ECONOMICS WORKING PAPER SERIES

EXTENDING THE CEPII GRAVITY DATA SET

Tamara Gurevich
Peter Herman
Serge Shikher
Ricky Ubee

Working Paper 2017–01–A

U.S. INTERNATIONAL TRADE COMMISSION
500 E Street SW
Washington, DC 20436
January 2017

Special thanks to Nabil Abbyad and Grace Kenneally for research assistance with this working paper.

Extending the CEPII Gravity Data Set
Tamara Gurevich, Peter Herman, Serge Shikher, and Ricky Ubee
Office of Economics Working Paper 2017–01–A

Abstract

This research note describes in detail the process and data sources used in the CEPII gravity data set update for the years 2007 to 2015. This data update preserves the nomenclature and the structure of the original CEPII data for easier integration into ongoing studies.

Tamara Gurevich, U.S. International Trade Commission
Tamara.Gurevich@usitc.gov
Peter Herman, U.S. International Trade Commission
Peter.Herman@usitc.gov
Serge Shikher, U.S. International Trade Commission
Serge.Shikher@usitc.gov
Ricky Ubee, U.S. International Trade Commission
Ravinder.Ubee@usitc.gov

1 Introduction

The gravity data set made available by CEPII (Head et al., 2010) has become a mainstay in trade research.[1] However, the data set has not been updated since 2006. This research note describes the process and data sources used in constructing an update that spans the years 2007–2015. We provide two data files, one containing only the updated variables for the years 2007–2015, and another containing all years of original CEPII and the updated years, spanning the entire period 1948–2015.[2]

This update preserves the nomenclature and structure of the original CEPII data set. The extension is constructed by first creating a balanced panel of bilateral relations between the 224 countries present in the original CEPII database for the years 2007–2015. We then separate CEPII variables into time varying and time invariant, consistent with the original CEPII definitions. For all but a handful of variables described in detail below, we assume the variables are time-invariant and assign the corresponding values from 2006 in the original CEPII data set to all subsequent years. Several variables that are expected to be significantly time-varying (GDP, GDP per capita, population, RTAs, common currency, and GATT/WTO membership) are updated with the appropriate data for each year, subject to data availability. For details on time-varying and time-invariant variables see tables 2 and 3 in the data appendix. The updates to time-varying variables are discussed in greater detail in the following section.

In some cases, such as *comlang_ethno* or *conflict*, time insensitivity is debatable, since linguistic composition of counties, as well as their conflict status, may change over time. Here, we have elected to use the 2006 values present in the original CEPII database. In most cases, we have done so because the original database has treated them as being time-invariant. In other cases, such as the variable for *conflict*, it is unclear what criteria the construction of the original variable followed, making a consistent extension difficult. Additionally, we have chosen not to introduce new countries that have gained independence after 2006 in order to preserve the balanced panel of 224 countries in the original database.

2 Methodology for Updated Variables

2.1 GDP and GDP per capita

We use data from the World Bank's World Development Indicators (WDI) database to update GDP and GDP per capita of both trading partners.[3] The CEPII variables we

[1] `http://www.cepii.fr/CEPII/en/bdd_modele/presentation.asp?id=8`

[2] "grav_data_2007to2015.csv" and "grav_data_1948to2015.csv" respectively

[3] The original WDI data are available through World Bank's website at `http://data.worldbank.org/data-catalog/world-development-indicators`.

are updating are *gdp_o*, *gdp_d*, *gdpcap_o*, and *gdpcap_d*. The values in these variables are measured in current U.S. dollars, consistent with the original CEPII data.

2.2 Population

Population data come from the WDI database described above. We update the variables *pop_o* and *pop_d* for each year in the period 2007–2015.

2.3 RTAs and Currency Unions

We update CEPII variables *rta* and *comcur* using a data set created by Egger and Larch (2008) based on the list of trade agreements provided by the WTO.[4] However, the WTO list of trade agreements and common currencies appears to be inconsistent with the CEPII original definitions of *rta* and *comcur*, as evident from a rather large increase in the number of RTAs and common currency unions between the years 2006 and 2007. For example, the final year of the original CEPII database, 2006, exhibits 2,734 country-pairs in an RTA. The data collected by Egger and Larch (2008), used here to extend the CEPII data set, shows 6,470 country-pairs in an RTA in 2006, suggesting a significant disagreement between the two data sources as to what qualifies as an RTA.

In light of this discrepancy, we create two additional variables, *rta_larch* and *comcur_larch*, which replace the CEPII variable with the WTO/Larch data for all years rather than only 2007–2015. The *rta* and *comcur* variables take values provided by CEPII for the years 1948–2006 and values provided by the WTO/Larch for the years 2007–2015.

2.4 WTO Membership

We update the variables *gatt_o* and *gatt_d* for countries that joined the WTO after 2006 using the WTOs membership timeline.[5] The *gatt* variables were updated to take the value of one for all years greater than or equal to the year in which they joined the WTO. For the full list of countries and accession dates see table 1 in the Data Appendix below.

3 Conclusion

This research note provides documentation for our updated CEPII data set that extends the set of gravity variables from 2006 to 2015. We attempt to do so in a way that preserves consistency with the original CEPII data set and readily permits its

[4] `http://rtais.wto.org/UI/PublicAllRTAList.aspx`

[5] `https://www.wto.org/english/thewto_e/whatis_e/tif_e/org6_e.htm`

usage in gravity research, past and present. In particular, we update variables for GDP, GDP per capita, population, and WTO membership. Time-invariant variables are defined so as to be consistent with the original CEPII data.

We are faced with a greater challenge when updating *rta* and *comcur* variables. CEPII does not provide documentation describing origins of these two variables, and they appear inconsistent with the WTO list of accepted trade agreements and common currency unions. To reconcile this issue, we use the WTO data for years 2007–2015 and provide two additional variables that are WTO-consistent, one for the trade agreements and one for the common currency unions.

We expect that this updated CEPII data will prove helpful to researchers at USITC and to trade researchers and gravity modelers outside the USITC.

Data Appendix

On request, the authors can provide a data archive that consists of several source files, a code to combine the data, and two final data sets.

The data were generated using the R code "Cepii Extender v1.R" that requires the following files as inputs:

"GDPs_From_World_Development_Indicators_CurrentUSdollars.csv"
"gravdata_cepii.dta"
"rta_20160215_stata12.dta"
"WDI_population.dta"
"WTO_Members_post2006.csv"

The GDP and GDP per capita variables *gdp_o*, *gdp_d*, *gdpcap_o*, and *gdpcap_d* come from the World Bank's WDI data set. They are all measured in current U.S. dollars and can be found in the file "GDPs_$\cdots$_CurrentUSdollars.csv"

The population variables *pop_o* and *pop_d* were updated using figures from the World Bank's WDI. The file "WDI_population.dta" contains these data.

The variables *rta* and *comcur* were updated using data made available by Mario Larch.[6] Larch's data are in the file "rta_20160215_stata12.dta".

WTO Membership was determined based on the list provided by the WTO.[7] The new entrants were:

[6] See detailed discussion in the Methodology section

[7] `https://www.wto.org/english/thewto_e/whatis_e/tif_e/org6_e.htm`

Table 1: New Entrants

Country	WTO ISO	Entry Date
Afghanistan	AFG	Jul 29, 2016
Cape Verde	CPV	Jul 23, 2008
Kazakhstan	KAZ	Nov 30, 2015
Liberia	LBR	Jul 12, 2016
Montenegro[†]	MNE	Apr 29, 2012
Russian Federation	RUS	Aug 22, 2012
Samoa	WSM	May 10, 2012
Seychelles	SYC	Apr 26, 2015
Tajikistan	TJK	Mar 2, 2012
Tonga	TON	Jul 27, 2007
Ukraine	UKR	May 16, 2008
Vanuatu	VUT	Aug 24, 2012
Vietnam	VNM	Jan 11, 2007
Yemen	YEM	Jun 26, 2014

[†] Montenegro is not present in the data as it gained independence after 2006 and is not included in the original panel of 224 countries.

Table 2: CEPII Gravity Time-Invariant Variables

Variable	Description
iso3_o	ISO3 code alphanumeric of origin[†]
iso3_d	ISO3 alphanumeric of destination[†]
iso2_o	ISO2 code alphanumeric of origin
iso2_d	ISO2 alphanumeric of destination
area_o	Area of origin in sq. kms
area_d	Area of destination in sq. kms
heg_o	Origin is current or former hegemon of destination
heg_d	Destination is current or former hegemon of origin
leg_o	Legal origin of iso3_o
leg_d	Legal origin of iso3_d
col_to	Binary indicator for trade from heg_o to colony
col_fr	Binary indicator for trade from colony to heg_o
contig	Contiguity – the two countries share a border
comlang_off	The two countries share a common official primary language
comlang_ethno	Common language spoken by at least 9% of the population in both countries
comcol	Both countries have a common colonizer post 1945
col45	The two countries are in a colonial relationship post 1945
distw	Population-weighted distance between the two countries, in kms
tdiff	Time difference between the two countries, in hrs
conflict	Binary indicator of war between the two countries
indepdate	Independence date, if former colony
colony	The two countries are or have ever been in colonial relationship
curcol	The two countries are currently in colonial relationship
empire	Empire to which both countries belong
comleg	The two countries have common legal origin
acp_to_eu	Binary indicator for ACP to EU
eu_to_acp	Binary indicator for EU to ACP

[†] Origin refers to the first country in a country-pair, destination refers to the second.

Table 3: CEPII Gravity Time-Varying Variables

Variable	Description
gdp_o	GDP of origin, in millions current U.S. dollars[†]
gdp_d	GDP of destination, in millions current U.S. dollars[†]
gdpcap_o	GDP per capita of origin, in millions current U.S. dollars
gdpcap_d	GDP per capita of destination, in millions current U.S. dollars
pop_o	Population of origin, total in mn
pop_d	Population of destination, total in mn
gatt_o	Binary indicator for origin's membership in GATT/WTO
gatt_d	Binary indicator for destination's membership in GATT/WTO
rta	Binary indicator for regional trade in force
comcur	The two countries have common currency

[†] Origin refers to the first country in a country-pair, destination refers to the second.

Bibliography

Egger, P. H. and M. Larch (2008). Interdependent Preferential Trade Agreement Memberships: An Empirical Analysis. *Journal of International Economics 72*(2), 384–399.

Head, K., T. Mayer, and J. Ries (2010). The erosion of colonial trade linkages after independence. *Journal of International Economics 81*(1), 1–14.