

# TradeProd. The CEPII Trade, Production and Bilateral Protection Database: Explanatory Notes

Thierry Mayer <sup>\*</sup>    Rodrigo Paillacar <sup>†</sup>    Soledad Zignago <sup>‡</sup>

August 4, 2008

**Abstract.** This dataset provides an updated version of the worldwide data used in Mayer and Zignago (2005), and available on CEPII's website. The database contains three groups of information covering 28 industrial sectors in the ISIC (International Standard Industrial Classification) classification Revision 2. The first group is bilateral trade for the 1980-2004 period, based on BACI, one of the most exhaustive worldwide dataset publicly available. The second is an extension of industrial production figures from the *Trade, Production and Protection* database by Alessandro Nicita and Marcelo Olarreaga (World Bank). The third group is bilateral data on trade policy at the industry level (tariffs and NTBs) taken from TRAINS and from CEPII's MAcMap Project, which both cover the 1989-2001 period.

*Keywords:* Database, Trade, National Production, Tariffs, Non-tariff barriers (NTBs).

*JEL classification:* F10, F14, F13, C80.

---

<sup>\*</sup>Université de Paris 1, Paris School of Economics, CEPII and CEPR.

<sup>†</sup>Université de Paris 1. Corresponding author. (rodrigo.paillacar@cepii.fr)

<sup>‡</sup>CEPII.

# 1 Introduction

In recent years, research in international economics has benefited from detailed and consistent data on bilateral trade volumes, industrial production and tariffs. This note provides information about the *CEPII Trade, Production and Bilateral Protection Database* (Henceforth, *TradeProd*), which updates a previous version initially developed for the article by Mayer and Zignago (2005). The result is a database that expands figures in production, provides bilateral trade based on a new and highly disaggregated dataset, and adds bilateral data on trade policy (tariffs and NTB) at the industry level. Information is disaggregated at the ISIC rev2 3-digit industry level (28 industrial sectors) over the period 1980-2004 for production and bilateral trade, and 1989-2001 for bilateral protection data. This note provides details about the characteristics of the information, sources, methods and a comparison with a similar dataset provided by the World Bank: *Trade, Production and Protection* database put together by Alessandro Nicita and Marcelo Olarreaga (Nicita and Olarreaga, 2006, henceforth N&O 2006<sup>1</sup>).

## 2 Sources and Procedures

This dataset is presented in four files. The first (*Prod\_cepïi8004.dta*) contains information at national level for production, wages, number of workers, etc. Two files provide bilateral trade volumes for a period of 25 years (*Trade\_cepïi8004.dta*) and 10 years (*Trade\_cepïi9504.dta*), respectively. The last file delivers several bilateral trade policy variables (*Tar\_cepïi8901.dta*). Figure 1 summarizes the different sources employed.

### 2.1 Trade Data

Trade data is based on a new database developed by CEPII, called BACI (for *Base pour l'Analyse du Commerce International*), which is built using COMTRADE, from the United Nations Statistical Department, as a primary source. The advantage of BACI is the use of mirror flows (harmonized to warrant consistency), which increases the coverage of the trade data, specially for developing countries.<sup>2</sup>

---

<sup>1</sup>Also in Nicita and Olarreaga (2007).

<sup>2</sup>Although it is not the objective here to develop in detail the methodology used to build BACI (see Gaulier and Zignago, 2008), two important aspects of the methodology deserve mentioning. The first consideration is the use of mirror flows. When both exporting and importing countries report to COMTRADE, researchers usually consider only imports, or average both values. While in the first case a lot of information is discarded, in the second the CIF component (cost, insurance and freight) is not properly removed. BACI develops a methodology estimating a mean CIF ratio, in order to obtain an estimation of FOB values, which is applied to every flow. The second consideration is the weighting criteria for mirror values, based on an estimation of the “quality of declaration” of reporting countries. Consequently, the weights can differ from a 50/50 average.

We use correspondence tables between the international trade nomenclatures and the ISIC Rev. 2 categories. Like Nicita & Olarreaga, a version covering a long period is available, by extracting COMTRADE data in SITC Rev. 2 categories. This version is in the file *Trade\_ceprii8004.dta*, and goes back until 1980, with trade for 222 countries/territories.<sup>3</sup>

Although using SITC Rev 2 certainly offers the best trade-off between time coverage and product aggregation, a better correspondence between traded products and industrial sectors can be obtained if the Harmonized System (HS) classification is used instead of the SITC. This can be done at the cost of a shorter coverage over time.<sup>4</sup>

A great number of the flows exhibit figures similar to those of N&O 2006. As the 1995-2004 version offers more differences in terms of methodology, we present some descriptive statistics for that dataset.<sup>5</sup> The mean value of the difference between World Bank and *TradeProd* datasets is 905,000 USD. Among the 1,843,160 non-zero, non-missing flows common to both datasets, 71.6% have differences that are below 20% of the value or smaller than 100,000 USD (57.6% with a threshold of 1,000 USD). Nonetheless, these moderate differences (and the remaining 16% of large ones) follow a clear pattern. Table 1 exhibits data for *TradeProd* at the industry level in the first column. The second column contains the same information from the World Bank dataset. The third column shows the differences in terms of total trade by industry (after discarding destinations that are not countries). The last column corresponds to differences without considering China, which exhibits extremely high values in trade flows in the World Bank database, specially with the US and Hong Kong. It should be noted that China is classified as a relatively less reliable reporter in BACI (because of systematic and wide divergences with a great number of trade partners), which entails less weight for its figures. We see important differences in the value of trade, with most of the industries showing higher values for *TradeProd* (the total difference in terms of values is more than USD 3,000 billions for the 10 years considered). However, 7 out of 28 industries exhibit a lower total trade. A closer look at the individual flows, reveals that they are in great part driven by some flows where China is the exporter or importer. When this country is not taken into account in the comparisons, the gap is reduced for all these industries except one (354, *Other Chemicals*).

A second important aspect is the number of flows. Table 2 shows the distribution by continent of exporter and importer of the flows that are positive

---

<sup>3</sup>A previous version of this dataset was already used to complete figures in the first version of *TradeProd*. In the new version, BACI is the only source.

<sup>4</sup>HS goes back to 1988, but given the delay in the implementation of this nomenclature by reporting countries, information provided by BACI is available starting in 1995 (Gaulier and Zignago, 2008).

<sup>5</sup>It is important to mention that a direct comparison between both databases is not an easy task, because BACI received a treatment not only in terms of reconciliation of COMTRADE mirror flows, and different product categories from COMTRADE (HS vs SITC), but also in terms of eliminating re-exports and all destinations that are not countries (i.e. Free zones, etc., defined with the codes FRE, SPE and UNS in the N&O 2006 dataset, and corresponding to 46,892 flows for the 1995-2004 period).

in *TradeProd*, but zero or missing in N&O 2006. However, these 690,000 flows should not be considered as a net gain, because there are 164,017 flows that are considered as zero or missing in *TradeProd*, but are positive in the World Bank data. In Table 2 we see that these new flows in *TradeProd* are concentrated in exports from Europe to African, Asian and American countries, and in “intra” Asian trade.

## 2.2 Production data

Information for the Industrial Production and related variables (number of firms, value added, number of employees and wages) are taken from United Nations Industrial Development Organization (UNIDO).<sup>6</sup> This source provides information at the three and four digits levels. Data at the 3-digit level uses ISIC Rev. 2, allowing for a direct merge with bilateral trade, and covering 1980-2003. In the case of 4-digit data, it is available from 1985 to 2004, but in ISIC Rev. 3 only, so a correspondence table is employed.<sup>7</sup> Additionally, we took advantage of information from OECD STAN Industry Database. Production data was converted to indexes, and used to fill some missing data.<sup>8</sup> The total number of countries with some information is 182, and with some information on production it is 170.

As these data are very similar to that provided in N&O 2006, we consider important to mention two differences in our methods. The first issue is the treatment for Belgium and Luxembourg data. World Bank present their information aggregated under the code BLX (Belgium-Luxembourg), because trade flows are reported at this level. The problem is that aggregating data can affect the national information and produce important distortions. Table 3 shows figures for the World Bank source (BLX) and from the original sources for Belgium (BEL) and Luxembourg (LUX) for a specific industry (ISIC 311) in 1993. Only the number of firms and value added are actually the sum of available data from both countries. Production and the number of employees correspond to Luxembourg, while wage bill corresponds to Belgium. This produces two problems. First, internal flows (production minus total exports) are negative, because exports (should) represent both countries. Second, wage per employee (which results from the ratio of wage bill over the number of employees) is over-estimated. We opted to keep both groups of data (BLX from N&O 2006 and BEL and LUX data from UNIDO), and let the researcher chose according to her purposes.<sup>9</sup>

---

<sup>6</sup>In the previous version of *TradeProd*, data was taken initially from the World Bank (N&O 2006), which is in turn originated from UNIDO. N&O 2006 data, available in the file *TPP\_rev1.csv* cover until 2003. Recently, a new release from UNIDO information has been available, with data covering years until 2004.

<sup>7</sup>We checked data to ensure consistency. The correlation between both production data is 0.995.

<sup>8</sup>The correlation between both production data sources is 0.985. The gain in data availability in this case is relatively small: 1,037 missing values are filled by this method.

<sup>9</sup>To obtain an estimation of internal flows for Belgium and Luxembourg, we allocated total imports and exports according to the GDPs. We are aware that this is not an ideal solution.

A second issue is data from Germany for the reunification period. Information available for Eastern Germany (DDR) ends in 1989, and for Unified Germany (GER or DEU according to the codification) starts in 1991. Original figures from UNIDO allow to add data from both countries for all the period except 1990 (there are only figures for Western Germany). A similar methodology was followed by Nicita and Olarreaga (2001), but in the recent version, they opted to present under the same code (DEU) data for Western Germany until 1990, and for the reunified Germany since then. We preferred and implemented the previous methodology.

Tables 4 and 5 exhibit the data availability in time span and industry coverage, respectively. Differences with former version of *TradeProd* and with N&O 2006 come from the fact that (1) we considered mirror total exports from BACI (*cf.* previous section) to generate the total exports variable (*expw*) to complete the figures of trade with self<sup>10</sup> and (2) we included additional data from UNIDO and OECD, in the case of number of firms (*firms*), number of employees (*lab*), wages per employee (*wage*), national production (*prod*) and value added (*va*).<sup>11</sup>

## 2.3 Bilateral Trade Policy

The third group of variables provided in the dataset concerns bilateral trade policy. They include tariffs as well as several indicators of Non-tariff Barriers (NTBs). N&O 2006 also provide data on protection (mainly coming from the same original source than *TradeProd*<sup>12</sup>) at the country level.<sup>13</sup> In *TradeProd* the information is provided *at the bilateral level* and covers the period 1989-2001.

Tariffs, initially available at the HS 6-digit product level, are matched to the ISIC Rev. 2 classification using world imports as weights. We use the information from Jon Haveman's treatment of TRAINS data (UTBC database) for the period 1989-2000. For the year 2001, we employ CEPII's dataset MAcMap (Market Access Map) that takes into account the complex system of bilateral preferences across countries in the world at a detailed product level.

Concerning NTBs, we set five different types: (1) those with a price effect; (2) those with a restriction on quantity; (3) those with a restriction on quality; (4) threatening measures and; (5) those subject to advanced payment. For each HS6 product,  $NTB = 1$  when there is a NTB at a lower level or 0, when there

---

In fact, in the same example from the Table 3, we obtained a positive value for internal trade flows for Belgium, but not for Luxembourg, which suggests that internal trade flows for Belgium are over-estimated. In sum, caution is needed when information for these two countries is employed.

<sup>10</sup>Using only World Bank data gives 3.6% of negative flows in trade with self for the 1980-2004 period. Using BACI information gives 3.9%, and the correlation between both measures is 0.997. We opted to keep both estimates.

<sup>11</sup>We found 117 negative figures for value added.

<sup>12</sup>UNCTAD provides detailed information in the TRAINS database. Jon Haveman has extracted some of these data, currently available on Raymond Robertson's Home Page.

<sup>13</sup>In N&O 2006, tariffs are (1) the simple average applied tariff rate or (2) the import weighted average applied tariff rate. The NTBs consider (1) the percentage of import subjected to NTBs and (2) the percentage of tariff lines (within each ISIC product) subject to NTBs. Tariff data is available since 1988 and NTBs for the period 1992-2001.

is none. These HS6 level data are aggregated in ISIC Rev. 2, in order to match our trade and production data, through frequency and covering indexes.<sup>14</sup>

### 3 Final considerations

This Dataset is available on CEPII’s website. It is provided on an “as is” basis and free of charge. The access is in the database section.<sup>15</sup> Other data useful for gravity equations is provided in separate files on CEPII’s website, like bilateral distances, official languages, colony links, etc. Thank you for giving us the information about the work for which you need our data and for comments about errors found.

### References

- Gaulier, G. and S. Zignago (2008). BACI: A World Database of International Trade Analysis at the Product Level. Technical report, CEPII Working Paper (Forthcoming).
- Mayer, T. and S. Zignago (2005). Market Access in Global and Regional Trade. CEPII Working Paper 2005-02, CEPII.
- Nicita, A. and M. Olarreaga (2001). Trade and Production, 1976-1999. Technical Note, Development Research Group, Trade (DECRG-TR). World Bank.
- Nicita, A. and M. Olarreaga (2006). Trade, Production and Protection 1976-2004. Technical Note, Development Economics Research Group (Trade) World Bank.
- Nicita, A. and M. Olarreaga (2007). Trade, Production, and Protection Database, 1976-2004. *World Bank Economic Review* 21(1), 165–171.

---

<sup>14</sup>In the case of the frequency indexes we made a simple ratio at the ISIC level of the number of HS subsectors subject to the NTB over the total HS subsectors. In the case of coverage indexes, we first weighted the frequency ratio at HS6 level using the world imports for that category, and we then aggregated at the ISIC level. This indicator was finally divided by the world imports at the corresponding ISIC level.

<sup>15</sup>For the English version: <http://www.cepii.fr/anglaisgraph/bdd/TradeProd.htm> and for the French version: <http://www.cepii.fr/francgraph/bdd/TradeProd.htm>

Table 1: Total Trade for the 1995-2004 period, by Industry Sector (ISIC 3 digits Rev. 2).

ind3d	<i>TradeProd</i>	<i>N&amp;O2006</i>	Difference All values	Difference w/o China
311	2956.6	2472.2	484.4	449.8
313	462.8	379.6	83.2	78.5
314	167.1	141.1	25.9	24.3
321	2010.8	1374.2	636.7	496.0
322	1555.9	1471.6	84.3	80.3
323	340.7	421.8	-81.1	-17.3
324	478.5	365.4	113.2	102.7
331	594.3	602.3	-8.0	2.1
332	77.5	551.2	-473.7	-410.6
341	1252.5	1165.2	87.3	90.7
342	479.7	338.0	141.7	145.1
351	3583.8	4082.1	-498.3	-457.7
352	3071.3	2239.3	832.0	784.4
353	1533.0	1380.0	153.0	140.1
354	33.3	53.5	-20.2	-30.9
355	553.4	471.9	81.5	73.7
356	1008.3	718.3	290.0	365.9
361	112.5	95.5	17.0	20.4
362	320.8	287.0	33.8	33.7
369	368.0	344.3	23.7	30.0
371	1453.4	1607.0	-153.5	-124.3
372	1817.7	1560.5	257.1	245.6
381	1979.8	1565.1	414.7	437.5
382	8453.4	8420.7	32.7	81.4
383	8233.4	8203.2	30.2	155.1
384	8150.1	7196.7	953.4	921.0
385	1938.8	1830.7	108.1	150.3
390	1249.4	1774.4	-525.0	-401.5
Total	54236.7	51112.7	3124.1	3466.4

Source: *Trade\_cep9504.dta* from *TradeProd* and *bilateraltrade.zip* from N&O2006 (5 files).  
Trade in billions of USD. Difference is trade value in *TradeProd* minus trade value in N&O 2006 (after discarding origin/destinations that are not countries).

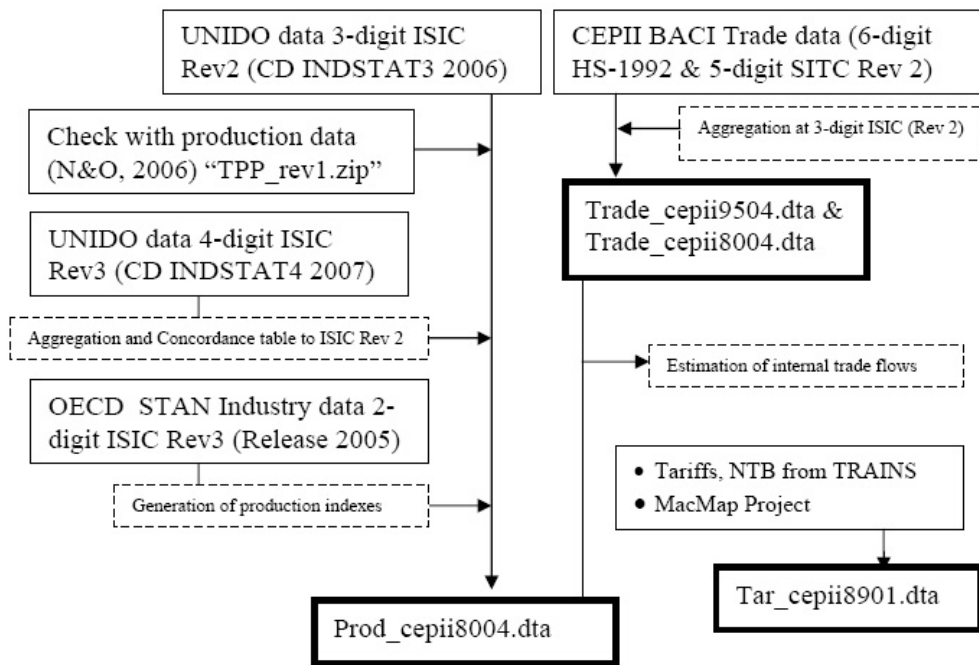


Figure 1: Sources and procedures to build *TradeProd*.

Table 2: **Additional flows by Continent for the 1995-2004 period.**

Continent of Exporter	Continent of Importer						TOTAL
	Africa	America	Asia	Europe	Pacific	N.C.	
Africa	39374	9760	20503	22605	1968	294	94504
America	16006	48121	21467	19671	4808	292	110365
Asia	45005	28501	60762	28946	16967	968	181149
Europe	73001	57666	82407	48221	14962	1342	277599
Pacific	3198	3472	6212	3659	9434	137	26112
N.C.	164	140	503	372	2	-	1181
<b>TOTAL</b>	<b>176748</b>	<b>147660</b>	<b>191854</b>	<b>123474</b>	<b>48141</b>	<b>3033</b>	<b>690910</b>

Source: *Trade\_cepii9504.dta* from *TradeProd* and *bilateraltrade.zip* from N&O2006 (5 files).

N.C.: Not classified.

Table 3: **Data for Belgium, Luxembourg and BLX.**

Country	year	Industry	Imports	Exports	Production	Number of Firms	Number of Employees	Wage bill	V.A.	Internal Flow	Wage per employee
BEL	1993	311			3.32e+07	6813		1602242	6317719		
BLX	1993	311	6877884	9812855	372552.1	7022	2974	1665804	6418915	-9440303	560.1223
LUX	1993	311			372552.1	209	2974	63561.58	101195.7		21.37242

BEL=Belgium, LUX=Luxembourg, BLX is the aggregation of both countries.

Table 4: **Production data available by year**

<b>year</b>	<b>firms</b>	<b>lab</b>	<b>wage</b>	<b>prod</b>	<b>va</b>	<b>intflow</b>
1980	29	2799	2560	2573	2705	2142
1981	2443	2800	2563	2628	2745	2181
1982	2369	2699	2515	2630	2627	2167
1983	2396	2678	2524	2521	2542	2079
1984	2439	2722	2601	2604	2510	2135
1985	2548	2656	2509	2524	2490	2043
1986	2484	2737	2492	2559	2542	2064
1987	2477	2726	2452	2530	2514	2011
1988	2510	2742	2423	2451	2461	1963
1989	2519	2680	2403	2568	2522	2034
1990	2669	2756	2450	2585	2512	2011
1991	2955	2864	2502	2694	2522	2007
1992	2944	2745	2422	2673	2360	2116
1993	3083	2858	2508	2721	2509	2203
1994	3043	2827	2544	2743	2500	2158
1995	2853	2778	2405	2684	2417	2093
1996	2828	2729	2376	2701	2378	2075
1997	2529	2497	2303	2482	2132	1904
1998	2554	2532	2289	2456	2120	1868
1999	2342	2292	1996	2195	1901	1654
2000	2344	2333	2047	2209	1957	1657
2001	2156	2094	1857	2098	1817	1576
2002	1865	1737	1596	1868	1584	1414
2003	1430	1327	1144	1492	1127	1145
2004	558	537	500	533	405	398
<b>Total</b>	<b>58367</b>	<b>62145</b>	<b>55981</b>	<b>59722</b>	<b>55899</b>	<b>47098</b>

Sources: *Prod\_cepri8004.dta*. Firms: Number of establishments; lab: Number of Employees; wage:wage per employee (wage bill/number of employees); prod: Value of goods produced; va: Value Added; intflow: Internal flows (prod - total exports). Only positive flows were retained.

Table 5: **Production data available by Industry Sector (ISIC 3 digits Rev. 2).**

Industry	firms	lab	wage	prod	va	intflow
311	2228	2481	2240	2376	2127	2186
313	2136	2395	2154	2311	2072	2044
314	1985	2133	1890	1997	1864	1621
321	2215	2415	2208	2328	2124	1941
322	2007	2280	2037	2149	1899	1336
323	2091	2221	2027	2156	1951	1515
324	1943	2107	1902	2021	1810	1596
331	2182	2424	2193	2324	2117	1978
332	1826	2122	1868	2029	1785	1779
341	2193	2419	2228	2335	2139	2067
342	2062	2302	2055	2186	1931	2033
351	2099	2252	2034	2200	2033	1710
352	1961	2212	2010	2182	1916	1915
353	1841	1733	1548	1712	1743	1455
354	1558	1316	1074	1369	1484	968
355	2073	2224	2023	2124	2008	1743
356	2010	2177	1953	2075	1872	1860
361	1922	1875	1659	1835	1868	1557
362	2007	2124	1922	2014	1902	1745
369	2010	2196	1963	2115	1880	1919
371	2010	1932	1767	1885	1940	1594
372	1715	1515	1353	1530	1654	1011
381	2163	2383	2152	2259	2029	2023
382	1951	2088	1912	1988	1923	1474
383	1951	2133	1928	2011	1871	1632
384	1979	2151	1943	2036	1893	1607
385	1919	1781	1602	1673	1816	1207
390	2064	2194	1996	2086	2006	1582
Total	56101	59585	53641	57306	53657	47098

Source: *Prod\_cepü8004.dta*. For variable definitions, see Table 4.