TradeProd V202401 Documentation

tradeprod@cepii.fr

Released June 2024

The 2024 version of the CEPII Trade and Production database (TradeProd) provides data on international and domestic trade flows and trade protection at the bilateral level for **165** countries and **9** industrial sectors over the period **1966-2020**. When using this dataset please cite:

Thierry Mayer, Gianluca Santoni & Vincent Vicard, 2023. The CEPII Trade and Production Database, CEPII Working Paper 2023-01, January 2023, CEPII.

The database consists of 3 different files:

TPe This version is intended for estimation purposes. The dataset is not squared, and domestic production is not extrapolated. The *TPe* database includes three different series for trade: i) *trade_i* based on importing country declarations (trade flow from origin *o* to destination *d* as reported by country *d*); ii) *trade_e* based on exporting country declarations (trade flow from *o* to *d* as reported by country *o*); and iii) *trade_comb* which combines import declaration flows with export declarations whenever the import data is missing (combined trade series ensure broader coverage).

TPc This version is intended for counterfactual exercises using new quantitative trade models. The simulation-oriented database includes two different series for trade based on *trade_comb*: i) *trade_sq* which is squared by industry and year, i.e., it is non-missing for the same set of origin and destination countries by industry and year; ii) *trade_sq_yr* which is squared by year, i.e., it is non-missing for the same set of origin and destination countries for all 9 industries in a year. The *TPc* version also includes a rest of the World aggregate, "ROW". Gross production for the "ROW" is extrapolated using the average gross output to total export ratio in a given industry and year.

TradeProd_Gravity_country_key To facilitate merging with the CEPII-Gravity database, we provide both the TradeProd (*cnum*) and the Gravity country iso codes (*iso*3*num*) in an additional file. The merge with the CEPII-Gravity database must be done by iso3num and year.

Descriptive statistics

Table 1 reports the list of variables in the dataset. Trade flows are identified by country of origin (*iso3_o*), country of destination (*iso3_d*), industry, and year. In the TPc version, the dummy variable *flag_extra_cty* identifies the domestic sales observations based on extrapolated gross output using the adjusted country-specific output to export ratios, whereas the dummy variable *flag_extra_avg* identifies the domestic sales observations extrapolated using industry averages, as for the ROW aggregate. Finally, the dummy variable *flag_extra_neg* identifies the observations (year-country-industry) for which the production reported in INDSTAT results in negative domestic sales, which are then set as missing and extrapolated. [1]

Figure 1 reports a summary of the coverage of the database comparing the total manufacturing output in TradeProd, computed as yearly $\sum_{odk} trade_{odk,t}$, with the aggregated figures reported in INDSTAT (ISIC D aggregate). Overall, TradeProd ensures great coverage: over the period 2010-2020, TradeProd traces around 97 percent of world manufacturing production in the TPe version and 98 percent in the TPc version.

Table 2 reports the country coverage of the database by decade and industries, for both the TPe and the TPc version.

Table 1: List of variables included in TradeProd

version	Variable	type	Description	Note
Common to TPe and TPc	year	int		1966-2018
	industry	str3	Based on 2-digit ISIC Rev. 3	9-industry aggregates
	iso3_tp_o	str4	origin country	ISO3 alphabetic code,
	iso3_tp_d	str4	destination country	territorial changes conform to CEPII-gravity
	cnum	str4		ISO3 numeric code from ComTrade
		str4		only in *TradeProd_Gravity_country_key*
	$tariff_{MFN}$	double	MFN tariff rate	simple average WITS HS 6-digit, starts in 1988
	$tariff_{pref}$	double	Preferential tariff rate	simple average WITS HS 6-digit, starts in 1988,
	$tariff$	double	combines MNF & Pref rate	Min($tariff_MFN$, $tariff_pref$)
TPe	*trade_i*	double	value of trade (Mln US $)	trade flow from o to d as reported by country d
	trade_e	double	value of trade (Mln US $)	trade flow from o to d as reported by country o
	trade_comb	double	value of trade (Mln US $)	combines *trade_i* with *trade_e*
TPc	*trade_sq*	double	value of trade (Mln US $)	squared by industry and year
	trade_sq_yr	double	value of trade (Mln US $)	squared by year
	flag_extra_neg	double	= 1 extrapolated negative domestic sales use country export to output ratios	
	flag_extra_cty	double	= 1 extrapolated domestic sales use country export to output ratios	
	flag_extra_avg	double	= 1 extrapolated domestic sales use average export to output ratios	

Note: $tariff_{MFN}$ and $tariff_{pref}$ are computed starting from HS 6-digit from the World Bank World Integrated Trade Solution (WITS) database. 6-digit values are aggregated to match the 9-industry grouping taking simple averages.

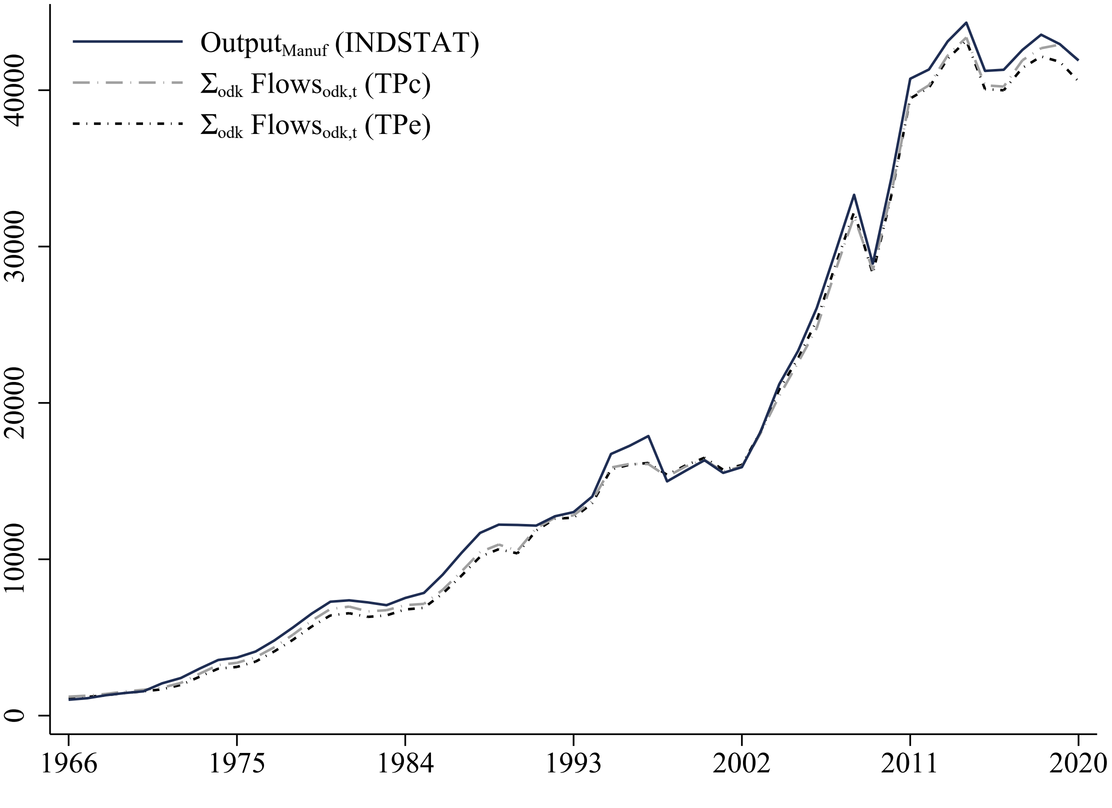
[1] Notice that the “flag” dummy variables are expanded to identify the country with extrapolated domestic sales across all its bilateral observations, hence a simple *drop(keep) if* condition selects the relevant sample.

Table 2: Coverage by decade, countries with non missing domestic sales

Decade	Version	# Origin and Destinations, industry−by−year Average	 Max (industry)	 Min (industry)
1966-1979		81	99 (15t16)	57 (34t35)
1980-1989		93	106 (15t16)	78 (34t35)
1990-1999	*TPe*	105	125 (15t16)	77 (29t33)
2000-2009	*trade_comb*	104	123 (15t16)	74 (29t33)
2010-2020		100	122 (15t16)	62 (17t18)
1966-1979		111	126 (23t25)	87 (34t35)
1980-1989		118	127 (23t25)	104 (34t35)
1990-1999	*TPc*	131	149 (15t16)	103 (34t35)
2000-2009	*trade_sq*	136	155 (15t16)	108 (29t33)
2010-2020		136	155 (15t16)	108 (29t33)
			Max (year)	Min (year)
1966-1979		129	134 (1979)	121 (1966)
1980-1989	*TPc*	134	134	134
1990-1999	*trade_sq_yr*	151	157 (1999)	133 (1990)
2000-2009		161	162 (2006)	161 (2005)
2010-2020		161	162 (2010)	161 (2020)

Note: The Max and Min columns also indicate the industry with the narrower and broader coverage by decades' average, or the year with the narrower/broader coverage for all industries.

Figure 1: World Manufacturing Output



Note: TThe graph reports the total manufacturing output in TradeProd and INDSTAT. Total output in TradeProd is computed as the yearly $\sum Flows_{odk,t}$; whereas INDSTAT total manufacturing output refers to the ISIC D aggregate.

Benchmark Estimates

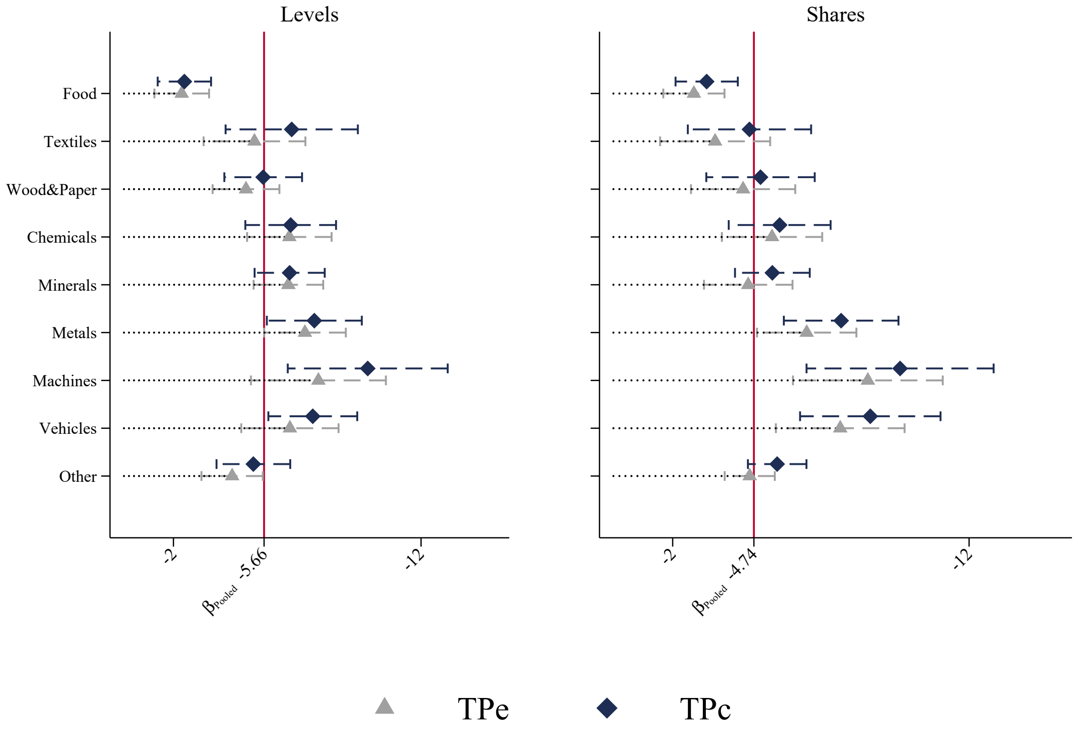
In this section, we report a simple benchmark exercise on: i) the log$(1 + tariff)$; ii) the border effect; iii) the log(distance). We estimate a standard gravity equation using a PPML estimator as follows:

$$T_{odkt} = \exp(\beta_1 \log(1 + t_{odkt}) + \beta_2 B_{dk} + \beta_3 \log(d_{od}) + \beta_4 X_{od} + \omega_{okt} + \omega_{dkt}) + \epsilon_{odkt}. \tag{1}$$

We present both the results using the dependent variable T_{odkt} in level (T_{odkt} represents exports from country o to country d in sector k and year t) and in share of destination country absorption (T_{odkt} is exports divided by total imports of country d in sector k and year t).

t_{odkt} is the import tariff rate, B_{dk} a border effect dummy equal to one when $o \neq d$, and d_{od} is weighted distance. X_{od} include the usual dyadic trade cost components: common language, contiguity, and colonial ties. Finally, ω_{okt} and ω_{dkt} are fixed effects by country-industry-year that control for Multilateral Resistance Terms. Standard errors are two-way clustered at the origin country and destination country levels.

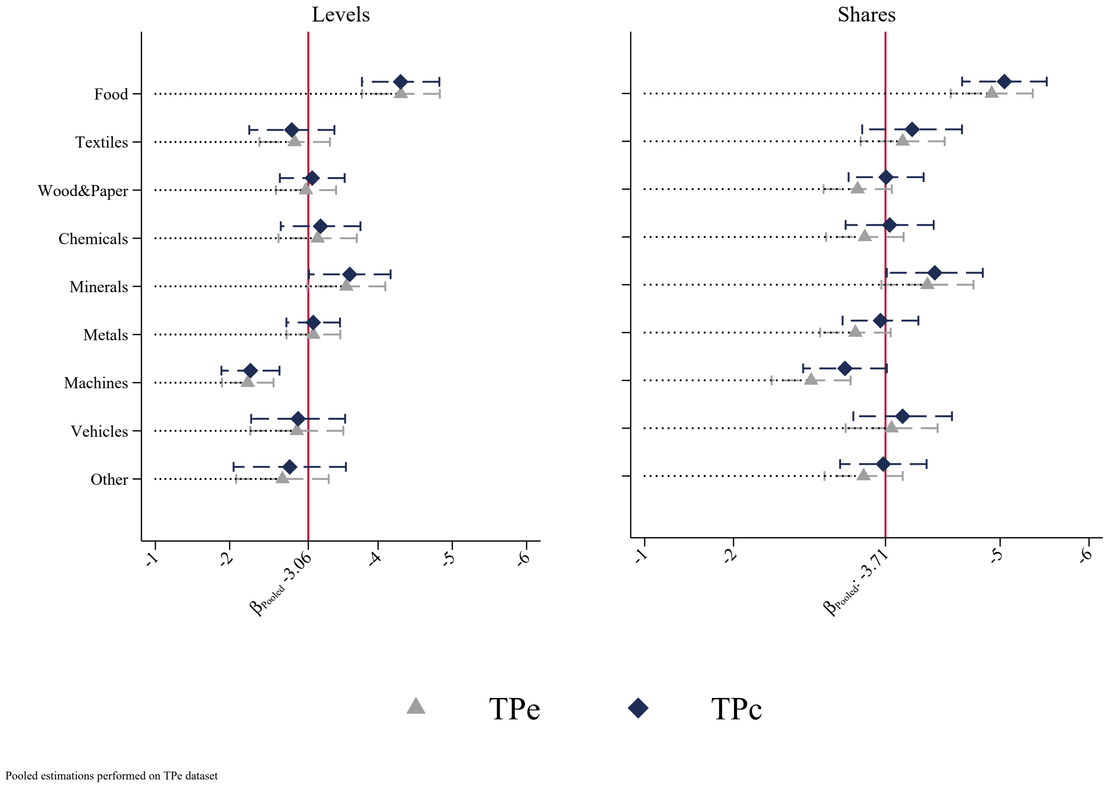
Figure 2: Tariff point estimates by industry



Pooled estimations performed on TPe dataset

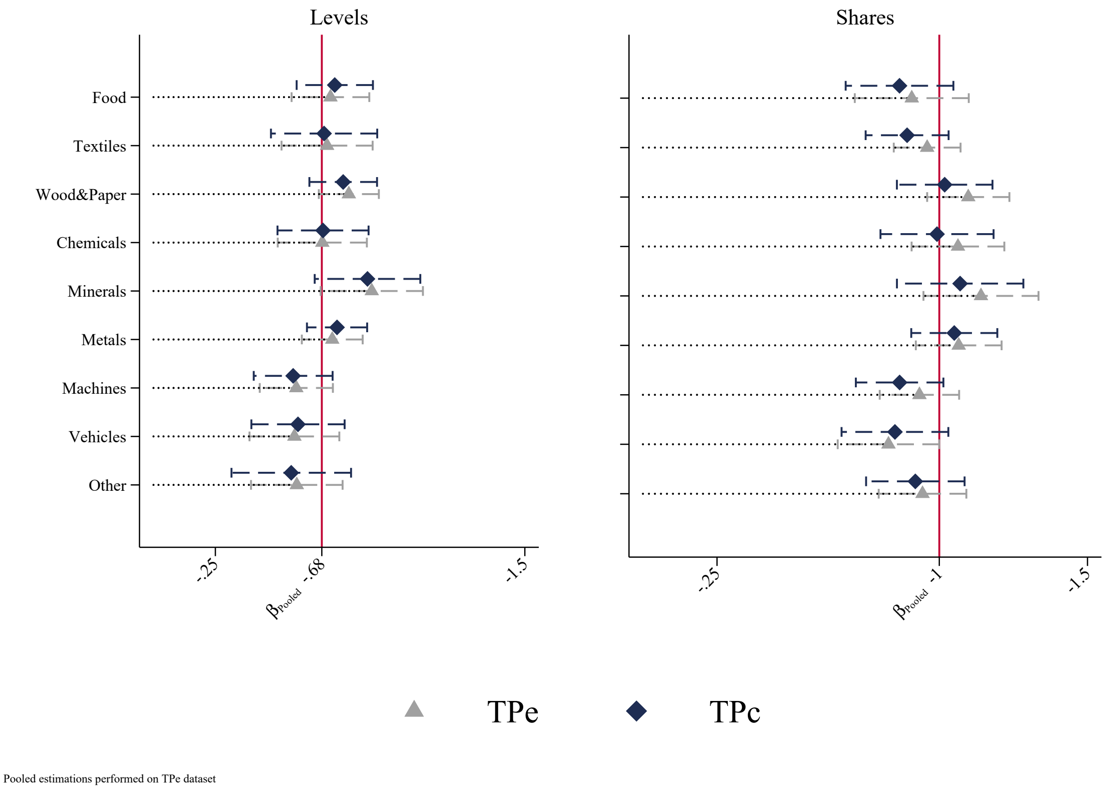
Note: The graph reports the estimated coefficients for $log(1 + tariff)$ from industry-specific regressions on the period 1988-2018 controlling for both outward and inward multilateral resistance terms, i.e., with origin-by-year and destination-by-year fixed effects, as well as dyadic fixed effects (origin-by-destination). The left panel plots coefficients from a PPML regression in levels, while the right panel reports the estimated coefficients from PPML regression in shares of destination absorption. Whiskers display 95% confidence intervals ($\pm 1.96 * SE$), where standard errors, SE, are two-way clustered at the origin and destination level. β Pooled refers to the estimated effects in the industry pooled sample with origin-industry-year, destination-industry-year, and origin-destination-industry fixed effects.

Figure 3: International Border point estimates by industry



Note: The graph reports the estimated international border coefficients from industry-specific regressions on the period 1966-2018 controlling for both outward and inward multilateral resistance terms, i.e., with origin-by-year and destination-by-year fixed effects. The estimated equation also controls for common language, contiguity, and colonial ties. The left panel plots coefficients from a PPML regression in levels, while the right panel reports the estimated coefficients from PPML regression in shares of destination absorption. Whiskers display 95% confidence intervals ($\pm 1.96 * SE$), where standard errors, SE, are two-way clustered at the origin and destination level. β Pooled refers to the estimated effects in the industry pooled sample with origin-industry-year and destination-industry-year fixed effects.

Figure 4: Distance point estimates by industry



Pooled estimations performed on TPe dataset

Note: The graph reports the estimated coefficients for $log(distance)$ from industry-specific regressions on the period 1966-2018 controlling for both outward and inward multilateral resistance terms, i.e., with origin-by-year and destination-by-year fixed effects. The estimated equation also controls for common language, contiguity, and colonial ties. The left panel plots coefficients from a PPML regression in levels, while the right panel reports the estimated coefficients from PPML regression in shares of destination absorption. Whiskers display 95% confidence intervals ($\pm 1.96 * SE$), where standard errors, SE, are two-way clustered at the origin and destination level. β Pooled refers to the estimated effects in the industry pooled sample with origin-industry-year and destination-industry-year fixed effects.