-8.6607	3	-10.2128	2	Burundi	73.953	144	102.282	146
Croatie	-8.5137	5	-8.9361	5	Cap-Vert	82.258	145	65.218	130
Chypre	-5.8825	15	-6.7202	14	Comores	126.878	151	123.41	151
République tchèque	-6.6639	10	-7.2544	13	Emirats arabes unis	61.807	138	88.734	140
El Salvador	-3.7163	20	-6.6888	15	Gambie	113.745	150	103.34	147
Estonie	-3.6939	21	-2.7881	26	Groenland	67.84	141	67.867	131
Finlande	-4.8972	18	-7.2636	12	Guyane française	176.577	158	187.539	157
Grèce	-2.428	28	-2.7567	27	Haïti	69.951	142	40.602	118
Hongrie	-7.7819	6	-8.866	6	Îles Cook	232.822	162	1.587	52
Islande	-6.8856	9	-7.2688	11	Kiribati	153.295	156	329.304	164
Lituanie	-3.5872	22	-6.2079	17	Mali	212.63	160	226.793	160
Maroc	-4.5294	19	-4.2037	21	Martinique	100.968	147	94.568	143
Nouvelle-Zélande	-2.8796	27	-3.1848	23	Montserrat	417.525	167	.	.
Pérou	-6.6156	11	-6.2753	16	Myanmar	157.676	157	166.249	155
Portugal	-2.3684	29	-2.7533	28	Niger	213.336	161	91.306	141
Roumanie	-5.719	16	-5.5478	19	Nigéria	64.144	139	79.466	137
Slovaquie	-6.0736	13	-4.5377	20	Papouasie-Nouvelle-Guinée	109.041	148	200.434	158
Slovénie	-8.8797	2	-8.3037	7	Qatar	66.271	140	74.043	134
Turquie	-3.4543	23	-3.7631	22	République centrafricaine	139.371	153	165.232	154
L`ex-Rép. Yougoslave					Rwanda	258.624	165	254.844	162
de Macédoine	-7.2056	7	-7.8966	8	Saint-Kitts-et-Nevis	130.151	152	181.624	156
Uruguay	-7.0103	8	-9.1826	4	Saint-Vincent-et-les Grenadines	86.752	146	104.395	148

Concerning the bad reporting countries (export), there is no surprise. The concerned countries are in the majority (or exclusively) South countries (report to Table 8). About the quality of import declaration the lists of best reporting countries and bad one's do not basically differ from these, because the correlation between quality of export declaration and quality of import declaration is not insignificant (0.766 for value and 0.818 for quantity). However, it is to be noted the apparition among the best reporting countries, some developed countries like Italy followed by Switzerland and the majority of industrialised countries. Some emerging and developing countries get good ranking, in particular Latin American as well as East-European countries. Import and export quality indicators are transformed in order to sum to 1 and be used as weights. For each bilateral trade flow we use those weights to compute an harmonised flow.

The last stage consists in taking the two values of the same flow (the value reported by exporter and the one reported by importer without freight costs) with levels not basically different and to generate a new one. The new value is the mean of those two values, pooled by a factor function of the quality of declaration.

3. References

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

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

DE SAINT-VAULRY A. (2008), "Base de données CHELEM – commerce international du CEPII", CEPII Working Paper 09. 25

GAULIER G. AND S. ZIGNAGO (2002), "La discrimination commerciale revelee comme mesure desagregee de l'acces aux marches," Economie Internationale, CEPII research center, issue 1Q-2Q, pages 262-280.

HUMMEL, S. D. (2001), "Toward a Geography of Trade Costs", GTAP 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.

MAYER, T. and ZIGNAGO, S. (2006), "Notes on CEPII's distance measures", MPRA Paper 26469.

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

YEATS, A. (1978).