The CEPII Gravity Database

Maddalena Conte[*] Pierre Cotterlaz[†] Thierry Mayer[‡]

May 26, 2021

Abstract

The Gravity database aims at gathering in a single place a set of variables that could be useful to researchers or practitioners willing to understand the determinants of international trade. Each observation corresponds to a combination of exporter-importer-year (i.e. origin-destination-year), for which we provide trade flows, as well as geographic, cultural, trade facilitation and macroeconomic variables.

[*]Bocconi University
[†]CEPII and Sciences Po
[‡]Sciences Po, Banque de France, CEPII and CEPR

Contents

1 Introduction **2**

2 The *Countries* dataset: Static country-level information **6**
2.1 Data Sources . . . 6
2.2 Variables . . . 6
2.3 Variable construction . . . 7

3 The *Gravity* dataset **8**
3.1 Country identifiers . . . 8
3.1.1 Data Sources . . . 8
3.1.2 Variables . . . 8
3.1.3 Variable construction . . . 10
3.2 Geographic variables . . . 10
3.2.1 Data Sources . . . 10
3.2.2 Variables . . . 10
3.2.3 Variable construction . . . 11
3.3 Cultural variables . . . 12
3.3.1 Data Sources . . . 12
3.3.2 Variables . . . 12
3.3.3 Variable construction . . . 14
3.4 Macroeconomic Indicators . . . 16
3.4.1 Data Sources . . . 16
3.4.2 Variables . . . 17
3.4.3 Variable construction . . . 18
3.5 Trade facilitation variables . . . 20
3.5.1 Data Sources . . . 20
3.5.2 Variables . . . 20
3.5.3 Variable construction . . . 21
3.6 Trade flow variables . . . 23
3.6.1 Data Sources . . . 23
3.6.2 Variables . . . 24
3.6.3 Variable construction . . . 25

A Appendix **31**
A.1 Country codes . . . 31
A.2 Territorial changes . . . 33
A.3 GDP and population data . . . 33
A.4 RTA data from the WTO . . . 37

1 Introduction

The gravity database aims at gathering in a single place a set of variables useful to researchers or practitioners willing to understand the determinants of international trade. Each observation corresponds to a combination of "exporter-importer-year" (i.e. "origin-destination-year"), for which we provide trade flows, as well as geographic, cultural, trade facilitation and macroeconomic variables.

Data spans from 1948 to 2019, and includes 252 countries, some of which only exist for a shorter period of time. The term "country" includes territories that are not formally independent, as well as past territorial configurations of countries.[1] The dataset is dynamic in the sense that it follows the ways in which countries have changed over time. It is "squared", meaning that each country pair appears every year, even if one of the countries in the pair actually does not exist. Nevertheless, when either destination or origin country do not exist, variables are set to missing, and dummy variables allow the users to easily get rid of the concerned observations.

Gravity is the main dataset, which contains the core information. In *Gravity*, countries are referred to using the variable *country_id*, which combines a country's ISO3 code with a number identifying potential territorial transformations of the country. We also include simple ISO3 codes (numeric and alphabetic) as additional country identifiers. We provide an additional dataset, *Countries*, that associates to each *country_id* the corresponding ISO3 country codes, as well as country name, and a set of variables enabling to track territorial changes.

Variables included in *Gravity* may correspond to unilateral characteristics (GDP, population...), or to bilateral characteristics (distances, trade flows...). In the case of unilateral variables, the name ends with *_o* when the information refers to the origin country, and with *_d* when it refers to the destination country. For instance, *country_id* (the variable identifying each country/territory) becomes *country_id_o* when referring to the origin and *country_id_d* when referring to the destination. Table 1 provides an exhaustive overview of the variables included in the main *Gravity* dataset. Note that the unilateral variables do not appear twice: for simplicity, we do not repeat their definition both for the origin and the destination.

Table 1: List of variables included in *Gravity*

Variable Name	Content	Level
iso3	ISO3 alphabetic code	unilateral
iso3num	ISO3 numeric code	unilateral
countrygroup_iso3	Largest entity of which the country is/was part (ISO3 alphabetic)	unilateral
countrygroup_iso3num	Largest entity of which the country is/was part (ISO3 alphabetic)	unilateral
country	Country name	unilateral
countrylong	Offcial country name	unilateral

[1] For more details on the universe of countries included in the data, see Section 2.

first_year	First year of territorial existence of the country	unilateral
last_year	Last year of territorial existence of the country	unilateral
country_exists	1 if the country actually exists	unilateral
gmt_offset_2020	GMT offset in 2020 of the country (hours)	unilateral
contig	Dummy equal to 1 if countries are contiguous	bilateral
dist	Distance between most populated city of each country (km)	bilateral
distw	Population-weighted distance between most populated cities (km)	bilateral
distcap	Distance between capitals (km)	bilateral
distwces	Population-weighted distance between most populated cities (km) using CES formulation with $\theta = -1$	bilateral
dist_source	Dummy variable indicating the source of distance data (1 if taken directly from CEPII's GeoDist and 0 if based on close country)	bilateral
comlang_off	1 if countries share common official or primary language	bilateral
comlang_ethno	1 if countries share a common language spoken by at least 9% of the population	bilateral
comcol	1 if countries share a common colonizer post 1945	bilateral
col45	1 if countries are or were in colonial relationship post 1945	bilateral
legal_old	Historical origin of a country's laws before 1991	unilateral
legal_new	Historical origin of a country's laws after 1991	unilateral
comleg_pretrans	1 if countries share common legal origins before 1991	bilateral
comleg_posttrans	1 if countries share common legal origins after 1991	bilateral
transition_legalchange	1 if common legal origin changed in 1991	bilateral
comrelig	Religious proximity index	bilateral
heg_o	1 if origin is current or former hegemon of destination	bilateral
heg_d	1 if destination is current or former hegemon of origin	bilateral
col_dep_ever	1 if pair ever was in colonial or dependency relationship (including before 1948)	bilateral
col_dep	1 if pair currently in colonial or dependency relationship	bilateral
col_dep_end_year	Independence year from concerned hegemon (includes colonial ties before 1948)	bilateral

col_dep_end_conflict	1 if independence from the concerned hegemon involved a conflict	bilateral
empire	Common colonizer	bilateral
sibling_ever	1 if pair ever had the same colonizer (including before 1948)	bilateral
sibling	1 if pair currently has the same colonizer	bilateral
sever_year	Severance year for pairs that ever had the same colonizer (includes colonial ties before 1948): corresponds to the independence year of the first independent sibling	bilateral
sib_conflict	1 if pair ever had the same colonizer and independence involved a conflict with the hegemon (includes colonial ties before 1948)	bilateral
pop	Population (in thousands)	unilateral
gdp	GDP (current thousands US$)	unilateral
gdpcap	GDP per capita (current thousands US$)	unilateral
gdp_source	GDP data source	unilateral
pop_source	Population data source	unilateral
gdp_ppp	GDP PPP (current thousands international $)	unilateral
gdpcap_ppp	GDP per capita PPP (current thousands international $)	unilateral
pop_pwt	Population (in thousands) (source: Penn World Tables)	unilateral
gdp_ppp_pwt	Deflated GDP at current PPP (2011 thousands US$) (source: PWT)	unilateral
gatt	1 if country currently is a GATT member	unilateral
wto	1 if country currently is a WTO member	unilateral
eu	1 if country currently is a EU member	unilateral
rta	1 if the pair currently has a RTA (source: WTO)	bilateral
rta_coverage	Indicates whether the RTA covers goods only or goods and services (source: WTO)	bilateral
rta_type	Indicates the type of RTA (customs union for instance)	bilateral
entry_cost	Cost of business start-up procedures (% of GNI per capita)	unilateral
entry_proc	Number of start-up procedures to register a business	unilateral
entry_time	Days required to start a business	unilateral
entry_tp	Days required to start a business + number of procedures to start a business	unilateral
tradeflow_comtrade_o	Trade flow as reported by the exporter (in thousands current US$) (source: Comtrade)	bilateral

tradeflow_comtrade_d	Trade flow as reported by the importer (in thousands current US$) (source: Comtrade)	bilateral
tradeflow_baci	Trade flow (in thousands current US$) (source: BACI)	bilateral
manuf_tradeflow_baci	Trade flow of manufactured goods (in thousands current US$) (source: BACI)	bilateral
tradeflow_imf_o	Trade flow as reported by the exporter (in thousands current US$) (source: IMF)	bilateral
tradeflow_imf_d	Trade flow as reported by the importer (in thousands current US$) (source: IMF)	bilateral

Gravity is obtained by assembling data from many different sources. In particular, we use data produced by the CEPII, as well as data from institutional sources such as the World Bank, the WTO and the IMF. We also include data produced by a variety of researchers, which has been made publicly available. Table 2 gathers information on all sources used and papers that need to be cited when using these variables, along with the names of variables they were used to construct. Detailed explanations on how variables have been generated based on these sources are provided in the following sections.

We provide the dataset in three different formats: .csv (that can be read by any software), .dta (which requires Stata), and .Rds (which can be read in R). The data is distributed under the Etalab Open Licence 2.0, meaning that it can be freely used, modified, and shared as long as a proper reference is made to the source. The name of each file contains a version identifier: for instance *Gravity_V202010* refers to the October 2020 version of the *Gravity* dataset.

In the .csv and .Rds versions of *Gravity*, categorical variables are in a numeric format, meaning that we use a numeric code to refer to each category, instead of a set of characters. Therefore, we provide a set of files that associate their label to each numeric code. These files are named after the variable they describe (for instance, rta_type.csv describes the labels of the variable rta_type).

Note that the Gravity dataset is squared in terms of the variable *country_id*. As mentioned above, this variable combines a country's ISO3 code with a number identifying potential territorial transformations of the country. Alphabetic and numeric ISO3 codes are also included but these are not able to identify a country precisely. In particular, there are some cases in which countries experience territorial changes without this being reflected by a different ISO3 alphabetic code. This happens when i) a country merges with another and the unified country adopts the ISO3 alphabetic code of one of the two pre-existing countries; ii) a part of a country becomes independent, but the original country continues to exist. An example of the first case is West Germany, which had ISO3 alphabetic code "DEU" before its unification with East Germany, a code that was later adopted by the unified Germany. An example of the second case is Sudan, which had ISO3 alphabetic code "SDN" before the independence of South Sudan in 2011, a code that remained unchanged after this date. These issues are resolved when using the variable *country_id* as a country identifier because territorial changes can be identified pre-

cisely. For instance, *country_id* becomes DEU.1 for West Germany and DEU.2 for the unified Germany, while *country_id* becomes SDN.1 for Sudan before the independence of South Sudan in 2011, and SDN.2 after the independence of South Sudan. For further details, see Section 2, and Sections A.1 and A.2 of the Appendix.

2 The *Countries* dataset: Static country-level information

Countries is the dataset that includes static country-level variables, allowing for a full identification of each country included in *Gravity* and, if relevant, for a tracking of its territorial changes (splits and merges). Some of the variables provided in *Countries* are also included in the main *Gravity* dataset (see below and in Section 3.1).

Countries includes one observation for each territorial configuration, mapping the full set of territorial changes that are accounted for in *Gravity*. For example, *Countries* includes one observation for West Germany, one for East Germany and one for the unified Germany. Similarly, it includes one observation for Sudan before the split of South Sudan, one observation for South Sudan, and one observation for Sudan after the split of South Sudan.

2.1 Data Sources

The universe of *Countries* (and of the *Gravity* dataset) is based on CEPII's GeoDist dataset (Mayer and Zignago 2011). This dataset is augmented with some countries and territories that either appear in the World Bank's World Integrated Trade Solution (WITS) or that are necessary to construct the full chain of territorial changes that have led to the creation of countries appearing in the GeoDist dataset. In addition, some names are updated, as well as ISO3 alphabetic numeric codes, by comparing the GeoDist dataset with the WITS dataset and with the official source for ISO country codes. Countries' official names also come from the WITS dataset, augmented by Wikipedia for countries or territories that are not present in the WITS dataset but that appear in GeoDist.

Countries (and the *Gravity* dataset) carefully tracks territorial changes, i.e. the country's previous membership (in case of a split) and the country's new membership (in case of a unification of two territories). We only take into account the modifications that occurred over the time span of the database, i.e 1948-2019. This is done using the CIA World Factbook and Wikipedia.

2.2 Variables

- ***country_id***: country identifier, *unilateral*.
- ***iso3***: ISO3 alphabetic code, *unilateral*.
- ***iso3num***: ISO3 numeric code, *unilateral*.

- ***countrygroup_iso3***: Country group (ISO3 alphabetic code), *unilateral*.
- ***countrygroup_iso3num***: Country group (ISO3 numeric code), *unilateral*.
- ***country***: Country name, *unilateral*.
- ***countrylong***: Country official name, *unilateral*.
- ***first_year***: First year of territorial existence, *unilateral*.
- ***last_year***: Last year of territorial existence, *unilateral*.

2.3 Variable construction

The *Countries* datasets precisely identifies each country (in its current or past territorial configuration), including some territories that are not officially independent. As mentioned above the variable *country_id* uniquely identifies each country. Alphabetic and numeric ISO3 codes are also included. Differently from alphabetic ISO3 codes, a substantial territorial reconfiguration triggers a change in the numeric ISO3 code. Thus, looking at the numeric ISO3 code can also help to track territorial changes. In particular, if a country changes its name without any territorial change, the ISO3 numeric code remains the same, otherwise a new one is created. Section A.1 of the Appendix describes in detail instances where the ISO3 alphabetic or numeric code change over time. For instance, Sudan used ISO3 numeric code 736 before South Sudan split away in 2011. Since then, Sudan uses numeric code 729, while keeping the same alphabetic code ("SDN"). Similarly, West Germany already used "DEU" as alphabetic ISO3 code before reunification, but used to have 280 as numeric ISO3 code, which differentiates it from the current unified Germany, that has 276 as ISO3 numeric code.

Additional information on territorial re-configurations is provided by the following three variables:

1. The first year of territorial existence, that corresponds to the first year in which the country exists in its current territorial form.
2. The last year of territorial existence, that corresponds to the last year in which the country exists in its current territorial form.
3. The "country group" identifier (*countrygroup_iso3num*), that indicates the largest entity of which a country was or is part of. It therefore provides the country's previous membership (in case of a split) or the country's new membership (in case of a unification). *countrygroup_iso3num* (numeric ISO3 code) is complemented by *countrygroup*, that gives the alphabetic ISO3 code of the reference country, at the time of the territorial change.

For instance, in the case of Sudan, former Sudan has last year of existence 2011, and current Sudan and South Sudan both have first year of existence 2011. For all three countries, *countrygroup_iso3num* is the ISO3 numeric code of former Sudan, 736, because former Sudan is the

largest entity to which these countries belong. Similarly, East Germany (the German Democratic Republic) and West Germany both have 1990 as last year of existence, while current Germany has 1990 as first year of existence. The three countries have 276 as *countrygroup_iso3num*, which identifies the unified Germany.

This setup is only designed to track splits and unifications of countries, i.e. changes in territorial conformations. It does not take into account colonial/dependency links. Another set of variables is used to track such links, which is described in Section 3.2.

Figure 1 summarizes the territorial changes accounted for in *Gravity*. Most of the reconfigurations consist of countries splitting, with the USSR and Yugoslavia accounting for most of the new countries.

3 The *Gravity* dataset

3.1 Country identifiers

In *Gravity*, each observation is uniquely identified by the combination of the *country_id* of the origin country, the *country_id* of the destination country and the year. *Gravity* is "squared", meaning that each country pair appears every year, even if one of the countries actually does not exist. However, based on the territorial changes tracked in the *Countries* dataset (see Section 2), we set to missing all variables for country pairs in which at least one of the countries does not exist in a given year. Furthermore, we provide two dummy variables indicating whether the origin and the destination countries exist. These dummies allow users wishing drop non-existing country pairs from the dataset to do so easily. Users looking for a more detailed account of country existence should turn to the *Countries* dataset.

A few caveats on the identification of countries through *country_id* must be noted. Firstly, when countries merge, it is the new country or territorial configuration that exists during transition year but not the old country or territorial configuration. As an example DEU.1 (West Germany) has 1989 as last year, not 1990, while DEU.2 (the unified Germany) has 1990 as first year. This is consistent with the construction of underlying variables that varies over time, such as GDP, population, trade. Secondly, since the dataset is square in terms of *country_id*, there exist cases in which two configurations of the same alphabetic ISO3 code appear bilaterally, e.g. DEU.1 and DEU.2. While DEU.1 and DEU.2 never existed simultaneously, we still keep these null observations to ensure that the final dataset is square.

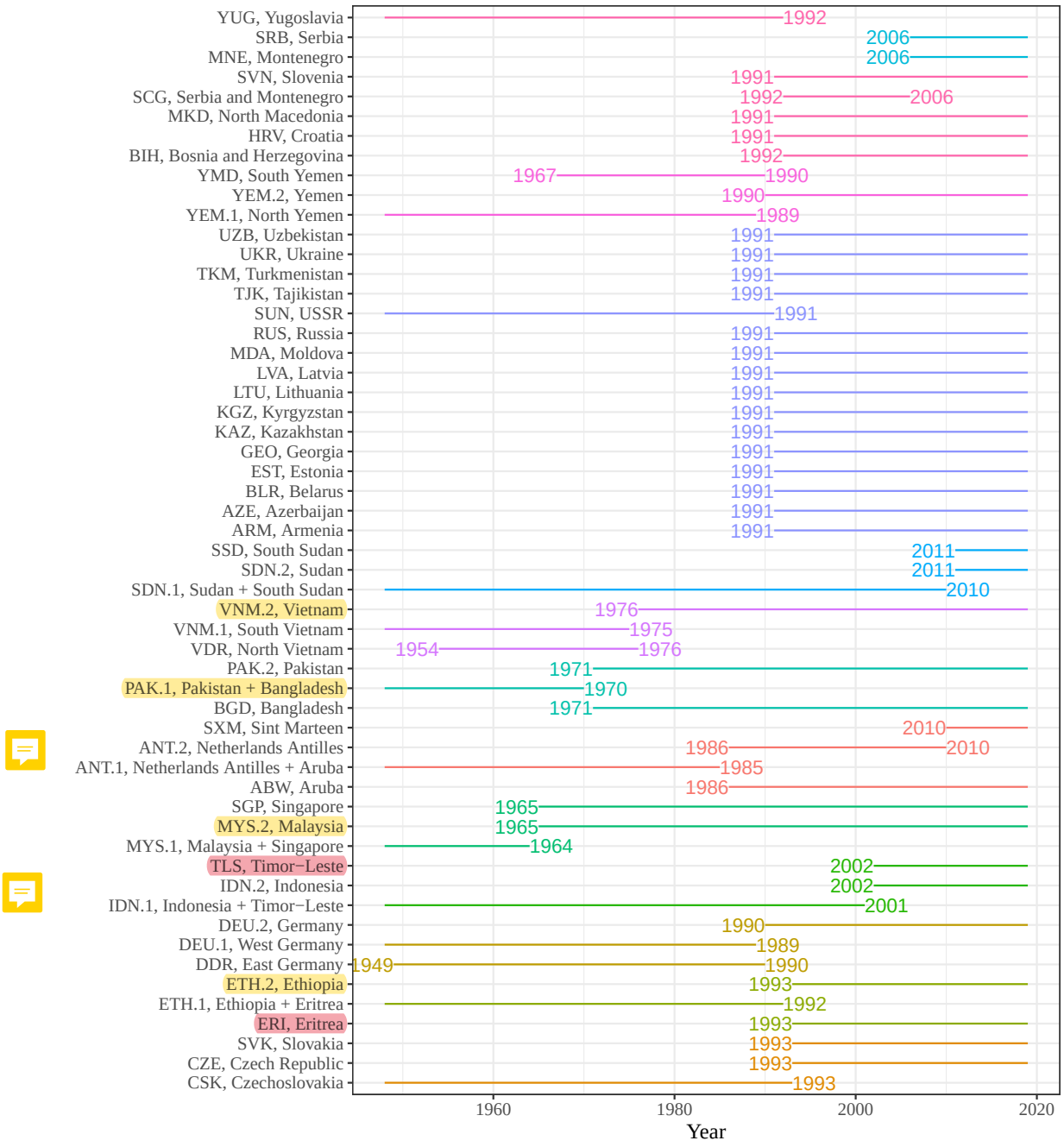
3.1.1 Data Sources

Data sources correspond to the same ones used in Section 2.

3.1.2 Variables

- ***country_id***: country identifier, *unilateral*.

Figure 1: Territorial changes.



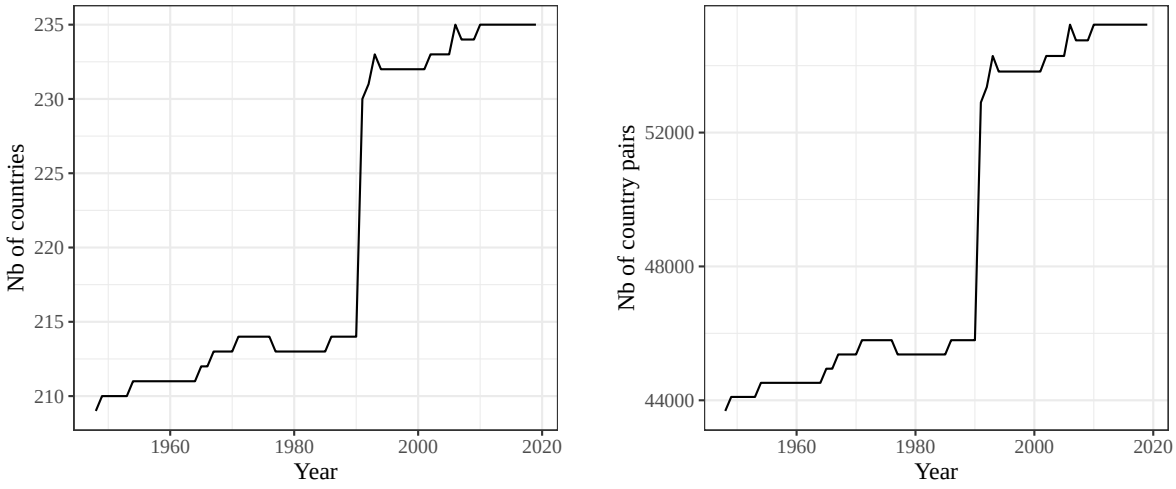
Notes: Lines represent the years during which countries actually exist. We include only the countries for which a territorial change is recorded.

- ***iso3***: ISO3 alphabetic code, *unilateral*.
- ***iso3num***: ISO3 numeric code, *unilateral*.
- ***country_exists***: Dummy equal to 1 if country exists in a given year *unilateral*.

3.1.3 Variable construction

$country\_exists$ dummies are based on the territorial changes detailed in the $Countries$ dataset. In figure 2, we plot the number of existing countries and the number of existing country pairs (excluding cases in which $iso3\_o = iso3\_d$).

Figure 2: Number of countries and country pairs in *Gravity* over time.



(a) Existing countries

(b) Existing country pairs

Notes: Existing country pairs are country pairs for which both *country_exists_o* and *country_exists_d* are equal to 1.

3.2 Geographic variables

3.2.1 Data Sources

Data on time zones (*gmt_offset_2020*) is taken from timezoneDB. We use the most up-to-date time zone as of July 2020. Further, we do not account for daylight saving time.

Geographic data on contiguity and distances (*contig*, *dist*, *distw*, *distcap*, *distwces*) comes from the CEPII's GeoDist database (for additional information on the construction of these variables, see Mayer and Zignago 2011). Note that this dataset has not yet been updated, hence some of the countries included in this version of the *Gravity* dataset are missing. We detail below the way we treated these missing countries.

3.2.2 Variables

- ***gmt_offset_2020***: GMT offset in 2020 of the country measured in hours, *unilateral*.
- ***contig***: Dummy equal to 1 if countries are contiguous, *bilateral*.

- ***dist***: Distance between most populated city of each country, measured in km, *bilateral*.
- ***distw***: Population-weighted distance between most populated cities, measured in km, *bilateral*.
- ***distcap***: Distance between capitals, measured in km, *bilateral*.
- ***distwces***: Population-weighted distance between most populated cities, measured in km, calculated using CES formulation with $\theta = -1$ (see Head and Mayer 2010 for more details), *bilateral*.
- ***dist_source***: Dummy variable indicating the source of distance data. It is equal to 1 if data is taken directly from CEPII's GeoDist and to 0 if we have filled it in based on past territorial configurations, as described below. *bilateral*.

3.2.3 Variable construction

Time zones:

No modification is made to the data provided by timezoneDB. Note that data corresponds to time zones as of July 2020, whereas time zones have often changed in the past: the time zone reported in *Gravity* for a given country in a given year does not correspond to the time zone in that year, but rather to the 2020 time zone. Further, we use standard time zones, i.e. not daylight saving time zones. For countries no longer existing, we add the time zone corresponding to the current time zone of their capital city. For countries with more than one time zone, we choose the timezone of their capital city.

Geographic data on contiguity and distances from CEPII:

Geographic data on contiguity and distances is taken from the CEPII's GeoDist database. When merging this dataset with *Gravity*, we checked that the geographic configurations of countries identified in both datasets were consistent, to ensure that contiguity and distance variables were properly attributed.

However, some countries have different territorial configurations in *Geodist* and *Gravity*, in particular among the countries affected by territorial changes (see Section 2 and Section A.2 of the Appendix). For a subset of these countries, we were able to fill in missing distance data (*dist*, *distcap distw*, *distwces*) with those of nearby countries that are or were part of the same country group.[2] Note that we only made adjustments for the variables *dist*, *distcap*, *distw* and *distwces*, leaving as missing data for *contig*. The only countries affected by territorial changes that remain with missing distance data are thus North Vietnam and Sint Marteen. The variable

[2]In particular, we made the following adjustments. We filled in Czechoslovakia's distance data with distance data from the Czech Republic. We filled in East Germany's distance data with distance data from Germany. We filled in USSR's distance data with distance data from Russia. We filled in Montenegro, Serbia and Serbia and Montenegro's distance data with distance data from Yugoslavia. Both current Sudan and South Sudan are assigned distance data corresponding to former Sudan, before the split of South Sudan in 2011. Both current (unified) Yemen and South Yemen are assigned distance data belonging to the former North Yemen.

dist_source helps to identify cases in which data is taken directly from the *GeoDist* database (when *dist_source* = 1), and cases in which we have replaced missing distance variables as described above (*dist_source* = 0).

3.3 Cultural variables

3.3.1 Data Sources

Data on common languages and on colonization (*comlang_off*, *comlang_ethno*, *comcol*, *col45*) comes from CEPII's GeoDist dataset. Since this dataset will no longer be updated in the future, these variables will also no longer be updated.

Data on the historical origin of a country's legal system (*legal_old*, *legal_new*, *comleg_pretrans*, *comleg_posttrans*, *transition_legalchange*) is taken from LaPorta et al. (1999) and LaPorta et al. (2008). Data on the share of religion by country (*cat*, *mus*, *pro*, *oth*) also comes from LaPorta et al. (1999).

Data on colonial ties (*heg*, *col_dep_ever*, *col_dep*, *col_dep_end_year*) is mostly derived from Head et al. (2010). We use the CIA World Factbook and Wikipedia to add data for countries not included in the original Head et al. (2010) dataset. We also compare Head et al. (2010)'s data with the Territorial Change (v6) dataset of the Correlates of War Project (COW) (Jaroslav et al. 1998). Whenever there are inconsistencies between these two sources, we use Wikipedia and the CIA World Factbook to make corrections. Data on sibling relationships (*empire*, *sibling_ever*, *sibling*, *sever_year*, *sib_conflict*,) is constructed based on data on colonial ties and independence of colonies from the above sources. Data used for *col_dep_end_conflict* is instead taken exclusively from the COW dataset on territorial changes.

3.3.2 Variables

- ***comlang_off***: Dummy equal to 1 if countries share common official or primary language, *bilateral*.
- ***comlang_ethno***: Dummy equal to 1 if countries share a common language spoken by at least 9% of the population, *bilateral*.
- ***comcol***: Dummy equal to 1 if countries share a common colonizer post 1945, *bilateral*.
- ***col45***: Dummy equal to 1 if countries are or were in colonial relationship post 1945, *bilateral*.
- ***legal_old***: Indicates historical origin of a country's laws, before transition of some countries from the Socialist legal system following the fall of the Soviet Union. It takes values of either fr (French), ge (German), sc (Scandinavian), so (Socialist), uk (English), *unilateral*.

- ***legal_new***: Indicates historical origin of a country's laws, after the above-mentioned transition, *unilateral*. Possible values are the same as for *legal_old*.
- ***comleg_pretrans***: Dummy equal to 1 if countries share common legal origins before transition, *bilateral*.
- ***comleg_pretrans***: Dummy equal to 1 if countries share common legal origins before transition, *bilateral*.
- ***comleg_posttrans***: Dummy equal to 1 if countries share common legal origins after transition, *bilateral*.
- ***transition_legalchange***: Dummy equal to 1 if common legal origin has changed since the above-mentioned transition, *bilateral*.
- ***comrelig***: Religious proximity index (Disdier and Mayer 2007): obtained by adding the products of the shares of Catholics, Protestants and Muslims in the exporting and importing countries. It is bounded between 0 and 1, and is maximum if the country pair has a religion which (1) comprises a vast majority of the population, and (2) is the same in both countries.
- ***heg_o***: Dummy equal to 1 if origin is current or former hegemon of destination, *bilateral*.
- ***heg_d***: Dummy equal to 1 if destination is current or former hegemon of origin, *bilateral*.
- ***col_dep_ever***: Dummy equal to 1 if country pair was ever in colonial relationship. This variable also takes into account colonial relationships before 1948 and is a *bilateral* variable.
- ***col_dep***: Dummy equal to 1 if country pair currently in colonial or dependency relationship, *bilateral*.
- ***col_dep_end_year***: Independence date from hegemon of the time, if pair was ever in a colonial or dependency relationship (*col_dep_ever* is equal to 1). Missing if the pair never was in a colonial or dependency relationship (*col_dep_ever* $= 0$). This variable also takes into account colonial relationships before 1948 and is a *bilateral* variable.
- ***col_dep_end_conflict***: Dummy equal to 1 if independence involved conflict and if *col_dep_ever* is equal to 1. Missing if the pair never was in a colonial or dependency relationship (*col_dep_ever* $= 0$). This variable also takes into account colonial relationships before 1948 and is a *bilateral* variable.
- ***sibling_ever***: Dummy equal to 1 if pair ever in sibling relationship (i.e. they ever had the same hegemon). This variable also takes into account colonial relationships before 1948 and is a *bilateral* variable.

- ***sibling***: Dummy equal to 1 if pair currently in sibling relationship (i.e. they have the same hegemon), *bilateral*.

- ***empire***: Hegemon if *sibling* equal to 1 for the time that country i and j are in current sibling relationship (i.e. *year* is smaller than *sever_year*), *bilateral*.

- ***sever_year***: Severance year for pairs if *sibling_ever* is equal to 1. Severance year corresponds to the year in which the first sibling in the sibling relationship became independent. This variable also takes into account colonial relationships before 1948 and is a *bilateral* variable.

- ***sib_conflict***: Dummy equal to 1 if pair ever in sibling relationship (*sibling_ever* = 1) and their independence from the hegemon involved conflict with hegemon. This variable also takes into account colonial relationships before 1948 and is a *bilateral* variable.

3.3.3 Variable construction

Colonisation and shared language data from CEPII:

Data from the CEPII's GeoDist dataset is directly added to the *Gravity* dataset. Note again that since this dataset will no longer be updated in the future, these variables will also no longer be updated.

Some adjustments are made. In particular, the variable *col45* is set to missing for Taiwan and Hong Kong with respect to China, and for Palestine with respect to Israel. Also, note that some countries are missing from *Geodist* and therefore have missing values for the corresponding variables.

Nevertheless, for a subset of countries affected by territorial configurations, we were able to fill in missing data on colonial relationships and shared languages (*comlang_off*, *comlang_ethno comcol*, *col45*) with those of nearby countries that are or were part of the same country group.[3]

Historical origin of a country's laws and religion shares from LaPorta et al. (1999) and LaPorta et al. (2008) :

As mentioned, data on countries' legal origin is available form La Porta and coauthors in two different versions. While LaPorta et al. (1999) is the original version of the dataset, LaPorta et al. (2008) contains major changes to the origins of countries' legal systems for the countries that used to have "socialist" legal origin and that switched to different legal structures after gaining independence from the Soviet Union. To distinguish pre- and post- transition values,

[3]In particular, we made the following adjustments. We filled in Czechoslovakia's data with data from the Czech Republic. We filled in East Germany's data with data from Germany. We filled in USSR's data with data from Russia. We filled in Montenegro, Serbia and Serbia and Montenegro's data with data from Yugoslavia. Both current Sudan and South Sudan were assigned data corresponding to former Sudan, before the split of South Sudan in 2011. Both current (unified) Yemen and South Yemen were assigned data belonging to the former North Yemen. The only countries affected by territorial changes that remain with missing data for these variables are thus North Vietnam and Sint Marteen.

we include two different variables that describe the historical origin of a country's legal system, "old" and "new", where "new" refers to the post-transition period.

Note that there are 29 countries with no legal origin data in the original dataset because they are not included in LaPorta et al. (1999) and LaPorta et al. (2008). However, for countries affected by territorial changes, we fill in missing data with data available for their respective country group.[4] Furthermore, we make some small corrections on the original dataset[5].

In addition, we take the original variables containing information on origin of countries' legal systems and we construct some additional variables. We add *comleg_pretrans*, which is equal to 1 if *legal_old_o* = *legal_old_d* before the transition of some countries away from Socialist regimes, and 0 otherwise. We add *comleg_posttrans*, which is equal to 1 if *legal_old_o* = *legal_old_d* after the transition of some countries away from Socialist regimes, and 0 otherwise. Finally, we add the dummy named *transition_legalchange*, which is equal to 1 if this transition has led to a change from common to different legal system, i.e. *comleg_pretrans* differs from *comleg_posttrans*, and 0 otherwise.

Regarding religion shares, we simply add the data to the *Gravity* dataset, as made available from LaPorta et al. (1999). Note that there are 46 countries with no religion data in the original dataset. However, for countries affected by territorial changes, we fill in missing data with data available for their respective country group, whenever possible.[6]

Colonial ties, based on Head et al. 2010 and COW:

Variables describing colonial ties and dependencies (*heg*, *col_dep_ever*, *col_dep*, *col_dep_end_year*) are constructed based on data from Head et al. (2010) as the main source, except for *col_dep_end_conflict* which is based exclusively on COW data. In particular, data from Head et al. (2010) identifies hegemon-colony or hegemon-dependency pairs and the independence date of colonies or dependencies. As the universe of coverage of Head et al. (2010) is less complete than that used for *Gravity*, we supplement the dataset using the CIA World Factbook and Wikipedia. Additionally, we use the COW dataset to check the consistency of independence dates. When we find disagreements between Head et al. (2010) and the COW data, we refer to the CIA World Factbook or to Wikipedia to input the independence date. We then use variables on colonial and dependency ties to construct variables on sibling relationships between countries.

Note that in some occasions colonies refer to past territorial configurations. For example, when the hegemon is TUR and the independence date is before 1923, TUR refers to the Ottoman Empire, which ceased to exist in 1923 (when the Republic of Turkey was established). Similarly, when the hegemon is AUT and the independence date is before 1918, AUT refers to the Austro-

[4] In particular, we fill in Czechoslovakia, North Vietnam and the Soviet Union data with *legal_old* = so (Socialist system). We fill Serbia and Sint Marteen data with *legal_new* = fr (French system). We fill in South Sudan with *legal_new* = uk (British system). We fill in South Yemen with *legal_old* = fr and *legal_new* = fr.

[5] In particular, we changed the origin of legal systems to French for French Guiana and French Polynesia. We also changed Northern Mariana Islands' origin of legal system to British, as it is under the US system, which has British origin. We added Serbia and Montenegro and included a French origin for its legal system

[6] In particular, we fill East Germany data with data from Germany. We fill in data for Sint Marteen with data from the Netherlands Antilles. We fill in data for South Yemen with data for Yemen.

Hungarian Empire, and when the hegemon is DEU and the independence date 1918, DEU refers to the German Empire.

Further, note that *col_dep_end_year* indicates independence from the coloniser (or hegemon in case of dependencies) at the time. This means that, in the case of countries with more than one coloniser in the past, such as Burundi, *col_dep_end_year* of 1918 from the German Empire does not correspond to the indepence date of Burundi, since Burundi later became a colony of Belgium. Instead, *col_dep_end_year* of 1962 from Belgium marks the independence of Burundi. Similarly, some territories that are currently dependencies of other colonies, such as Cocos Island, may have non-missing *col_dep_end_year* although this refers to independence from previous colonisers/countries.

We also include dependency ties for past Dominions of the British Empire which were not included in the original version of the dataset. The word Dominion was used from 1907 to 1948 to refer to one of several self-governing colonies of the British Empire. "Dominion status" was formally accorded to Canada, Australia, New Zealand, Newfoundland, South Africa, and the Irish Free State (modern Ireland). India, Pakistan and Sri Lanka were also dominions for a short period of time.[7]

Finally, for Taiwan and Hong Kong all variables on colonial or dependency ties with respect to China (and the corresponding variables on sibling relationships) are set to missing, as we do not take a stance on the relationship that these countries have with China. Similarly for all variables on colonial or dependency ties between Palestine and Israel. These observations are the only ones for which both countries exist but colonial ties variables are nevertheless missing.

Figure 3 shows the share of existing country pairs in colonial or dependency relationship over time, i.e. where $col\_dep = 1$.

3.4 Macroeconomic Indicators

3.4.1 Data Sources

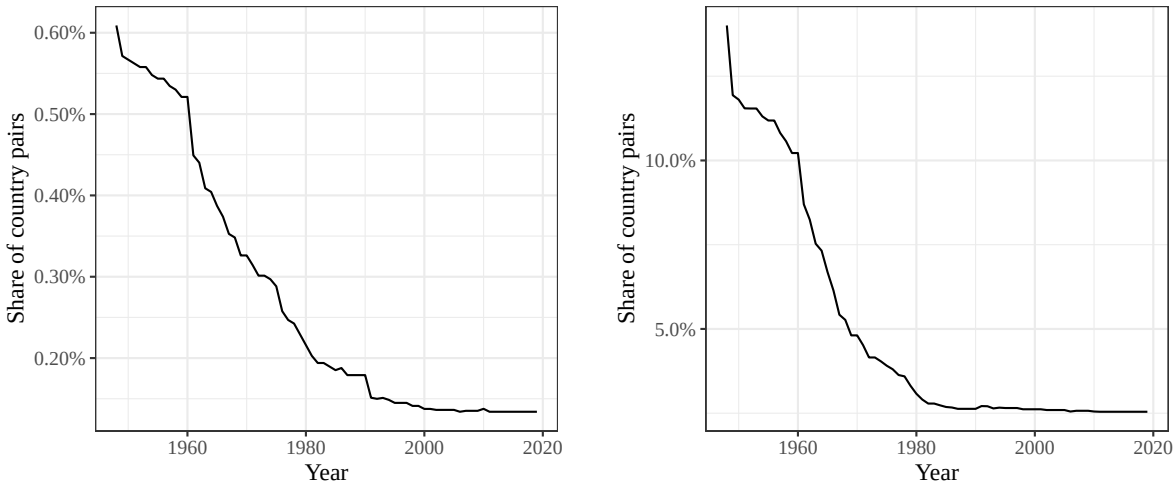
The main data source for GDP and population data (*pop*, *gdp*,*gdpcap*, *gdp_ppp*, *gdpcap_ppp*) is the World Bank's Development Indicators (WDI). However, WDI data does not include former countries but only refers to up-to-date territorial configurations. Further, it does not cover years prior to 1960. Therefore, we rely on two alternative sources for observations for which no WDI data is available:

1. For GDP: Katherine Barbieri's International Trade Dataset, which contains GDP figures for the period 1948-1992 (Barbieri 2005). In particular, Barbieri contains GDP for East and West Germany.

2. For population: Angus Maddison's Statistics on World Population, GDP and Per Capita GDP, 1-2008 AD (Horizontal file, copyright Angus Maddison, university of Groningen),

[7] Since Bangladesh was part of Pakistan until 1971, we also include Bangladesh as former colony of the British Empire.

Figure 3: Share of existing country pairs in colonial relationship



(a) Metropole-colony

(b) Siblings

Notes: The number of country pairs in a metropole-colony relationship in a given year is the number of country pairs for which $col\_dep = 1$. The share of existing country pairs in a metropole-colony relationship is the ratio between this number and the number of existing country pairs in a given year. Siblings are country pairs that share a common colonizer. The share of sibling country pairs is the ratio between this number and the number of existing country pairs in a given year. We exclude cases in which $iso3\_o = iso3\_d$.

both in its previous version and in the version updated as of 2010 that is currently available on the website of Groningen Growth and Development Centre.

The variables *gdp_source* and *pop_source* indicate the sources of each datapoint (WDI, Maddison, Barbieri or Taiwan Govt[8]).

As an alternative source of GDP and population data, we use the Penn World Tables (PWT) version 9.1 (Feenstra et al. 2015).[9]. In particular, we use PWT to obtain data on GDP in PPP and population (*gdp_ppp_pwt* and *pop_pwt*). PWT data also does not include former countries but only refers to current territorial configurations.

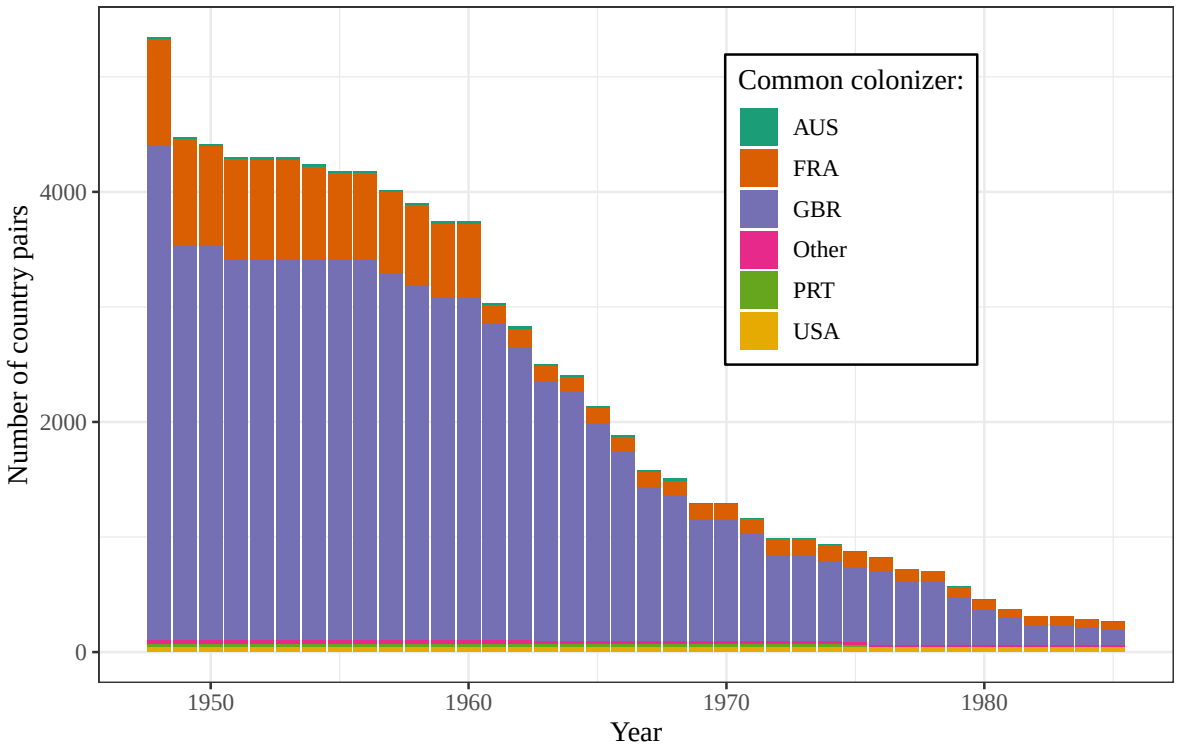
3.4.2 Variables

- ***pop***: Population, in thousands (source WDI/Maddison), *unilateral*.
- ***gdp***: GDP, in current thousands US$ (source WDI/Barbieri), *unilateral*.
- ***gdpcap***: GDP per cap, in current thousands US$ (source WDI/Barbieri), *unilateral*.

[8]Since Taiwan does not have data in the WDI, we use data from its national statistical agency (downloaded on 11/08/2020). This data is available from 1951, hence we complement it with Maddison's population data which is available for 1947-1950 (Barbieri historical GDP data is not available for Taiwan before 1951).

[9]This dataset was downloaded on 11/08/2020

Figure 4: Siblings: number of country pairs, by common colonizer (1948-1985)



Notes: Siblings are country pairs that share a common colonizer. We exclude cases in which $iso3\_o = iso3\_d$.

- ***gdp_ppp***: GDP PPP, in current thousands international $ (source WDI), *unilateral*.
- ***gdpcap_ppp***: GDP per cap PPP, in current thousands international $ (source WDI), *unilateral*.
- ***pop_pwt***: Population, in thousands (source PWT), *unilateral*.
- ***gdp_ppp_pwt***: GDP, current PPP, in 2011 thousands US$ (source PWT), *unilateral*.
- ***gdp_source***: Source of GDP data: 1 = WDI, 2 = Barbieri, 3 = Taiwan Govt), *unilateral*.
- ***pop_source***: Source of population data: 1 = WDI, 2 = Maddison, 3 = Taiwan Govt, *unilateral*.

3.4.3 Variable construction

WDI, Barbieri and Maddison data:

For GDP and population data, WDI is the main source. When WDI data is available, we use it to create the variables *gdp* and *pop*. If there is no WDI data on GDP, Barbieri's data is used to complement GDP data. Similarly, if no WDI population data is available, we use Maddison as an alternative source of population data. The variables *gdp_source* and *pop_source* identify the sources of data for the origin and the destination country in any given year. We then use data on GDP and population obtained from the above mentioned sources to compute GDP per capita, in current US $ (*gdpcap*). Thus, *gdp_source* and *pop_source* also identify the sources used to construct *gdpcap*. WDI also contains data on GDP in PPP (current international $, hence not deflated) and its per capita version (used to construct variables *gdp_ppp* and *gdpcap_ppp*). This data is not augmented with Maddison or Barbieri data.

To ensure that GDP and population data matches the dynamic nature of the *Gravity* dataset, and thus takes into account territorial changes, we occasionally have to aggregate data on countries (for instance in the case of Yugoslavia and the Soviet Union), or set some data to missing. Section A.3 in the Appendix describes in detail these modifications and aggregations. Note that in years in which territorial changes occur, only in some cases we are able to include GDP and population data both for countries in their first year of existence and countries in their last year of existence.[10]

Penn World Tables data:

We also include data on population and GDP measured at current PPP from the Penn World Tables (*pop_pwt* and *gdp_ppp_pwt*). For GDP, we use the expenditure-side real GDP at current PPPs, which enables to compare relative living standards across countries at a single point in time. Note that, in contrast to the GDP in PPP provided by the WDI (*gdp_ppp*), PWT GDP in PPP data is deflated. As with WDI data, to ensure that GDP and population data matches the dynamic nature of the *Gravity* dataset, and thus that it takes into account territorial changes, we occasionally have to aggregate data on countries (for instance in the case of Yugoslavia and the Soviet Union), or set some data to missing. Again, Section A.3 in the Appendix describes in detail these modifications and aggregations.

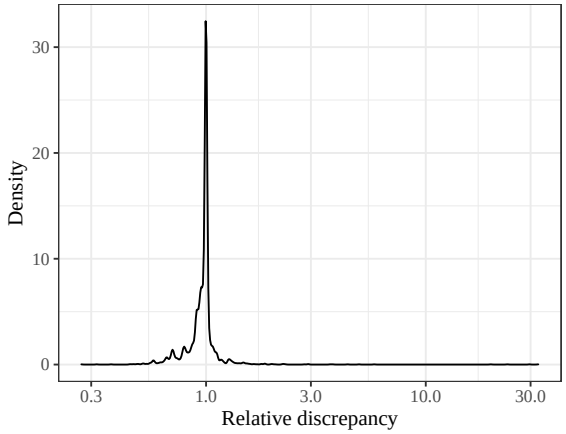
Comparison across sources:

Figure 5 plots the distribution of the ratio of the variables as reported by each source, for country years for which two sources are available. Concerning population and GDP, there are sometimes discrepancies across the data sources (especially for GDP), so users that are interested in time variations should be very cautious when the variables *gdp_source* and *pop_source* indicate a change in the source.

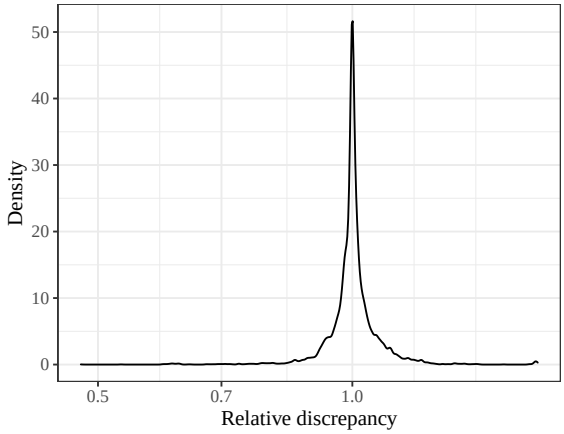
[10] In particular, we are not able to include GDP and population data for DEU.1 in 1990, YEM.1 in 1990 and VNM.1 in 1976.

Figure 5: Relative discrepancy across sources.

(a) GDP Barbieri/WDI



(b) Population Maddison/WDI



Notes: We compute the ratio of the variable as reported by each source for country-years for which both data sources are available, and plot the distribution of this ratio (kernel density).

3.5 Trade facilitation variables

3.5.1 Data Sources

Data on GATT and WTO membership (*gatt*, *wto*) is taken from the WTO. Data on EU membership (*eu*) is derived from information made available by the European Union, which provides data on its members and their accession dates.

Data on Regional Trade Agreements (RTAs) (*rta*, *rta_type*, *rta_coverage*) is taken from the WTO's (2020) "Regional Trade Agreements Information System (RTA-IS)".[11] For each RTA, this dataset lists the RTA name, the coverage (whether it's goods, services or both), the type of RTA, the date of entry into force for the part on goods and for the part on services (the two may differ), the original signatories, and specific entry or exit dates for additional signatories.

Data on entry costs (*entry_cost*, *entry_proc*, *entry_time*, *entry_tp*) is taken from the World Bank's Development Indicators (WDI).

Users wishing to add data on common currencies (*comcur*) can download De Sousa's Currency Unions dataset (available for the years 1948-2015) updated on December 2014 (de Sousa 2012).

3.5.2 Variables

- ***gatt***: Dummy equal to 1 if country is a GATT member in a given year, *unilateral*.
- ***wto***: Dummy equal to 1 if country is a WTO member in a given year, *unilateral*.
- ***eu***: Dummy equal to 1 if country is a EU member in a given year, *unilateral*.

[11] This dataset was downloaded on 12/08/2020.

- ***rta***: Dummy equal to 1 if origin and destination country are engaged in a regional trade agreement of any type within the given year (Source: WTO), *bilateral*.
- ***rta_coverage***: Coverage of the trade agreement. 0 = "no trade agreement" ($rta = 0$). 1 = "goods only", 2 = "services only", 3 = "goods and services". Source: WTO, *bilateral*.
- ***rta_type***: Categorical variable describing type of regional trade agreement if origin and destination country are engaged in a regional trade agreement within the given year and rta = 1 (Source: WTO, see Table **??** for a description of the values taken by this variable), *bilateral*.
- ***entry_cost***: Cost of business start-up procedures (% of GNI per capita), *unilateral*.
- ***entry_proc***: Number of start-up procedures to register a business, *unilateral*.
- ***entry_time***: Days required to start a business, *unilateral*.
- ***entry_tp***: Days required to start a business + number of procedures to start a business, *unilateral*.

3.5.3 Variable construction

GATT and WTO membership:

Data on GATT and WTO membership is taken directly from the WTO,[12] without additional modifications. We kept GATT and WTO membership distinct to maintain the ability to identify those (few) cases in which countries are part of GATT but not of WTO. These cases are Lebanon and Syria for countries that currently still exist. Concerning the years of entry and exit from GATT and WTO, the dummy variables are equal to one if countries enter or exit between 1st January and 31st December of the given year.

EU membership:

Data on EU membership is constructed based on information available on the European Union website. Concerning the year of entry into the EU, the dummy variable *eu* is set to one if countries enter between 1st January and 31st December of the given year.

Regional Trade Agreements based on the WTO:

We make many adjustments to the RTA-IS dataset of the WTO, since this database is originally structured with one observation per trade agreement, and has to be converted into the origin-destination-year structure of the *Gravity* dataset. These numerous and substantial adjustments are described in the appendix, section A.4.

[12] For WTO membership, data is available directly from the following WTO page https://www.wto.org/english/thewto_e/whatis_e/tif_e/org6_e.htm. For GATT members, we take data from the following WTO page https://www.wto.org/english/thewto_e/gattmem_e.htm.

We use the following time convention: a country pairs is considered as being in a RTA in a given year as soon as the RTA was in force at least one day during this year. Also, we do not consider that countries have RTAs with themselves, i.e. RTA variables are set to missing when $iso_o = iso_d$.

The WTO distinguishes 4 types of RTAs: Partial Scope Agreements (PSA), Free Trade Agreements (FTA), Customs Union (CU) and Economic Integration Agreements (EIA). PSAs typically involve the elimination of import tariffs in only a few sectors. FTAs entail the elimination of import tariffs in most sectors but FTA members retain independent trade policies. Customs unions build on FTAs by requiring participants to harmonize their external trade policy, including establishing a common external tariff. EIAs involve the liberalization of trade in services. These types may be combined, for instance a pair of countries can be both in a customs union and in trade agreement liberalizing services. This is why the categorical variable that describes the type of RTA takes may take more than one value (such as "CU & EIA").

Also, the WTO data allows to create a variable distinguishing between RTAs that cover only goods, only services, or both goods and services ($rta\_coverage$)

Descriptive statistics:

Figure 6 shows the share of country pairs engaged in RTAs by distinguishing the coverage of the RTA (goods only, services only, or both). While RTA focused entirely on services are almost non existent, we observe since the 2000s a strong rise in the share of RTA involving liberalization of trade both in goods and services.

In figure 7, we plot types of trade agreement. Free Trade Agreements represent the bulk of RTAs, especially since the early 2000s, whereas the share of Partial Scope Agreements start declining.

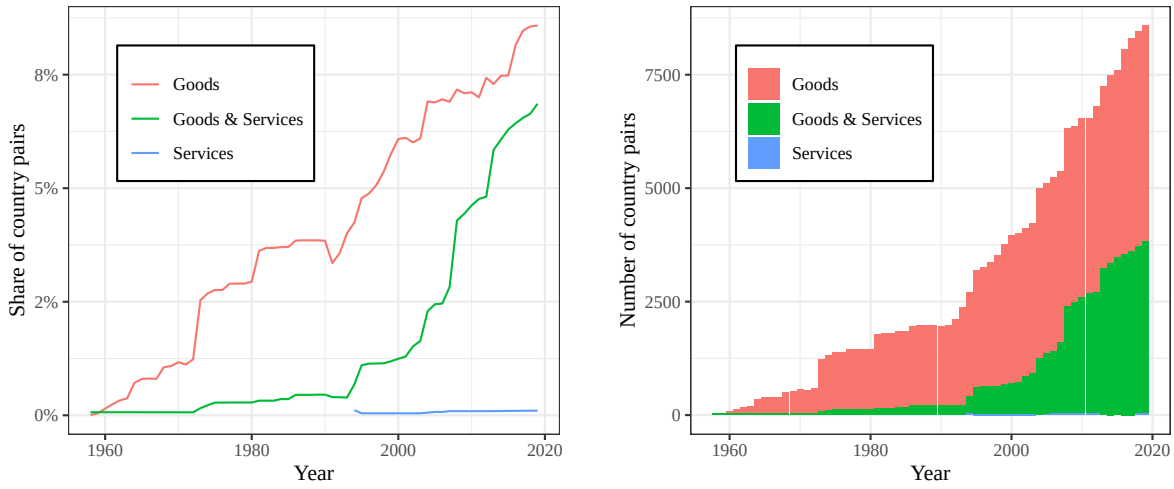
It is important to keep in mind that not all RTAs reported to the WTO are equally substantial as the effective coverage of RTAs depends on details that are not fully accounted for by the data provided to the WTO.

Comparison with Jeffrey Bergstrand's Database on Economic Integration Agreements:

An alternative source of RTA data is Jeffrey Bergstrand's dataset on Economic Integration Agreements. A comparison between this data and data constructed based on the WTO RTA-IS is instructive because the two sources differ significantly in their content and construction. While WTO variables only include trade deals officially notified to the WTO, Bergstrand's data also relies on other sources, such as the Council of the European Union and the Tuck Trade Agreements Database. Furthermore, it includes data on Generalized System of Preference agreements (GSPs), which are generally not included in the WTO RTA-IS dataset. Finally, WTO may include some agreements that are not substantial in practice. To construct *Gravity*, we do not analyse the depth of trade agreements, but rather chose to include all of the agreements listed by the WTO (with some exceptions detailed in the appendix, section A.4).

In figure 8, we compare the share of country pairs covered by a RTA between the two sources (WTO and Bergstrand). To improve comparability between the two series, we remove from the Bergstrand count the country pairs for engaged in a non-reciprocal Preferential Trade

Figure 6: RTA coverage over time (1958-2019), source: WTO



(a) Share (b) Number

Notes: The share of country pairs is the ratio "number of existing country pairs in a given RTA coverage in a given year"/"number of country pairs existing in a given year". Before 1958 there are virtually no country pairs involved in RTAs. We exclude cases in which $iso3\_o = iso3\_d$.

Agreements, as these RTAs are not accounted for in the WTO RTA-IS dataset. Figure 9 displays the share and the number of countries engaged in RTAs by type of RTA based on data from Bergstrand's Database on Economic Integration Agreements.

Entry Costs:

Data on entry costs for the variables *entry_cost*, *entry_proc*, and *entry_time* is taken directly from the World Bank Development Indicators API.[13] Note that the earliest year available from the WDI dataset on entry costs is 2003, and data is available up to 2019. Further, no data is available for Taiwan.

Based on the data taken from the WDI, we construct one additional variable *entry_tp* which is the sum of *entry_proc* and *entry_time*.

3.6 Trade flow variables

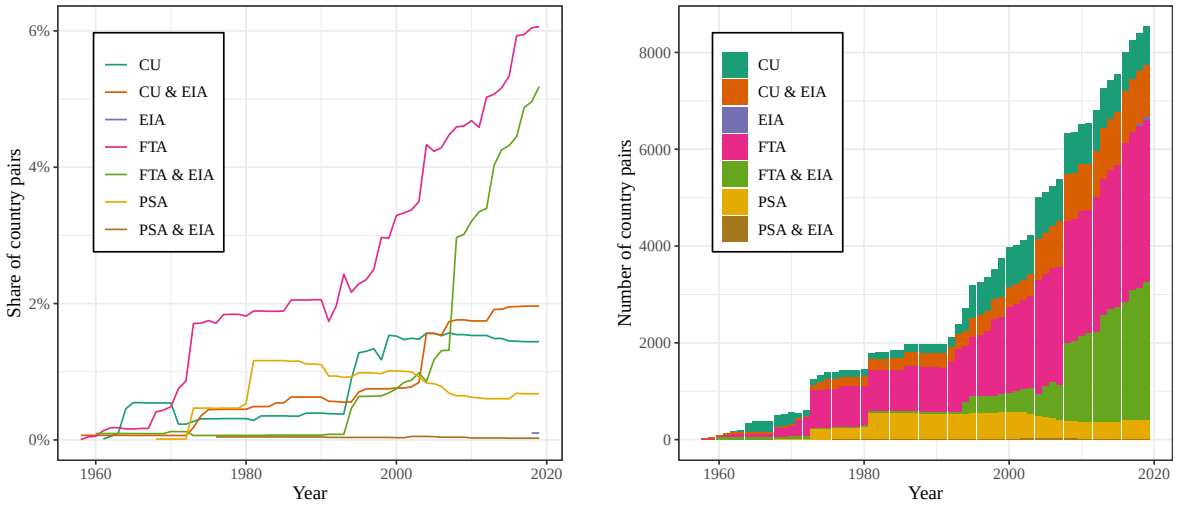
3.6.1 Data Sources

We provide trade flows data from three sources: the CEPII's BACI database, the UNSD's Comtrade data and the IMF's DOTS data.

UN Statistics Division data (*tradeflow_comtrade_o*, *tradeflow_comtrade_d*) is available via Comtrade. We download the bulk files in the first revision of the SITC product nomenclature,

[13] In particular, we take the following indicator codes IC.REG.COST.PC.ZS, IC.REG.PROC, IC.REG.DURS. See here for more details on the variables available.

Figure 7: RTA types over time (1958-2019), source: WTO



(a) Share (b) Number

Notes: PSA = "Partial Scope Agreement", FTA = "Free Trade Agreement", CU = "Customs Union", EIA = "Economic Integration Agreement". Before 1958 there are virtually no country pairs involved in RTAs. Shares of country pairs have the number of existing countries in a given year as denominator. We exclude cases in which $iso3\_o = iso3\_d$.

which provides the longest time coverage (from 1962 onwards).

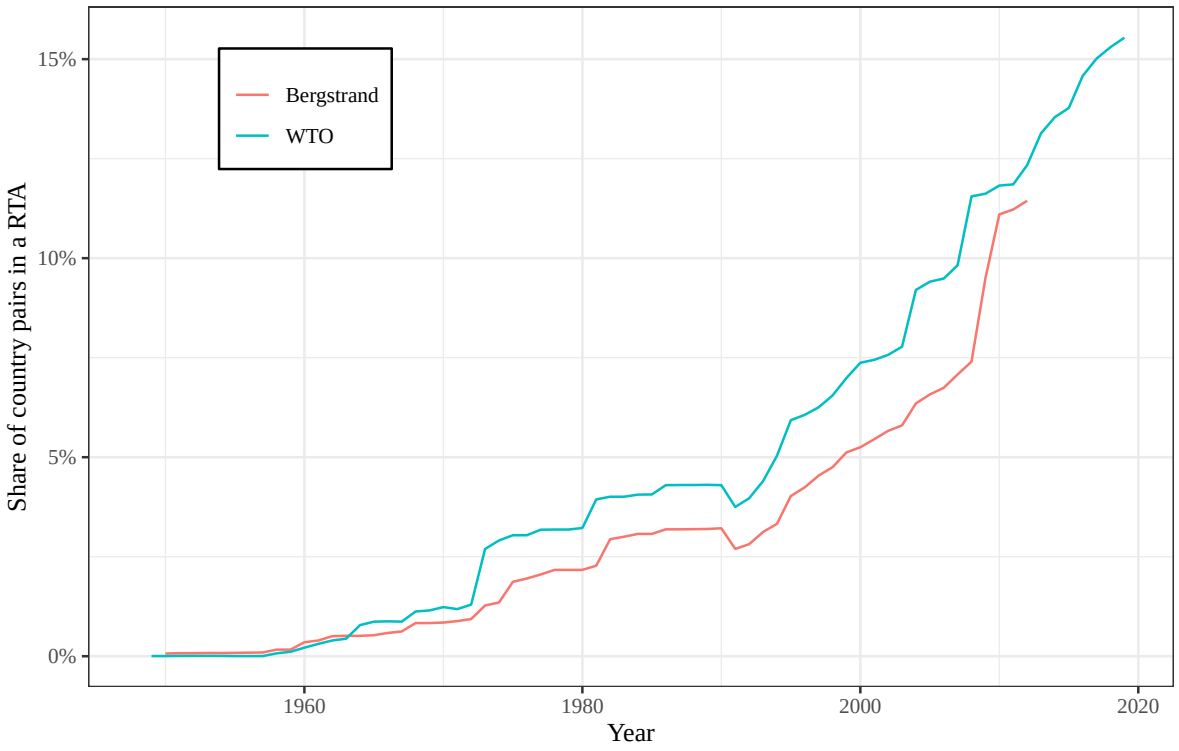
BACI trade flows (*tradeflow_baci*, *manuf_tradeflow_baci*) are taken from the April 2020 version of BACI (version 202004a). BACI provides a single harmonized trade flow for each exporter-importer-year, by reconciling Comtrade mirror flows. However, its time coverage is more limited than Comtrade since it is available only from 1996 onwards.

IMF data (*tradeflow_imf_o*, *tradeflow_imf_d*) is provided by the Direction of Trade Statistics (DOTS). The DOTS database contains official trade data reported by country authorities to the IMF, or collected by the IMF from official sources. For European countries, data is obtained from the COMEXT database maintained by Eurostat, while data from UN Comtrade is used for countries that do not report to the IMF. Official data is complemented with estimated data for individual countries that report (or publish) trade statistics with a delay, or do not publish trade statistics by partner country at all. Estimates for these countries are based on data reported by their trading partners or, when these are also unavailable, on their total level of exports and imports. Also note that IMF DOT data does not include former countries but only refers to current territorial configurations. We download bulk files of yearly trade flows.

3.6.2 Variables

- ***tradeflow_comtrade_o***: Trade flows as reported by the origin, 1000 current USD (Source: Comtrade), *bilateral*.

Figure 8: Share of country pairs engaged in RTAs, comparison WTO/Bergstrand.



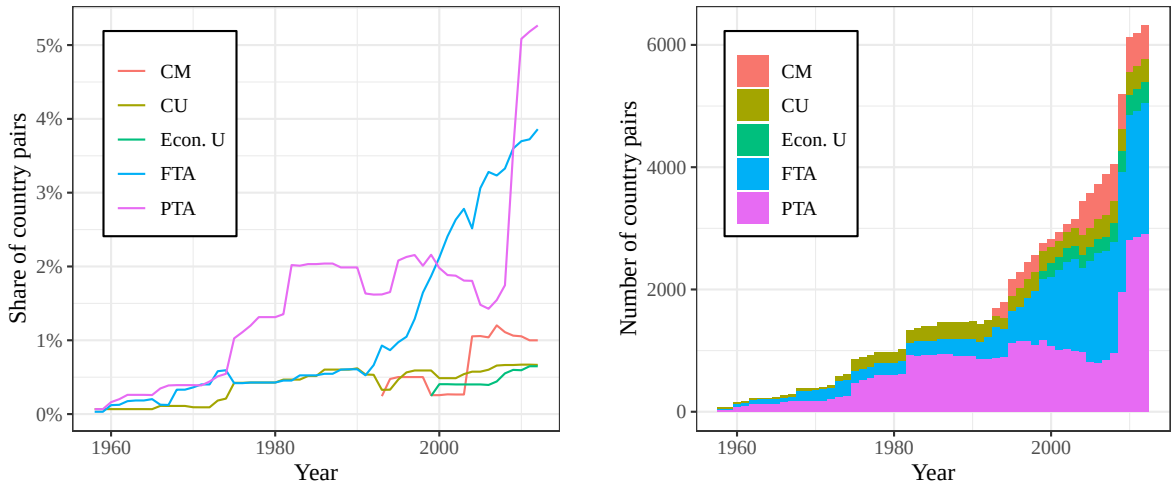
Notes: The share of country pairs indicates the ratio "number of existing country pairs engaged in an RTA in a given year"/"number of country pairs existing in a given year". For Bergstrand RTA data, we exclude Non Reciprocal Preferential Trade Arrangement (that are not included in the WTO dataset). In other words, we only count country pairs where *rta_bergstrand* = 1 and *rta_type_bergstrand is greater than 1*

- ***tradeflow_comtrade_d***: Trade flows as reported by the destination, 1000 current USD (Source: Comtrade), *bilateral*.
- ***tradeflow_baci***: Trade flow, 1000 current USD (Source: BACI), *bilateral*.
- ***manuf_tradeflow_baci***: Trade flow of manufactured goods, 1000 current USD (Source: BACI), *bilateral*.
- ***tradeflow_imf_o***: Trade flows as reported by the origin, 1000 current USD (source: IMF), *bilateral*.
- ***tradeflow_imf_d***: Trade flows as reported by the destination, 1000 current USD (source: IMF), *bilateral*.

3.6.3 Variable construction

UNSD's Comtrade:

Figure 9: RTA types over time (1958-2019), source: EIA database.



(a) Share (b) Number

Notes: PTA = " Two-Way Preferential Trade Agreement", FTA = "Free Trade Agreement", CU = "Customs Union", CM = "Common Market", Econ. U = "Economic Union". Shares of country pairs have the number of existing countries in a given year as denominator. We exclude cases in which $iso3\_o = iso3\_d$.

Comtrade data has a "reporter-partner" structure, meaning that each reporting country indicates how much it trades with each of its partner countries, both as exports and as imports. We reshape the data to fit the "origin-destination" structure of the Gravity dataset. For instance, a trade flow reported by France as exports towards its partner Germany will become a flow from France to Germany, as reported by the origin (i.e. it will appear in the *tradeflow_comtrade_o* variable). Note that trade flows reported by the exporter are FOB (Free on Board), while trade flows reported by the importer are CIF (i.e. they include Cost, Insurance and Freight).

CEPII's BACI:

The BACI data we use is originally expressed in the 1996 revision of the Harmonized System Nomenclature (henceforth HS). To single out manufactured goods, we need to incorporate information from two other nomenclatures: ISIC and BEC. Conversion from HS to ISIC is made using the CN2020-CPA2.1 correlation table provided by Eurostat, available on RAMON. Conversion from HS to BEC is made using the correlation table provided by the UNSD. Manufactured goods are then identified as goods that do not belong to the division 01 (agriculture) of the ISIC classification (revision 3.1), nor to the Primary Goods categories of the BEC classification (revision 4).

IMF DOTS:

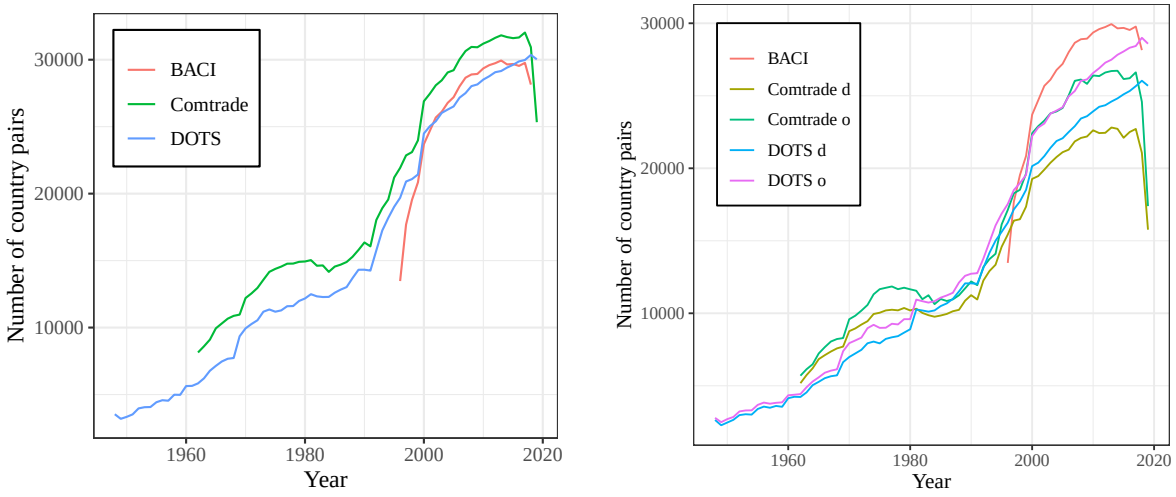
DOTS data also has a "reporter-partner" structure, meaning that each reporting country indicates how much it trades with each of its partner countries, both as exports and as imports.

As with Comtrade data, we reshape the data to fit the "origin-destination" structure of the Gravity dataset. Note that trade flows as reported by the exporter are FOB (Free on Board), while trade flows reported by the importer are CIF (i.e. they include Cost, Insurance and Freight). Also note that IMF DOTS data does not take into account the territorial changes that affected some countries.

Comparison across sources:

Figure 10 compares the coverage of each of these three sources, by counting the number of country pairs for which a strictly positive trade flow is recorded. This number strongly increases over the time period included in *Gravity*, suggesting an improvement in the coverage of trade flow data. However, note that this increase also reflects a decrease in the number of country pairs that do not trade.

Figure 10: Number of country pairs with trade flow data, by source.

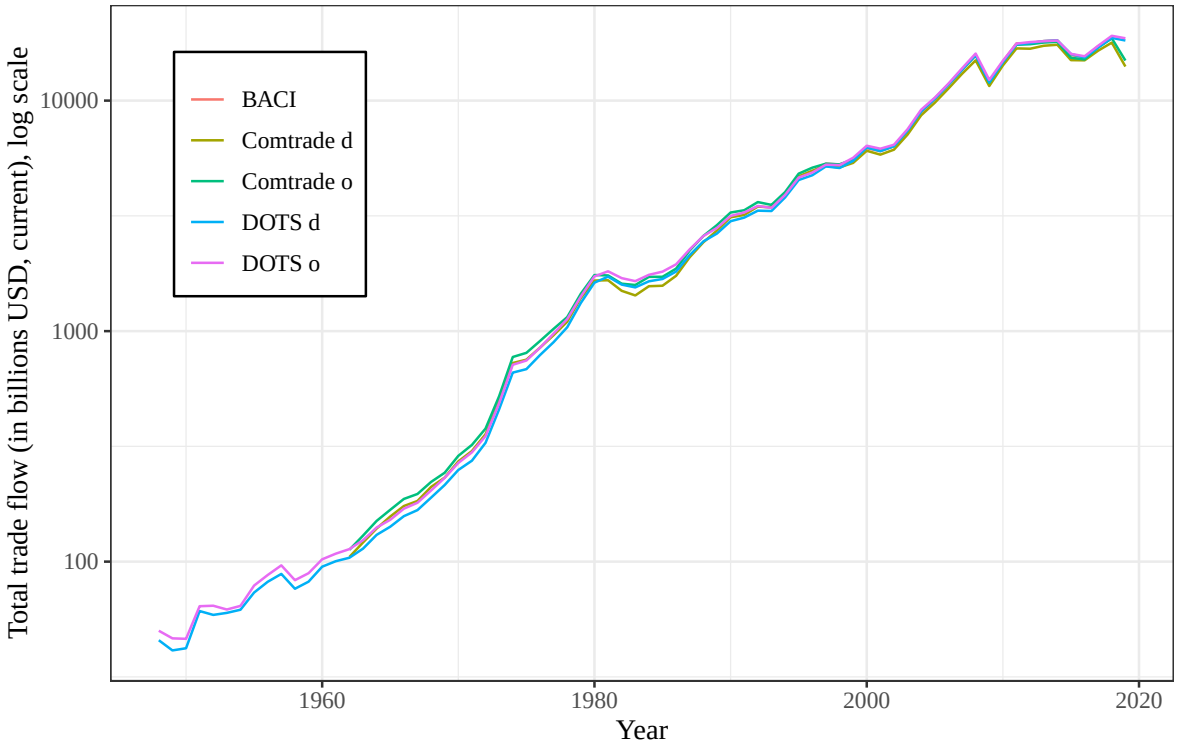


(a) (b)

Notes: (a) "Comtrade" refers to trade flows available in Comtrade, i.e. reported by at least one of two countries (either the importer, or the exporter, or both). Similarly, for DOTS, we count the number of trade flows reporter at least one of the two countries. We exclude cases in which $iso3_o = iso3_d$. (b) "Comtrade d" refers to trade flows reported by the destination (importer) in Comtrade, while "Comtrade o" refers to trade flows reported by the origin (exporter) in Comtrade. Similarly for trade flows reported in DOTS, we distinguish between those reported by the destination (importer) and those reported by the origin (exporter). We exclude cases in which $iso3_o = iso3_d$.

Figure 11 reports total trade flows summed across country pairs over each year, for each of the five sources in the *Gravity* dataset. Despite some slight differences, the overall picture is that there is no major divergence across sources in terms of coverage of the most important trade flows (in value).

Figure 11: Total trade flows, by source.



Notes: "Comtrade d" refers to trade flows reported by the destination/importer in Comtrade, while "Comtrade o" refers to trade flows reported by the origin/exporter in Comtrade. Similarly for trade flows reported in DOTS, we distinguish between those reported by the destination/importer and those reported by the origin/exporter.

Table 2: Sources used to construct the *Gravity* dataset

Source	**Variables created based on source**
CEPII's GeoDist (geo_cepii.dta), Mayer and Zignago (2011)	Country coverage, *comlang_off*, *comlang_ethno*, *comcol*, *col45*
World Bank's World Integrated Trade Solution	Country coverage
ISO	*country*, *iso3*, *iso3num*
CIA World Factbook	*first_year*, *last_year*, *country_exists*, *countrygroup_iso3*, *countrygroup_iso3num*
Wikipedia	*country*, *countrylong*, *countrygroup_iso3*, *countrygroup_iso3num*
UN World Urbanisation Prospect 2018, Wikipedia, https://latitude.to/, World Bank Surface Area, NASA's Urban-Rural Population and Land Area Estimates (v2)	*distw_harmonic*, *distw_arithmetic*, *dist*, *main_city_source*, *distcap*
Julian Hinz's Gravity distances, Hinz (2017).	*distw_harmonic_jh*, *distw_arithmetic_jh*
ARCGIS's World Countries (Generalized) dataset	*contig*
timezoneDB	*gmt_offset_2020*
United Nations General Assembly Voting data, Bailey et al. (2017).	*diplo_disagreement*
Social Connectedness Index data, Bailey et al. (2018).	*scaled_sci_2021*
LaPorta et al. (1999) and LaPorta et al. (2008)	*legal_old*, *legal_new*, *comleg_pretrans*, *comleg_posttrans*, *transition_legalchange*
LaPorta et al. (1999). Please cite this paper.	*comrelig*
Head et al. (2010) (please cite) and Correlates of War Project (Territorial Change, v6)	*heg*, *col_dep_ever*, *col_dep*, *col_dep_end_year*, *col_dep_end_conflict*, *empire*, *sibling_ever*, *sibling*, *sever_year*, *sib_conflict*
World Bank's Development Indicators, Barbieri (2005), Angus Maddison's Statistics on World Population, Taiwan's national statistical agency	*pop*, *gdp*,*gdpcap*, *gdp_ppp*, *gdpcap_ppp*
Penn World Tables version 9.1	*gdp_ppp*, *gdpcap_ppp*
WTO	*wto*, *gatt*
European Union	*eu*
WTO's Regional Trade Agreements Information System	*rta*, *rta_type*, *rta_coverage*
World Bank's Development Indicators	*entry_cost*, *entry_proc*, *entry_time*, *entry_tp*
CEPII's BACI	*tradeflow_baci*, *manuf_tradeflow_baci*
IMF's Direction of Trade Statistics	*tradeflow_imf_o*, *tradeflow_imf_d*
UN Comtrade	*tradeflow_comtrade_o*, *tradeflow_comtrade_d*

References

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, *3*(32), 259–80.

Bailey, M., Strezhnev, A., & Voeten, E. (2017). Estimating dynamic state preferences from united nations voting data. *Journal of Conflict Resolution*, *2*(61), 430–56.

Barbieri, K. (2005). The liberal illusion: Does trade promote peace? (A. A. U. of Michigan Press, Ed.).

de Sousa, J. (2012). The currency union effect on trade is decreasing over time. *Economics Letters*, *117*(3), 917–920.

Disdier, A.-C., & Mayer, T. (2007). Je t'aime, moi non plus: Bilateral opinions and international trade. *European Journal of Political Economy*, *23*(4), 1140–1159. https://doi.org/https://doi.org/10.1016/j.ejpoleco.2006.09.021

Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2015). The next generation of the penn world table. *merican Economic Review*, *105*(10), 3150–3182.

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

Head, K., & Mayer, T. (2010). Illusory border effects : Distance mismeasurement inflates estimates of home bias in trade (P. A. G. van Bergeijk & S. Brakman, Eds.). In P. A. G. van Bergeijk & S. Brakman (Eds.), *The gravity model in international trade*. Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511762109

Hinz, J. (2017). The view from space: Theory-based time-varying distances in the gravity model. *Kiel Working Paper*, (2059).

Jaroslav, T., Schafer, P., Diehl, P., & Goertz, G. (1998). Territorial changes, 1816-1996: Procedures and data. *Conflict Management and Peace Science*, *16*, 89–97.

LaPorta, R., Lopez-de-Silanes, F., & Shleifer, A. (2008). The economic consequences of legal origins. *Journal of Economic Literature*, *46*(2), 285–332.

LaPorta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. (1999). The quality of government. *Journal of Law, Economics and Organization*, *15*(1), 222–279.

Mayer, T., & Zignago, S. (2011). Notes on cepii's distances measures: The geodist database. *CEPII Working Paper*, *25*.

A Appendix

A.1 Country codes

This section describes instances in which the ISO3 alphabetic or numeric code of countries changes over time. In particular, we distinguish between cases in which territorial entities merged and cases in which territorial entities were affected by a split.

For cases in which territorial entities merged:

- West Germany used ISO3 alphabetic code of DEU before reunification in 1990, but 280 as ISO3 numeric code. The unified Germany has ISO3 numeric code of 276 (and has kept DEU has ISO3 alphabetic code).
- Following unification of North and South Yemen, the unified country inherited the ISO3 alphabetic code of North Yemen (YEM). Before reunification, North Yemen had ISO3 numeric code of 886, which changed to 887 after reunification.
- Following the unification of North and South Vietnam in 1976, the new unified country inherited the ISO3 alphabetic code of South Vietnam (VNM). However, we could not find the ISO3 numeric code for South Vietnam before unification. Similarly, we could not find the ISO3 numeric code for North Vietnam. However, it is possible to track territorial change through the variables *countrygroup_iso3* and *countrygroup_iso3num*.

For cases in which territorial entities were affected by a split and the country continued to exist:

- Sudan used alphabetic ISO3 code of SDN and numeric code 736 before South Sudan split away in 2011. Since then, Sudan has used numeric code 729, while keeping the same alphabetic code.
- Ethiopia used alphabetic ISO3 code of ETH and numeric code 230 before Eritrea split away in 1993. Since then, Ethiopia has used numeric code 231, while keeping the same alphabetic code.
- The Netherlands Antilles used ISO3 alphabetic code of ANT and numeric code of 532 before Aruba became independent in 1986. After Aruba's independence, the Netherlands Antilles has used ISO3 numeric code of 530, while keeping the same alphabetic ISO3 code.
- Pakistan currently has alphabetic ISO3 code of PAK and numeric code of 586. However, we could not find a different numeric code for Pakistan before the independence of Bangladesh from Pakistan in 1971. However, it is still possible to track this territorial change, as Bangladesh is linked to Pakistan through the variables *countrygroup_iso3* and *countrygroup_iso3num*.

- Malaysia currently has alphabetic ISO3 code of MYS and numeric code of 458. However, we could not find a different numeric code for Malaysia before the independence of Singapore from Malaysia in 1965. As with Pakistan and Bangladesh, it is possible to track territorial change through the variables *countrygroup_iso3* and *countrygroup_iso3num*.

- Indonesia currently has alphabetic ISO3 code of IDN and numeric code of 360. However, we could not find a different numeric code for Indonesia before the independence of Timor Leste from Indonesia in 2002. As with the above cases, it is possible to track territorial change through the variables *countrygroup_iso3* and *countrygroup_iso3num*.

For cases in which territorial entities were affected by a split and the country ceased to exist:

- Yugoslavia (the Socialist Federal Republic of Yugoslavia) has ISO3 alphabetic code of YUG and ISO3 numeric code of 890. Following the split of Yugoslavia, the Federal Republic of Yugoslavia inherited the ISO3 alphabetic code of YUG, but the ISO3 numeric code of 891. The ISO3 alphabetic code of YUG existed until the Federal Republic of Yugoslavia was renamed Serbia and Montenegro in 2003, adopting the ISO3 alphabetic code of SCG, while keeping the same ISO3 numeric code of 891. We thus replace the ISO3 alphabetic code of SCG for Serbia and Montenegro to YUG after the name change in 2003.

- The USSR had ISO3 alphabetic code of SUN and numeric code of 810. After the collapse of the Soviet Union, a number of countries emerged, all with distinct ISO3 alphabetic and numeric codes.

- Czechoslovakia had ISO3 alphabetic code of CSK and numeric code of 200. After the split of Czechoslovakia in Czech republic and Slovakia, both countries adopted distinct ISO3 codes.

A further case in which the numeric ISO3 code changed concerns Panama, which gained joint control with the United States over the Panama Canal Zone in 1980. Before it had ISO3 numeric code of 590, after of 591.

In some cases we did not change the alphabetic ISO3 code as it was only officially changed due to a change in the name of the country (without a corresponding territorial change):

- In the case of Burma, which changed its name to Myanmar in 1989 without any territorial change, the numeric ISO3 remains unchanged. We also leave the alphabetic ISO3 code unchanged at MMR, although the official code is BUR before 1989.

- In the case of the Democratic Republic of the Congo, which changed its name from Zaire in 1997 without any territorial change, the numeric ISO3 remains unchanged. We also leave the alphabetic ISO3 code unchanged at COD, although the official alphabetic ISO3 is ZAR before 1997.

A.2 Territorial changes

Table 3 below describes the countries whose territorial changes are tracked in the *Gravity* and *Countries* datasets using the *country_id*, country group variables, first and last year of territorial existence, or changing ISO3 numeric codes as described in Section A.1 of the Appendix. Remember that the *countrygroup_iso3* variable is used to track the country's previous membership (in case of a split) and the country's new membership (in case of a unification of two territories). In other words, *countrygroup_iso3* indicates the largest entity of which a country was or is part of, in case of territorial change. Also, note again that in the case of Indonesia, Malaysia, Pakistan and Vietnam we could not find alternative numeric ISO3 codes denoting their territorial changes (see Section A.1 of the Appendix for more details).

In addition, remember that for Germany, Vietnam and Yemen, *first_year* refers to the first year of existence of the unified country, but the same ISO3 alphabetic is also used for West Germany, South Vietnam and North Yemen respectively. However, West Germany, South Vietnam and North Yemen exist with *iso3* of DEU, VNM and YEM respectively, *before* the *first_year* indicated in Table 3. Thus, the variable *country_exists* is set to 1 since 1948 (i.e. from the beginning of the dataset) for these 3 specific cases where *iso3* is either DEU, VNM or YEM.

Also, remember that for countries that suffered a split but continued to exist (Pakistan, Ethiopia, Malaysia, Netherlands Antilles, Sudan and Indonesia), the same alphabetic ISO3 code refers to the country before and after the split, depending on the year in which the country is observed.

Further, it is important to note that we do not track the following territorial changes:

- Guadalupe was affected by territorial change when Saint-Barthelemy and Saint-Martin were separated from it in 2007. At the moment, we exclude Saint-Barthelemy and Saint-Martin and only have Guadalupe, hence we do not track this territorial change.

- Phoenix Islands and some of the Line Islands became part of Kiribati territory by the Treaty of Tarawa. We do not account for this territorial change, because the Phoenix Islands are not in the dataset. However, we do include Kiribati.

- Regarding Tanzania, Tanganyika united with Zanzibar to form the United Republic of Tanganyika and Zanzibar, then renamed Tanzania. Tanganyika and Zanzibar are currently not in the dataset, hence we do not account for this territorial change.

- Regarding Saudi Arabia, we do not track its relationship with the Saudi-Iraqi Neutral Zone, which is not included in the dataset.

A.3 GDP and population data

This section describes adaptions that we made in order to ensure that GDP and population data follows the dynamic nature of the *Gravity* dataset.

Table 3: Territorial Changes

country_id	*iso3*	*iso3num*	*country*	*first_year*	*last_year*	*countrygroup_iso3*	*countrygroup_iso3num*
ARM	ARM	51	Armenia	1991		SUN	810
ABW	ABW	533	Aruba	1986		ANT	532
AZE	AZE	31	Azerbaijan	1991		SUN	810
BGD	BGD	50	Bangladesh	1971		PAK	586
BLR	BLR	112	Belarus	1991		SUN	810
BIH	BIH	70	Bosnia and Herzegovina	1992		YUG	890
HRV	HRV	191	Croatia	1991		YUG	890
CCZE	ZE	203	Czech Republic	1993		CSK	200
CSK	CSK	200	Czechoslovakia		1993	CSK	200
DDR	DDR	278	East Germany	1949	1990	DEU	276
ERI	ERI	232	Eritrea	1993		ETH	230
EST	EST	233	Estonia	1991		SUN	810
ETH.1	ETH	230	Ethiopia + Eritrea		1992	ETH	230
ETH.2	ETH	231	Ethiopia	1993		ETH	230
GEO	268	Georgia	1991		SUN	810	
DEU.1	276	West Germany		1989	DEU	276	
DEU.2	276	Germany	1990		DEU	276	
IDN.1	IDN	360	Indonesia + Timor-Leste		2001	IDN	360
IDN.2	IDN	360	Indonesia	2002		IDN	360
KAZ	KAZ	398	Kazakhstan	1991		SUN	810
KGZ	KGZ	417	Kyrgyzstan	1991		SUN	810
LVA	LVA	428	Latvia	1991		SUN	810
LTU	LTU	440	Lithuania	1991		SUN	810
MYS.1	MYS	458	Malaysia + Singapore		1964	MYS	458
MYS.2	MYS	458	Malaysia	1965		MYS	458
MDA	MDA	498	Moldova	1991		SUN	810
MNE	MNE	499	Montenegro	2006		SCG	891
ANT.2	ANT	532	Netherlands Antilles + Aruba		1985	ANT	532
ANT.2	ANT	530	Netherlands Antilles	1986	2010	ANT	532
MKD	MKD	807	North Macedonia	1991		YUG	890
VDR	VDR		North Vietnam		1976	VNM	704
PAK.1	PAK	586	Pakistan + Bangladesh		1970	PAK	586
PAK.2	PAK	586	Pakistan	1971		PAK	586
RUS	RUS	643	Russia	1991		SUN	810
SRB	SRB	688	Serbia	2006		SCG	891
SCG	SCG	891	Serbia and Montenegro	1992	2006	YUG	890
SGP	SGP	702	Singapore	1965		MYS	458
SXM	SXM	534	Sint Marteen	2010		ANT	532
SVK	SVK	703	Slovakia	1993		CSK	200
SVN	SVN	705	Slovenia	1991		YUG	890
SSD	SSD	728	South Sudan	2011		SDN	736
YMD	YMD	720	South Yemen	1967	1990	YEM	887
SDN.1	SDN	736	Sudan + South Sudan		2010	SDN	736
SDN.2	SDN	729	Sudan	2011		SDN	736
TJK	TJK	762	Tajikistan	1991		SUN	810
TLS	TLS	626	Timor-Leste	2002		IDN	360
TKM	TKM	795	Turkmenistan	1991		SUN	810
SUN	SUN	810	USSR		1991	SUN	810
UKR	UKR	804	Ukraine	1991		SUN	810
UZB	UZB	860	Uzbekistan	1991		SUN	810
VNM.1	VNM	704	South Vietnam		1975	VNM	704
VNM.2	VNM	704	Vietnam	1976		VNM	704
YEM.1	YEM	886	North Yemen		1989	YEM	887
YEM.2	YEM	887	Yemen	1990		YEM	887
YUG	YUG	890	Yugoslavia		1992	YUG	891

WDI GDP and population data

WDI has data for some countries before their formal independence and does not account for territorial changes as is done in the *Gravity* dataset. For instance, the WDI does not have data

for Czechoslovakia, Serbia and Montenegro, and Yugoslavia, but for the underlying countries before their formal independence. We treat these cases as follows:

- WDI does not have data for Czechoslovakia, but it has Czech Republic and Slovakia pre-1993. We have summed data for Czech Republic and Slovakia to create Czechoslovakia before 1993, and replaced with missing data on Czech Republic and Slovakia before 1993.
- WDI does not have data for “Serbia and Montenegro" (1992-2006), but it has data for Montenegro and Serbia going back to 1960, rather than only from 2006. We have summed data for Montenegro and Serbia to create data for the unified country of “Serbia and Montenegro" in the period 1992-2006, and replaced with missing data on the two countries Serbia and Montenegro before 2006.
- WDI does not have data for Yugoslavia, but it has data for all its underlying countries. We have summed data for Serbia, Montenegro, Croatia, Slovenia, North Macedonia, Bosnia and Herzegovina to create Yugoslavia before 1993 (if data for each of these countries exists in every year), and replaced with missing data on these underlying countries before 1991 or 1992, depending on their date of independence.
- WDI does not have data on the Soviet Union, but it has data on countries that became independent from the Soviet Union before their formal independence (Russia, Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Tajikistan, Turkmenistan, Ukraine, Uzbekistan). In particular, it has population data going back to 1960 for many of these countries, while it has GDP data starting at different years for each of these countries. For countries that were born from the dissolution of the Soviet Union, we use as first year of independence 1991, hence we set as missing all observations pre-1991 and we sum population and GDP data for these countries until 1990 to create data for the Soviet Union (if data for each of these countries exists in every year).
- WDI has data for South Sudan since 1960, although South Sudan was established in 2011. We set data on South Sudan as missing if before 2011.
- WDI has data for Sint Marteen since 1960, although it should only be from 2010, as before then it was part of the Netherlands Antilles. Moreover, WDI has no data for the Netherlands Antilles. We set data on Sint Marteen as missing if before 2010.

Maddison population data

Maddison population data has Czechoslovakia. It also has data on the Czech Republic and Slovakia before 1993, which we set to missing. Further, it has data on the Soviet Union, also after 1990, which we set to missing. It has data on countries that became independent from the Soviet Union before their formal independence in 1991, which we set to missing before 1991. It has data on Yugoslavia also after 1993, which we set to missing, and data on some of

the countries that became independent from Yugoslavia (Croatia, Slovenia, North Macedonia, Bosnia and Herzegovina) which we replace with missing before 1991 or 1992, depending on their date of independence. However, Maddison population data does not have data on Serbia and Montenegro, hence we also cannot construct data for "Serbia and Montenegro". Further, it does not have data on South Sudan nor Sint Marteen.

Barbieri GDP data

Barbieri's historical GDP data includes data on Czechoslovakia until 1992, but it does not have data on Slovakia and Czech Republic. It also has data for 1991 on some countries that became independent from the Soviet Union (Estonia, Lithuania and Latvia). It has data that it defines as representing Russia before 1992. However, the magnitude of Russia's GDP data in Barbieri's dataset is more comparable to the Soviet Union. As a result, we replace the ISO3 code with that of the Soviet Union before 1992. Further, Barbieri's dataset has data on Yugoslavia until 1992, but it does not have data for countries that became independent from Yugoslavia in 1993 (Serbia, Montenegro, Croatia, Slovenia, North Macedonia, Bosnia and Herzegovina). In particular, since it does not have data on Serbia and Montenegro, we also cannot construct data for "Serbia and Montenegro". In addition, Barbieri does not have data on South Sudan nor Sint Marteen. However, Barbieri has GDP data for West Germany hence this is included in the dataset.

PWT GDP and population data

The PWT dataset also has data for some countries before their formal independence and does not account for territorial changes as is done in the *Gravity* dataset. For instance, it does not have data for Czechoslovakia, Serbia and Montenegro, and Yugoslavia, but for the underlying countries before their formal independence. We treat these cases as follows:

- PWT has no data on Czechoslovakia, but it has data on Slovakia and Czech Republic before their formal territorial existence in 1993 (in particular, data is available from 1990). Hence, we set data for Slovakia and Czech Republic to missing before 1993, and aggregate the countries into Czechoslovakia before 1993.

- PWT has no data on "Serbia and Montenegro", but it data on the two countries Serbia and Montenegro before 2006 (in particular, data is available from 1990). Hence we aggregate the two countries to create "Serbia and Montenegro" before 2006, and we set their respective data to missing before 2006.

- PWT has no data on Yugoslavia, but it has data on all countries that became independent from it (in particular, data is available from 1990). Hence we aggregate data for these countries before 1992 to generate data for Yugoslavia (for 2 years) and replace data on underlying countries with missing before their formal territorial independence.

- PWT has no data on the Soviet Union, but it has data for all countries that gained formal independence from the Soviet Union in 1991 (in particular, data is available from 1990).

Hence we aggregate data for these countries before 1991 to construct data for the Soviet Union (for 1 year) and set data for these countries to missing if before 1991.

General adjustments to GDP and population data

Further, for all population and GDP variables (from WDI, Barbieri and Maddison, as well as from PWT), we accounted for territorial changes of countries that suffered a split but did not cease to exist (e.g. Ethiopia who lost the part that is now Eritrea in 1993). In these cases, we have replaced the country's GDP and population data with the sum of GDP and population of the two countries before the split (e.g. Ethiopia up to 1993 is the sum of data we have for Ethiopia and Eritrea) only if data on both countries is non-missing before the split. If data on one of the two countries is missing, we have replaced the country's variable with missing. In some cases, we thus "lose" data in order to ensure the dynamic nature of the *Gravity* dataset is respected. In particular:

- Since we only have GDP data on Eritrea from 1993, we now have missing data for Ethiopia's GDP pre-1993, because the latter is set to missing when we do not have Eritrea's data.
- In the case of Sudan and South Sudan, we also lose a data because South Sudan data is only available from 1993.
- In the case of Indonesia and Timor Leste, we also lose GDP data pre-2002 on Indonesia, since GDP data for Timor Leste is only available from 2000.

We also set data to missing for all countries not mentioned above before their first year of formal territorial independence, as specified in Table 3.

A.4 RTA data from the WTO

Adjustments to listed RTAs:

The WTO dataset, as downloaded on 12/08/2020, contains information on 559 trade agreements, of which we discarded those with status "Early announcement-Under negotiation" and those that had a missing date of entry into force (34 observations in total). After doing so, 9 RTAs were left which did not have information on original signatories. We used Wikipedia to add this information for 3 of these, namely for the "Economic Community of West African States (ECOWAS)", the "European Union - Cote d'Ivoire" and the "West African Economic and Monetary Union (WAEMU)". We dropped the remaining 6 RTAs with no information on signatories.[14] The remaining dataset contains 525 RTAs.

[14] In particular, we drop the Borneo Free Trade Area signed between North Borneo and Sarawak between 1962 and 1969, because none of the member countries was independent prior to them joining Malaysia in 1963 and therefore these member countries are not included in our universe of countries. The three Conventions of Lomé Goods and the two Yaoundé RTAs are also excluded as they are categorised as Generalised System of Preference agreements (GSPs) and, apart from these 5 GSPs, the WTO dataset does not include GSPs, so that including them would create inconsistencies with the rest of the dataset.

Another important issue is that the date of entry into force sometimes differs for goods and services. Hence, in order to convert the dataset from its original from to the origin-destination-year form, we duplicates RTAs that are applied to both goods and services to obtain two different RTAs, of different coverage, with different active years.

Adjustments to listed member countries:

To construct the list of members subject to an RTA at any point in time, we started with the column denoting the original signatories (*Original signatories*). We checked cases in which the original signatories differed from the current signatories and identified 114 such cases. In many cases, the addition of new member countries is accounted for through the addition of new trade agreements that reflect accessions, such as the "Central American Common Market (CACM) - Accession of Panama" of 2013 that followed the original "Central American Common Market (CACM)" agreement of 1961. In some cases, the discrepancy between original and current signatories is due to the fact that original signatories are then described as a country group in *Current signatories*, for instance in the case of EFTA agreements with third parties. In other cases, the variable *Specific Entry Exit dates* identifies this discrepancy and contains information on the changes that took place between *Original signatories* and *Current signatories*. We thus used the information contained in *Specific Entry Exit dates* to identify cases in which countries entered later or exited earlier than other members.

When *Original signatories* and *Current signatories* coincide, there may still be cases in which signatories are country groups (for example the EU or ASEAN) and their member countries need to be adjusted depending on their date of accession to the country group. In the case in which agreements were signed between specific countries and country groups (e.g. between India and the ASEAN in the case of the "ASEAN - India RTA", trade deals involving the European Economic Area, or Agreements between the EU and third parties), we identified members of the country group in the year of entry into force of the Goods and Services sections of the RTA (as that the two may differ) and we added the corresponding members to the list of signatories. We then updated the countries included in the country group, depending on their year of accession to the country group. Further, we had to address a number of specific cases for some RTAs, which are detailed below:

- Commonwealth of Independent States (CIS): Turkmenistan is included in *Currentsi gnatories* but not in *Original signatories*. We include Turkmenistan among the original signatories because, like Ukraine, Turkmenistan ratified the CIS Creation Agreement, making it a "founding states of the CIS", but did not ratify the subsequent Charter that would make it member of the CIS. Nevertheless, both Ukraine and Turkmenistan, while not being formal members of CIS, were allowed to participate in CIS. The second discrepancy is the Kyrgyz Republic, which does not appear in *Current signatories*. However, checking the WTO it also seems that Kyrgyz Republic is still in the CIS, hence we correct for this and include Kyrgyz Republic as current member of the CIS.

- EU Overseas Countries and Territories (OCT): There is a discrepancy between *Current signatories* and *Original signatories*. For those countries present in *Current signatories*

(i.e. only those overseas territories that are marked as being current members of RTA) we add their specific entry date as corresponding to the date of accession to EU of the country they depend from. For Aruba, we add its establishment date. We also include different EU members depending on their date of accession to the EU.

- Protocol on Trade Negotiations (PTN): *Original signatories* includes Yugoslavia, but *Current signatories* does not include Yugoslavia and includes Serbia instead. To match the dynamic nature of the *Gravity* dataset, we thus add the end date of Yugoslavia, the beginning and end date of "Serbia and Montenegro" and the beginning date of Serbia to *Specific Entry Exit dates*.

We also corrected a number of small issues regarding listed member countries in the WTO dataset. Firstly, the WTO does not appear to list Czechoslovakia as a member country during years in which it existed but does include Slovakia and the Czech Republic prior to their existence. For example, the "EFTA - Czechoslovakia" agreement, which lasted from 1992-1993, does not list Czechoslovakia as a member but does list the Czech Republic, despite it not yet existing. To address this, we have classified both Slovakia and and the Czech Republic as Czechoslovakia between 1948 and 1992. Also, several agreements list member countries prior to their territorial existence. In these cases, we have not included countries as members until they gain territorial existence.