The CEPII Gravity Database

Maddalena Conte* Pierre Cotterlaz† Thierry Mayer‡

October 12, 2023

Abstract

The *Gravity* database gathers a set of variables useful to researchers or practitioners estimating gravity equations. Each observation corresponds to a combination of an exporting country, an importing country and a year (i.e. "origin-destination-year"), for which we provide trade flows, as well as geographic, cultural, trade facilitation and macroeconomic variables.

*CREST
†CEPII
‡Sciences Po, CEPII and CEPR

Contents

1 Introduction

The *Gravity* database gathers a set of variables useful to researchers or practitioners estimating gravity equations. Each observation corresponds to a combination of an exporter country, an importing country and a 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" may refer to 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 pair of countries appears every year, even if one of the countries actually does not exist. Nevertheless, when either the destination or the origin country do not exist, variables are set to missing, and dummy variables allow to easily identify these 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 alphabetic ISO3 code with a number identifying potential territorial transformations of the country. We also include the conventional ISO3 codes (numeric and alphabetic) to allow easier integration with external data sources. An additional dataset, *Countries*, associates each *country_id* with the corresponding ISO3 country codes, as well as country name, and a set of variables enabling to track territorial changes.

We created our own country identifier, *country_id*, which combines a country's alphabetic ISO3 code with a number identifying potential territorial transformations of the country. We had to do this because alphabetic and numeric ISO3 codes are not able to identify a country precisely. In particular, there are some cases in which countries experience territorial changes without getting a new, different, ISO3 alphabetic code. This happens when a country merges with another and the unified country adopts the ISO3 alphabetic code of one of the two pre-existing countries, or when 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. For instance, *country_id* becomes DEU.1 for West Germany and DEU.2 for the unified Germany, and 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.

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

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

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[2].

Table 1: List of the variables included in *Gravity*

Variable Name	Content	Level
country_id	Gravity country ID	unilateral
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
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
distw_harmonic	Population-weighted distance between most populated cities (harmonic mean)	bilateral
distw_arithmetic	Population-weighted distance between most populated cities (arithmetic mean)	bilateral
distw_harmonic_jh	Weighted distance by J. Hinz (harmonic mean)	bilateral
distw_arithmetic_jh	Weighted distance by J. Hinz (arithmetic mean)	bilateral
dist	Geodesic distance between most populated cities (km)	bilateral
main_city_source	Source of most populated city	unilateral
distcap	Geodesic distance between capital cities (km)	bilateral
diplo_disagreement	UN diplomatic disagreement score	bilateral
scaled_sci_2021	Social connectedness index in 2021	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

[2]Note that the unilateral variables do not appear twice: for simplicity, we do not repeat their definition both for the origin and the destination.

(continued)

Variable Name	Content	Level
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

(continued)

Variable Name	Content	Level
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
fta_wto_raw	1 if pair currently engaged in a regional trade agreement (source: WTO)	bilateral
fta_wto	1 if pair currently engaged in a regional trade agreement (source: WTO supplemented by Thierry Mayer)	bilateral
rta_coverage	Indicates whether the RTA covers goods only or goods and services (source: WTO)	bilateral
rta_type	Indicates the type of RTA (source: WTO)	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: from the CEPII, as well as from institutional sources such as the World Bank, the WTO and the IMF, but also from a

variety of researchers. Table 2 lists all the sources that were used and the papers that need to be cited when using these variables. Detailed explanations on how variables have been generated based on these sources are provided in the following sections.

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

Source	**Variables created based on source**
CEPII's GeoDist	Country coverage, *contig*, *dist*, *distw*, *distcap*, *distwces*, *dist_source*, *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*
timezoneDB	*gmt_offset_2020*
LaPorta et al. (1999) and LaPorta et al. (2008)	*legal_old*, *legal_new*, *comleg_pretrans*, *comleg_posttrans*, *transition_legalchange*
LaPorta et al. (1999)	*comrelig*
Head et al. (2010) 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*
Jeffrey Bergstrand's NSF-Kellogg Institute Database on Economic Integration Agreements	*rta_bergstrand*, *rta_type_bergstrand*, *gsp*
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*

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 with 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 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).

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

Countries is the dataset that includes time unvarying country-level variables. It allows a full identification of each country included in *Gravity* and, if relevant, a tracking of its territorial changes. Territorial changes refer in this context to situations where two countries combine to form one single country (East and West Germany) or where a single country splits into several independent countries (Yugoslavia). In this case, we indicate either the country's previous membership (when the country used to be part of a much larger entity) or 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.

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.

The universe of *Countries* (and of the *Gravity* dataset) is based on CEPII's GeoDist dataset (Mayer and Zignago, 2011), 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.

2.1 Variables

- ***country_id***: country identifier, *unilateral*.
- ***iso3***: ISO3 alphabetic code, *unilateral*.
- ***iso3num***: ISO3 numeric code, *unilateral*.
- ***countrygroup_iso3***: Largest entity of which the country was/is part of (ISO3 alphabetic code), *unilateral*.
- ***countrygroup_iso3num***: Largest entity of which the country was/is part of (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*.

2.2 Data Sources

- CEPII's GeoDist, World Bank's WITS and ISO website: official country ISO codes for ***iso3*** and ***iso3num***
- World Bank's WITS and Wikipedia: official country names for ***countrylong***.
- CIA World Factbook and Wikipedia: track territorial changes for ***countrygroup_iso3***, ***countrygroup_iso3num***, ***first_year*** and ***last_year***

Notes:

ISO3 codes mainly come from GeoDist. Some are updated based on WITS and on the official source for ISO country codes.

Country official names mainly 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.

2.3 Methodology and descriptive statistics

The *Countries* datasets precisely identifies each country (in its current or past territorial configuration), including some territories that are not officially independent. The identifier is *country_id*. Alphabetic and numeric ISO3 codes are also included.

A substantial territorial reconfiguration triggers a change in the numeric ISO3 code (contrary to the alphabetic ISO3 codes, that *can* remained unchanged). Thus, looking at the numeric ISO3 code can help to track territorial changes. 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. Section A.1 of the Appendix describes in detail instances where the ISO3 alphabetic or numeric code change over time.

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 5.

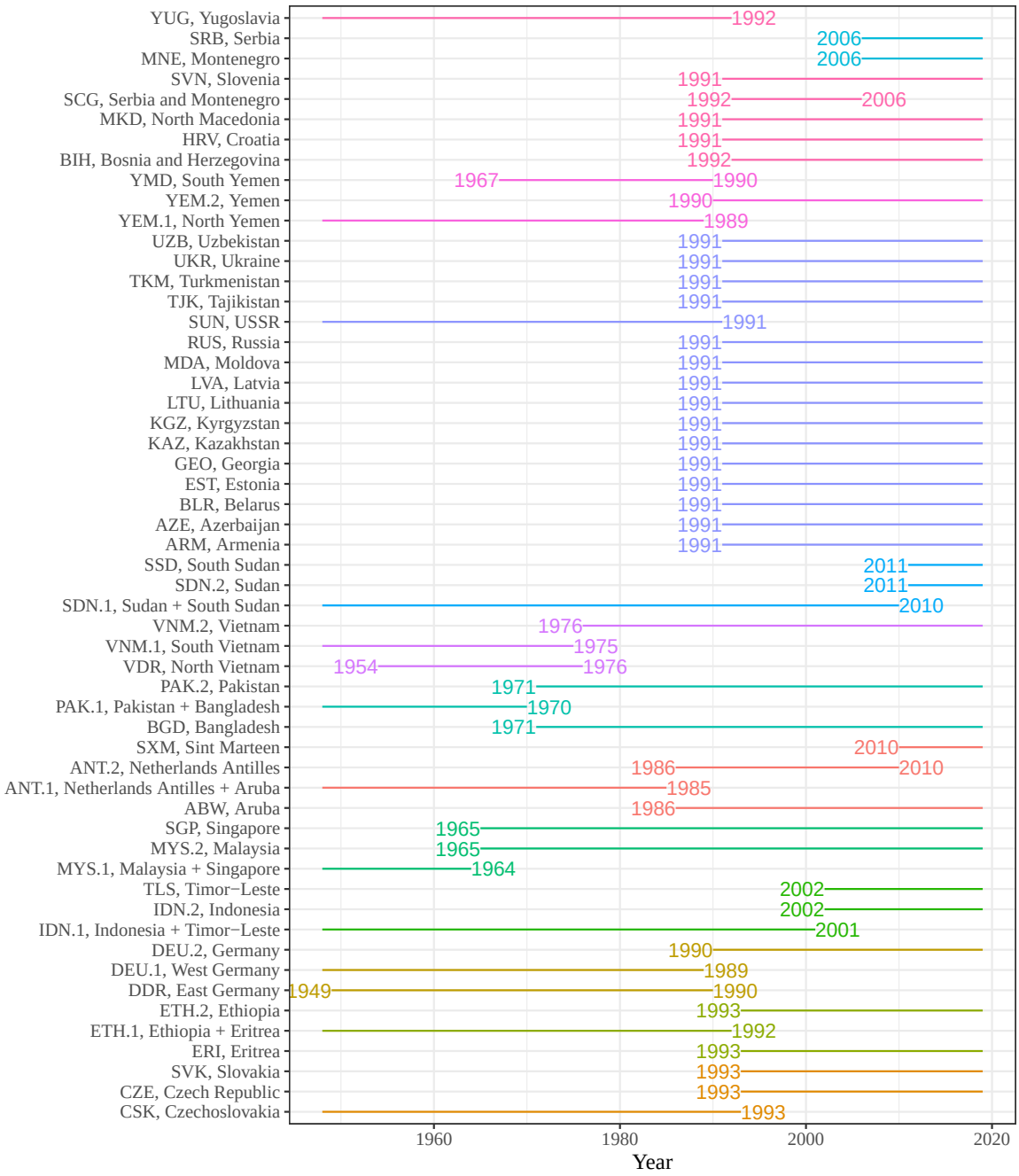
Figure 1 summarizes the territorial changes accounted for in *Gravity*. Most of the reconfigurations consist of countries splitting, with the USSR and Yugoslavia accounting for most of the new countries.

3 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 *country_id* must be noted. First, 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. Second, 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

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.

DEU.2 never existed simultaneously, we still keep these null observations to ensure that the final dataset is square.

3.1 Variables

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

3.2 Data Sources

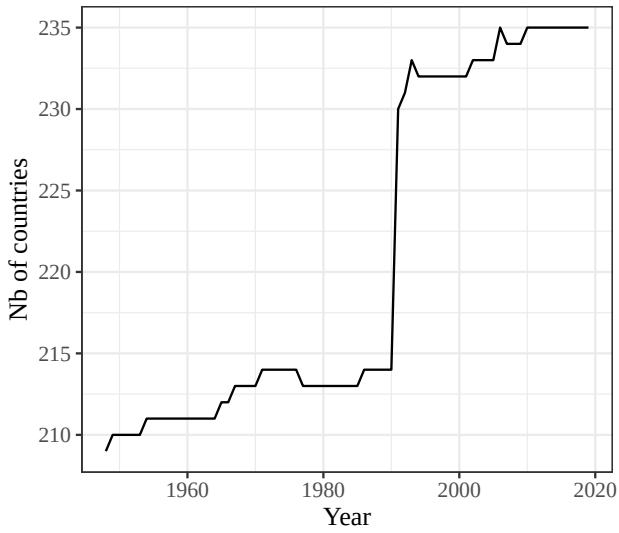
We use the same sources as in Section 2.

3.3 Methodology and descriptive statistics

country_exists dummies are constructed based on the territorial changes documented in the *Countries* dataset.

The number of existing countries experiences a sharp increase in the early 90s, corresponding to the end of socialist regimes (figure 2).

Figure 2: Number of countries in *Gravity*



Notes: Existing countries are countries for which *country_exists* = 1. We exclude cases in which origin = destination.

4 Geographic distances

We provide bilateral distances between countries. These are either simple distances, where we measure the distance between either the most populated city (*dist*) or the capital city in each country (*distcap*), or weighted distances (*distw_harmonic* and *distw_arithmetic*), where we use city-level data to account for the geographic distribution of population within each country. In addition, we include a dummy denoting whether countries are contiguous (*contig*) and two weighted distance measures computed by Julian Hinz (*distw_harmonic_jh* and *distw_arithmetic_jh*).

4.1 Theory

Head and Mayer (2010) show that the effective distance between countries i and j is given by:

$$d_{ij} = \left(\sum_{k \in i} \frac{y_k}{y_i} \sum_{l \in j} \frac{y_l}{y_j} d_{kl}^{\theta} \right)^{\frac{1}{\theta}}$$

where y denotes the economic activity, subscripts k and l denote cities, and d_{kl} denotes the geographic distance between cities k and l. θ is the distance elasticity of trade flows, which is often found to be close to -1.

One way to approach the ideal distance measure is to approximate the GDP shares with population shares, which are more easily available at a sub-national level. Weighted distances are then :

$$d_{ij} = \left(\sum_{k \in i} \frac{\text{pop}_k}{\text{pop}_i} \sum_{l \in j} \frac{\text{pop}_l}{\text{pop}_j} d_{kl}^{\theta} \right)^{\frac{1}{\theta}}$$

When the value of θ is set to 1, the distance is just an arithmetic mean of city to city distances (*distw_arithmetic*). When the value of θ is set to -1, the distance corresponds to an harmonic mean. The harmonic mean (*distw_harmonic*) is the theory consistent way to measure distance between two countries, since the empirically measured distance elasticity is close to -1 (and θ should reflect this estimated parameter).

4.2 Variables

- ***contig***: 1 if the countries are contiguous (neighbors), *bilateral*.
- ***distcap***: distance between the capital city of each country, in km, *bilateral*.
- ***dist***: distance between the most populated city of each country, in km, *bilateral*.

- ***distw_harmonic***: population-weighted average distance between the most populated cities of each country, harmonic mean, in km *bilateral*.
- ***distw_arithmetic***: population-weighted average distance between the most populated cities of each country, arithmetic mean, in km *bilateral*.
- ***distw_harmonic_jh***: weighted average distance based on satellite nightlight data, computed by Julian Hinz, harmonic mean, in km *bilateral*.
- ***distw_arithmetic_jh***: weighted average distance based on satellite nightlight data, computed by Julian Hinz, arithmetic mean, in km *bilateral*.
- ***main_city_source***: flag variable, indicates the source used to identify the most populated city in the construction of *dist*, or to obtain the country surface used in the computation of internal distances (both for *dist* and *distcap*), *unilateral*.

4.3 Data sources

Capital cities

We rely on the UN World Urbanisation Prospect 2018 to identify the capital of each country, and obtain its geographic coordinates (file WUP2018-F13-Capital_Cities.xls). This dataset will henceforth be referred to as *UN WUP Capitals*.

Population at the city level

The UN World Urbanisation Prospect 2018 is also used to obtain populations at the city level (file WUP2018-F12-Cities_Over_300K.xls). This dataset will henceforth be referred to as *UN WUP Largest*. *UN WUP Largest* contains data on 1860 cities whose population was above 300k inhabitants in 2018. It provides time variations of these populations from 1950 to 2020, based either on Census data, estimations or projections. Figures are provided with a 5 years time interval, i.e. we have population data for 1950, 1955, 1960, etc. The dataset is perfectly balanced: all cities are tracked over time from 1950 to 2020.

Coverage of the UN World Urbanisation Prospect 2018

The *UN WUP Capitals* dataset is very exhaustive. Of the 252 countries included in *Gravity*, only 22 are not included in *UN WUP Capitals*. We identify these 22 missing capitals based on external sources to ensure complete coverage. This ensures that every country included in *Gravity* has a capital and geographic coordinates for this capital. Among the 22 countries whose capital is missing from the original dataset:

- 16 correspond to former territorial configurations. For those countries, we fill in information either from the *UN WUP Capitals* dataset itself (e.g. the capital of the USSR was Moscow, which is currently the capital of Russia), or from the *UN WUP Largest* dataset (e.g. the capital of West Germany was Bonn, and we have data on this city from the *UN WUP Largest* dataset), or from Wikipedia (for the Netherlands Antilles).

- 6 correspond to small countries. We input capitals manually for them, using Wikipedia, and https://latitude.to/ for their latitude and longitude. A list of these cases is provided in the appendix, Section A.5.

Coverage of the *UN WUP Largest* is less exhaustive: there are 154 unique countries in this dataset, meaning that 98 of the 252 countries included in *Gravity* are missing. 16 of these are past territorial configurations, and 82 are small territories. For past territorial configuration, we are able to fill in missing information by assigning a past territory to each city when relevant. For example, among cities located in Germany, we can distinguish between cities that were in East Germany and cities that were in West Germany.

Geographic location of each country

This information is required to construct the contiguity dummy (*contig*). We use ARCGIS's World Countries (Generalized) dataset, which provides country boundaries in their January 2020 configuration. Again, former countries are not included in the dataset, as well as some small islands and states that are not-officially independent[3], but we manually input the contiguity information for them, based on Wikipedia data.

Surface area of each country

This information is required to compute internal distances. We use the surface area from the World Bank. Surface area is defined as a country's total area, including areas under inland bodies of water and some coastal waterways. The data corresponds to current or most up-to-date territorial configurations. Therefore, former countries are not included, but we are able to infer surfaces for them.[4] Some small island countries are also missing from the World Bank's surface data and we complement this information based on Wikipedia.

Urban surface area of each country

We use data from NASA's Urban-Rural Population and Land Area Estimates (v2) to calculate urban internal distance when constructing weighted distances. See Section 4.4.4 for more details.

[3]The missing countries are Bonaire, Sint Eustatius and Saba; Sint Marteen; Hong Kong; Macao; Taiwan and Western Sahara.

[4]For this purpose, we use two approaches:

1. For former countries that split, we sum areas from the two resulting countries, using World Bank's data. This is done for Pakistan before the independence of Bangladesh, Ethiopia before the independence of Eritrea, Malaysia the before independence of Singapore, and Indonesia before the independence of Timor-Leste.
2. For the former countries that merged, we use Wikipedia's country pages. This is done for West Germany, North Yemen, South Vietnam, Netherlands Antilles in its past territorial configurations, Sudan and South Sudan. Note that Wikipedia reports latest available surface area reached by countries, when referring to past territorial configurations. When multiple values are available for different years, we always choose the latest.

4.4 Methodology

4.4.1 General information

For both *dist* and *distcap*, we use the *geosphere* R package to calculate "city to city" distances, i.e. distances between two sets of geographic coordinates. Specifically, we use its *distGeo* function, which provides a highly accurate estimate of the shortest distance between two points on an ellipsoid.

Distance measures are rounded to the nearest km.

4.4.2 Capital to capital distance: *distcap*

General principles

The *UN WUP Capitals* dataset allows us to identify the capital of each country in 2018. We then calculate simple distances between the capitals of each country pair, in km. Changes in capitals are taken into account, so *distcap* may exhibit time variation. Since the source data only provides capitals as of 2018, we manually added information on previous capitals when relevant. A list of all the instances in which the capital of a country changes can be found in the appendix, Section A.5.

Identification of capitals

In 8 countries, the *UN WUP Capitals* dataset lists more than one capital. This happens for instance when one city hosts the government, and another one the parliament. For these 8 countries, we choose the capital on a case by case basis. We detail these choices in the appendix, Section A.5. Also, for 22 countries that were missing from the original dataset, we added the information ourselves (see Section 4.3 on data sources for more details).

Within-country distances

For internal distances, we apply the formula suggested in Head and Mayer (2010): distance $= 2/3\sqrt{\text{area}/\pi} \approx 0.38\sqrt{\text{area}}$, where area refers to the surface area of the country, taken mostly from the World Bank dataset, or from Wikipedia in case of missing information (see Section 4.3 on data sources for more details).

4.4.3 Main city to main city distance: *dist*

General principles

The *UN WUP Largest* dataset provides data on city populations every 5 years, starting from 1950. We use this data to identify the largest city for each country in each period. The distance between the largest cities of each country pair is then computed every five years from 1950 onwards. The obtained distance is rolled over until the next update year. For instance, the distance computed based on the main cities in 1950 is applied to 1950, 1951, 1952, 1953 and 1954. Then, a distance is computed based on the main cities in 1955, which is in turn applied

to the years 1955-1959. For the years 1948 and 1949, we fill this data backwards, i.e. we populate 1948 and 1949 with 1950 data. This is because the *UN WUP Largest* dataset only goes back to 1950 and not 1945. Since largest cities may vary over time, *dist* may also exhibit time variation (in cases where the main city changes in at least one of the two countries). This situation occurs very rarely: only in 19 out of 252 countries in *Gravity* (7.54%) does the main city change over the period 1948-2020. Section A.4 in the Appendix provides a list of these changes.

Identification of largest cities

In most cases, determining the most populated city in a country is straightforward, we simply retain the city that has the largest population in the *UN WUP Largest* dataset in a given period. However, in some circumstances we need to proceed differently because some existing countries are missing from the *UN WUP Largest* dataset or because the country underwent a territorial transformation, and the *UN WUP Largest* only provides data for current territorial configurations. As explained in Section 4.3, to obtain information on the largest city of former countries, we take the *UN WUP Largest* dataset and identify which cities where in which past territorial configuration. For example, we distinguish between cities that were in East Germany and cities that were in West Germany[5]. For the remaining small existing countries that are missing from *UN WUP Largest*, we use their capital, i.e. we make the implicit assumption that the capital is the most populated city of these relatively small countries.

We provide a flag variable, *main_city_source*, that keeps trace of this treatment, for both origin and destination. It takes value:

- 1 (UN_WUP_Largest) if the largest city is determined based on the *UN_WUP_Largest* dataset. This also includes past territorial configurations, whose largest city we have identified as detailed above.
- 2 (UN_WUP_Capitals) if it is taken from the *UN_WUP_Capitals dataset*, with modifications as described in Section 4.4.2.
- 3 (WB_Area) if data for internal distances is taken from the World Bank surface area dataset.
- 4 (Wikipedia_Area) if data for internal distances is taken from Wikipedia.

Note that sources 3-4 refer to source data for internal distances, i.e. when origin and destination are the same. Sources 3-4 also apply to the construction of the variable *distcap*.

Within-country distances

Same as for *distcap*.

[5]However, in the case of the Netherlands Antilles (before and after the independence of Aruba), the *UN WUP Largest* does not contain any data on largest cities in this territory. Thus, for the Netherlands Antilles, we use data on its capital city in order to construct *dist*

4.4.4 Population weighted distances: *distw_harmonic* and *distw_arithmetic*

General principles

Our purpose is to construct country-to-country distances that take into account the spatial distribution of economic activity within each country. We proxy economic activity by population, and compute a weighted average of distances between each existing pair of cities (i.e. each combination of cities in the origin and destination countries). The weight reflects the population share of the pair. More precisely, Head and Mayer (2010) derived a theory consistent aggregation formula to compute the distance from country i to country j based on distances between cities (d_{kl}):

$$d_{ij} = \left(\sum_{k \in i} \frac{\text{pop}_k}{\text{pop}_i} \sum_{l \in j} \frac{\text{pop}_l}{\text{pop}_j} d_{kl}^{\theta} \right)^{\frac{1}{\theta}}$$

where pop denotes the population, d_{kl} the distance from city k to city l, and θ is the distance elasticity of trade flows, i.e. the distance coefficient estimated in standard cross-sectional gravity equations, whose value is often found to be close to -1.

We consider two values for the θ parameter. When setting $\theta = 1$, the computed distance is an arithmetic mean of the distances between all possible pairs of cities. We name this geographic distance *distw_arithmetic*.

$$distw\_arithmetic_{ij} = \sum_{k \in i} \frac{\text{pop}_k}{\text{pop}_i} \sum_{l \in j} \frac{\text{pop}_l}{\text{pop}_j} d_{kl}$$

When setting $\theta = -1$, the computed distance is an harmonic mean of the distances between all possible pairs of cities. We name this geographic distance *distw_harmonic*.

$$distw\_harmonic_{ij} = \left(\sum_{k \in i} \frac{\text{pop}_k}{\text{pop}_i} \sum_{l \in j} \frac{\text{pop}_l}{\text{pop}_j} d_{kl}^{-1} \right)^{-1}$$

Even though this second formula is less intuitive, *distw_harmonic* is a more appropriate measure than *distw_arithmetic*, since the empirical estimates of the distance elasticity of trade flows are unambiguously closer to -1 than to 1.

The *UN WUP Largest* dataset provides data on city populations every 5 years, starting in 1950, so the weighted distances are in fact computed every five years and rolled over until the next update year[6]. Since the spatial distribution of population varies over time, population weighted distances also exhibit time variation.

All final distance measures are rounded to the nearest km.

Special cases

Distances for past territorial configurations can be computed since we assigned cities to

[6]Only for the years 1948 and 1949 do we fill the data backward, i.e. we populate those years with 1950 figures, given that the *UN WUP Largest* dataset only goes back to 1950 and not 1945.

former countries when relevant (see Section 4.3). If a country does not have any city in the *UN WUP Largest* dataset, we use its capital as single city.

Within-country distances are computed using the same formula, i.e. we form all possible combinations of cities within the country, and take a weighted average of these combinations. Some of the combinations correspond to distance from a city to itself (within-city distances). We infer city surfaces using an additional data source.

4.4.5 The contiguity dummy: *contig*

To construct the variable *contig*, we use ARCGIS's World Countries (Generalized) dataset, which provides optimized country boundaries as of January 2020. In particular, the dataset contains .dbf and .shp files, from which we can extract contiguity information using Stata's $spshape2dta$ and $spmatrix$ commands, with the option $contiguity$.

In 22 cases, we have no contiguity data as the source file does not cover all countries. 4 of these have no contiguous country, while for the remaining missing countries and for past territorial configurations, we augment the dataset manually using Wikipedia as source, as explained in Section 4.3. When origin and destination coincide, we set $contig = 0$.

4.4.6 Julian Hinz's weighted distances: *distw_arithmetic_jh* and *distw_harmonic_jh*

We include Julian Hinz's weighted distances, which are computed using nighttime lights data as weights (Hinz, 2021). Night luminosity is a good way to proxy the spatial distribution of economic activity within a country, but the data covers a more limited time span than the other distance measures, from 1992 to 2012.

5 Cultural variables

5.1 Variables

- ***gmt_offset_2020***: GMT offset in 2020 of the country measured in hours, *unilateral*.
- ***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***: Origin of the country's legal system. Possible values are british, french, german, scandinavian, or socialist. *unilateral*

- ***legal_new***: Origin of the country's legal system, after the fall of the USSR, *unilateral*. Possible values are british, french, german, scandinavian, or socialist. *unilateral*
- ***comleg_pretrans***: 1 if countries share common legal origins before transition, *bilateral*.
- ***comleg_posttrans***: 1 if countries share common legal origins after transition, *bilateral*.
- ***transition_legalchange***: 1 if common legal origin has changed since the fall of the USSR, *bilateral*.
- ***comrelig***: Religious proximity index (Disdier and Mayer, 2007): obtained by summing the products of the shares of Catholics, Protestants and Muslims in the origin and destination countries. Varies between 0 and 1, increases when the country pair shares a common religion practised by a large share of the population.
- ***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 country pair was ever in colonial relationship. Takes into account colonial relationships before 1948, *bilateral*.
- ***col_dep***: 1 if country pair currently in colonial or dependency relationship, *bilateral*.
- ***col_dep_end_year***: Independence date from hegemon, if the pair was ever in a colonial or dependency relationship (if *col_dep_ever* is equal to 1). Missing if the pair never was in a colonial or dependency relationship (*col_dep_ever* $= 0$). Takes into account colonial relationships before 1948, *bilateral*.
- ***col_dep_end_conflict***: 1 if independence involved conflict. Missing if the pair never was in a colonial or dependency relationship (*col_dep_ever* $= 0$). Takes into account colonial relationships before 1948, *bilateral*.
- ***sibling_ever***: 1 if pair ever in sibling relationship, i.e. if they ever had the same hegemon. Takes into account colonial relationships before 1948, *bilateral*.
- ***sibling***: 1 if pair currently in sibling relationship, i.e. if they have the same hegemon, *bilateral*.
- ***sever_year***: Severance year for sibling pairs. Corresponds to the year in which the first sibling in the pair became independent. Takes into account colonial relationships before 1948, *bilateral*.
- ***empire***: Name of the hegemon if the pair is currently sibling (i.e. *year* is smaller than *sever_year*), *bilateral*.

- ***sib_conflict***: 1 if pair ever in sibling relationship (*sibling_ever* = 1) and if their independence from the hegemon involved a conflict with the hegemon. Takes into account colonial relationships before 1948, *bilateral*.
- ***scaled_sci***: Social Connectedness Index, *bilateral*.
- ***diplo_disagreement***: Diplomatic disagreement, measured through UN votes, *bilateral*.

5.2 Data Sources

- TimeZoneDB: time zones → *gmt_offset_2020*
- CEPII's GeoDist: common languages → *comlang_off*, *comlang_ethno*. Colonial ties → *comcol*, *col45*
- LaPorta et al. (1999) and LaPorta et al. (2008): historical origins of legal systems → *legal_old*, *legal_new*, *comleg_pretrans*, *comleg_posttrans*, *transition_legalchange*
- LaPorta et al. (1999): religion shares → *comrelig*
- Head et al. (2010), CIA World Factbook, Correlates of War Project (COW): colonial ties → *heg*, *col_dep_ever*, *col_dep*, *col_dep_end_year*, *col_dep_end_conflict*
- Bailey et al. (2017) (permanent link): UN General Assembly Voting → *diplo_disagreement*
- Bailey et al. (2018) (download page): Social Connectedness Index (SCI) → *scaled_sci_2021*
- Authors' computations based on colonial ties variables → *empire*, *sibling_ever*, *sibling*, *sever_year*, *sib_conflict*

5.3 Methodology and descriptive statistics

Time zones:

No modification is made to the data provided by timezoneDB. The information corresponds to time zones as they were in July 2020. Time zones have often changed in the past but the time zone reported in *Gravity* for a given country in a given year does not correspond to the time zone in that year. We use standard time zones, i.e. not daylight saving time zones. For countries that no longer exist, we use the time zone of the former capital city. For countries with more than one time zone, we choose the timezone of their capital city.

Colonisation and shared language data from CEPII:

Data from the CEPII's GeoDist dataset is directly added to *Gravity*. Note that since this dataset will no longer be updated in the future, these variables will also no longer be updated. 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.[7]

Historical origin of a country's legal system:

Data on countries' legal origin is available in two different versions. LaPorta et al. (1999) is the original version of the dataset, and 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 switched to different legal structures after gaining independence from the USSR. To distinguish pre- and post- transition values, 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.

There are 29 countries absent from the original datasets. However, for countries affected by territorial changes, we fill in missing data with data available for their respective country group.[8] Furthermore, we make some small corrections on the original dataset.[9]

We construct some additional variables based on the legal origins information: *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; *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; and *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.

Religion similarity

The religious similarity index uses religion shares provided in LaPorta et al. (1999). 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.[10]

[7] 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. Also, the variable *col45* is set to missing for Taiwan and Hong Kong with respect to China, and for Palestine with respect to Israel.

[8] 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.

[9] 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

[10] 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.

Colonial ties:

heg, *col_dep_ever*, *col_dep* and *col_dep_end_year* are constructed using data from Head et al. (2010) as the main source. *col_dep_end_conflict* is based exclusively on COW data. Data from Head et al. (2010) identifies hegemon-colony pairs and the independence date of colonies or dependencies. Since the coverage is less complete than in *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 (*empire*, *sibling_ever*, *sibling*, *sever_year*, *sib_conflict*).

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

col_dep_end_year indicates independence from the current coloniser (or hegemon in case of dependencies), not from the latest coloniser. 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.[11]

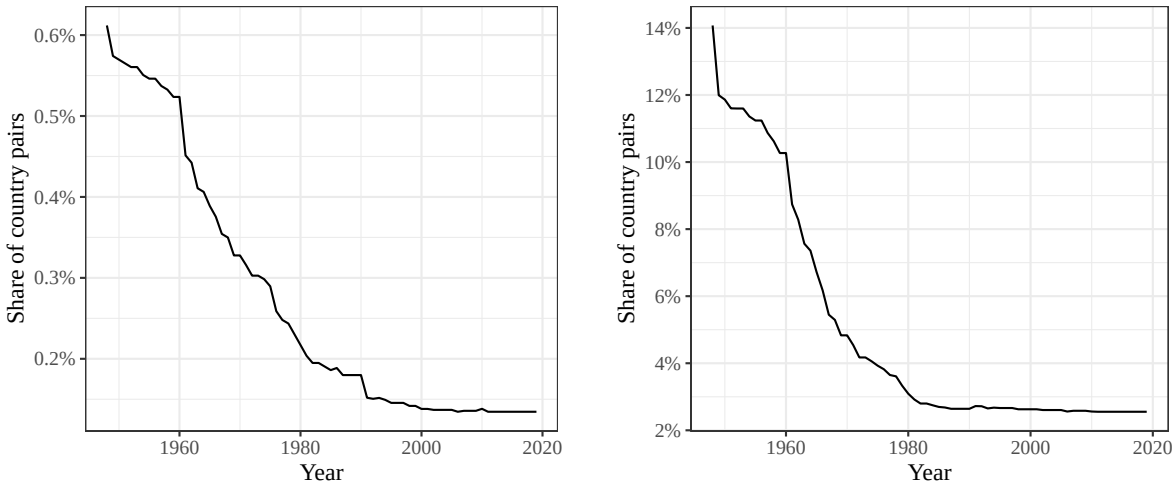
For Taiwan and Hong Kong, all variables on colonial or dependency ties with respect to China are set to missing. Dependency ties between Palestine and Israel are also set to missing. These observations are the only ones for which both countries exist but colonial ties variables are nevertheless missing.

Political distance: UN voting disagreement score:

We use the dyadic dataset constructed by Bailey et al. (2017). It provides ideal point esti-

[11] 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 origin = destination.

mates for each country and each UN session (there may be more than one session per year[12]). The ideal point captures the position of states vis-a-vis the US-led order. It is normalized to have a mean equal to zero and a standard deviation equal to 1. To construct *diplo_disagreement*, we first average ideal point estimates for each country and year. We then compute the absolute distance between average ideal point estimates in a given year of the two countries in a country pair.

Countries in the original data are identified with Correlates Of War country codes and not with ISO3 codes. Therefore we have to convert the former to ISO3 country country codes and then account for territorial variations for some countries.[13]

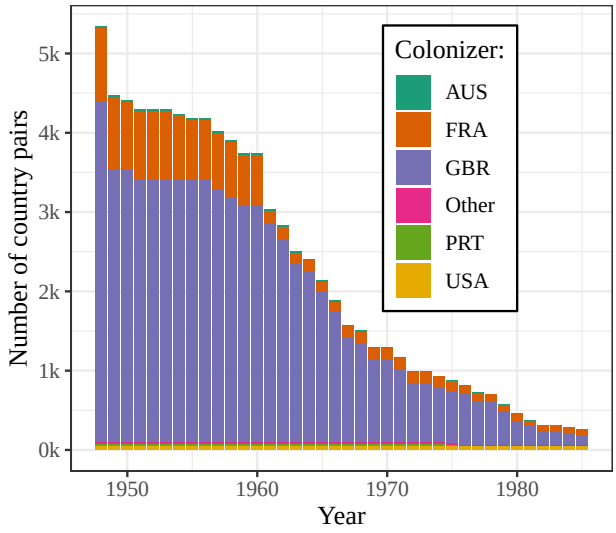
Social distance: Social Connectedness Index:

We use the country_country.tsv dataset, updated on 4 January 2021. This means that the data describes the situation in 2021. The SCI dataset is based on an anonymized snapshot of

[12]Furthermore, the year refers to the year of the session and not necessarily the year of voting, as voting may take place in a delayed fashion compared to the session, sometimes running into January or even spring of the following year. However, most votes take place in the autumn of the session year.

[13]In most cases, conversion from the Correlates Of War country code to *country_id* is straightforward. However, the dataset seems to refer to Serbia, Serbia and Montenegro and Yugoslavia with the same Correlates Of War country code of 345. In this case, we assume that the data refers to Serbia starting from 2006, to Serbia and Montenegro between 1992 and 2006, and to Yugoslavia before 1992.

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



Notes: Siblings are country pairs that share a common colonizer. We exclude cases in which origin = destination.

all active Facebook users and their friendship networks. It captures the intensity of the social connectedness between different countries. More precisely, it measures the relative probability that two individuals across two locations are friends with each other on Facebook. The measure is rescaled to have a maximum value of 10^9 and a minimum value of 1. The data excludes certain countries, e.g., countries where Facebook is banned or countries with only few active users. Overall, there are 185 unique countries in the dataset.

6 Macroeconomic Indicators

6.1 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*.
- ***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*.

6.2 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: the geographic entities correspond to the 2021 situation. Further, it does not cover years prior to 1960. We rely on two alternative sources to fill in 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), 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[14]).

As an alternative source of GDP and population data, we use the Penn World Tables (PWT) version 9.1 (Feenstra et al., 2015).[15] 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.

6.3 Methodology

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,

[14]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).

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

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.[16]

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.

7 Trade facilitation variables

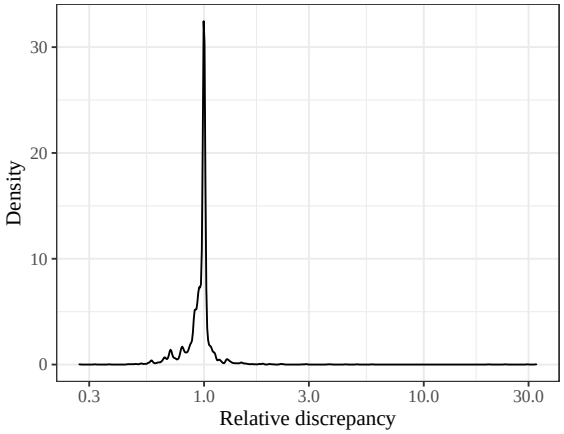
7.1 Variables

- ***gatt***: 1 if the country is a GATT member, *unilateral*.
- ***wto***: 1 if the country is a WTO member, *unilateral*.
- ***eu***: 1 if country is a EU member, *unilateral*.

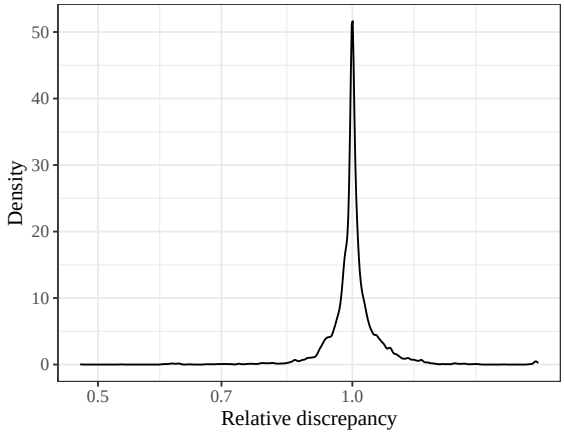
[16] In particular, we are not able to include GDP and population data for DEU.1 in 1990, YEM.1 in 1990 and VNM.1 in 1976.

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).

- ***fta_wto***: 1 if the country pair is engaged in a regional trade agreement, source WTO supplemented by Thierry Mayer, *bilateral*.
- ***fta_wto_raw***: 1 if the country pair is engaged in a regional trade agreement, source WTO, *bilateral*.
- ***rta_coverage***: Coverage of the trade agreement. 0 = "no trade agreement". 1 = "goods only", 2 = "services only", 3 = "goods and services", *bilateral*.
- ***rta_type***: Type of trade agreement. PSA = "Partial Scope Agreement", FTA = "Free Trade Agreement", CU = "Customs Union", EIA = "Economic Integration Agreement". See below for more detailed explanations , *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*.

7.2 Data Sources

- List of GATT members on WTO website → *gatt*
- List of WTO members on WTO website → *wto*

- WTO's Regional Trade Agreements database → *fta_wto*, *fta_wto_raw*, *rta_type*, *rta_coverage*
- World Bank's Development Indicators (WDI) → *entry_cost*, *entry_proc*, *entry_time*, *entry_tp*

7.3 Methodology and descriptive statistics

GATT and WTO membership:

Data on GATT and WTO membership is taken directly from the WTO, 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[17]. 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. 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:

For each RTA, the WTO dataset lists the RTA name, the coverage (i.e. whether it covers goods, services or both), the type of RTA (i.e a measure of the depth of the agreement), the date of entry into force, the original signatories, and specific entry or exit dates for additional signatories. The original dataset is structured with one observation per trade agreement, and has to be converted into the origin-destination-year structure of the *Gravity* dataset.

The conventions of Lomé and Yaoundé are 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.

We use the following time convention: a country pair is considered as being in a RTA in a given year as soon as the RTA was in force before July, 1st. Also, we do not consider that countries have RTAs with themselves, i.e. RTA variables are set to zero when origin = destination.

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

[17] These cases are Lebanon and Syria for countries that currently still exist.

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").

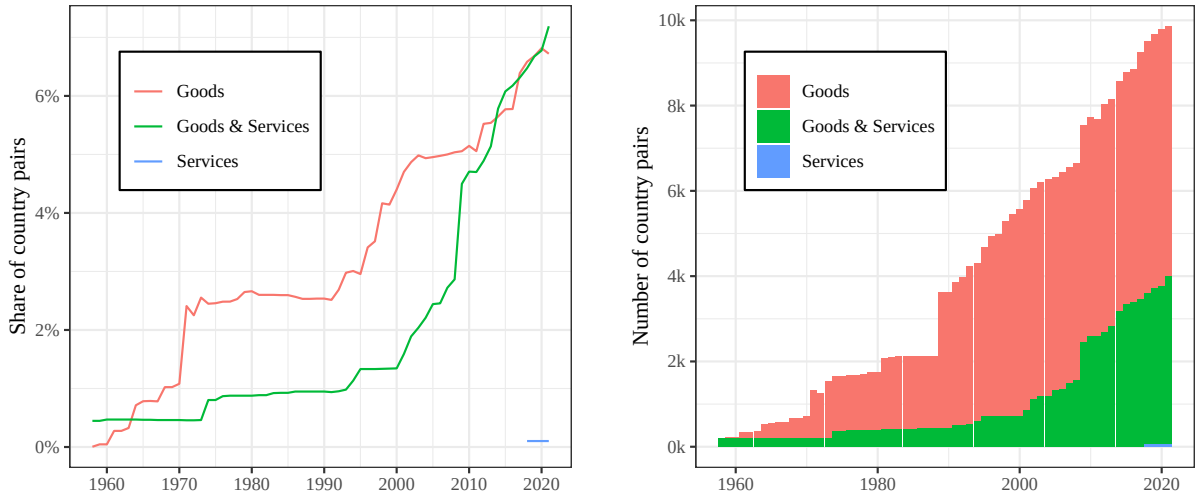
Also, the WTO data allows to create a variable distinguishing between RTAs that cover only goods, only services, or both goods and services (*rta_coverage*)

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.

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.

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



(a) Share (b) Number

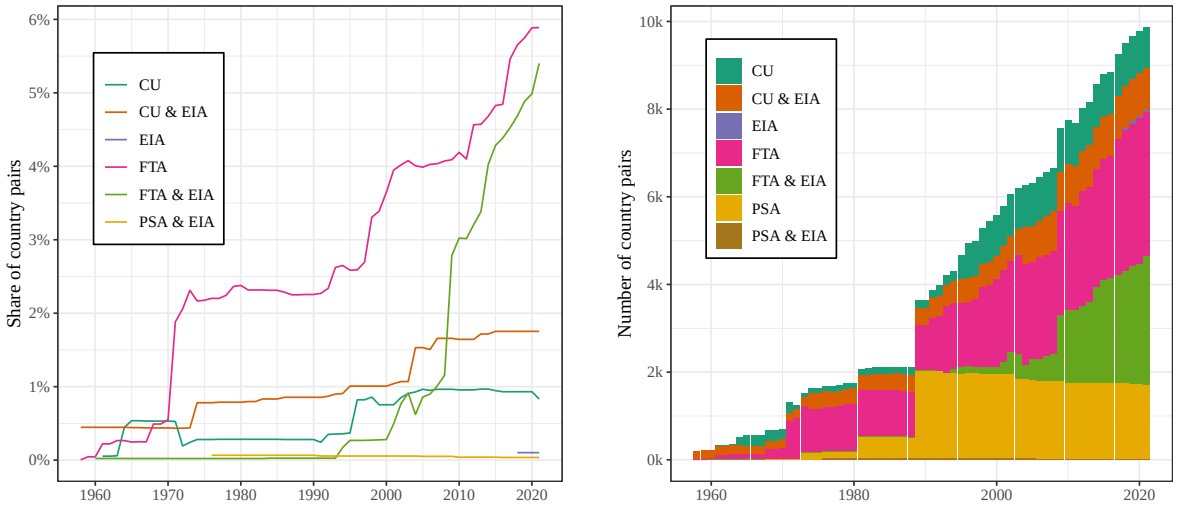
Notes: The share of country pairs is the ratio "number of existing country pairs in a given RTA coverage in a given year"/"number of country pairs existing in a given year". Before 1958 there are virtually no country pairs involved in RTAs. We exclude cases in which origin = destination.

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.[18] Note that the earliest year available from the WDI dataset on entry costs is 2003. No data is available for Taiwan. Based on the data taken

[18]We use 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, 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 origin = destination.

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

8 Trade flow variables

8.1 Variables

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

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

BACI trade flows (*tradeflow_baci*, *manuf_tradeflow_baci*) are taken from the CEPII's BACI dataset. BACI provides a single harmonized trade flow for each exporter-importer-year based on trade flows recorded in Comtrade. However, its time coverage is more limited, since it is only available for trade flows after 1996.

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. The IMF DOTS data does not include former countries but only refers to current territorial configurations.

8.3 Methodology

UNSD's Comtrade:

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:

BACI already has an "origin-destination" structure, so we do not need to reshape the data. BACI has detailed data at the product level. The product nomenclature is the Harmonized System (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 classi-

fication (revision 3.1), nor to the Primary Goods categories of the BEC classification (revision 4).

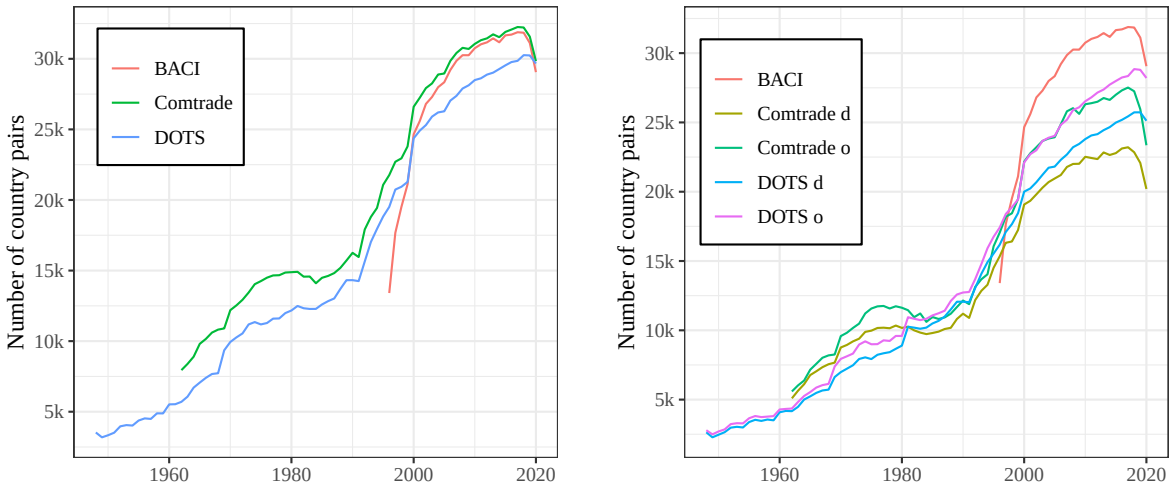
IMF DOTS:

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 8 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 8: Number of country pairs with trade flow data, by source.

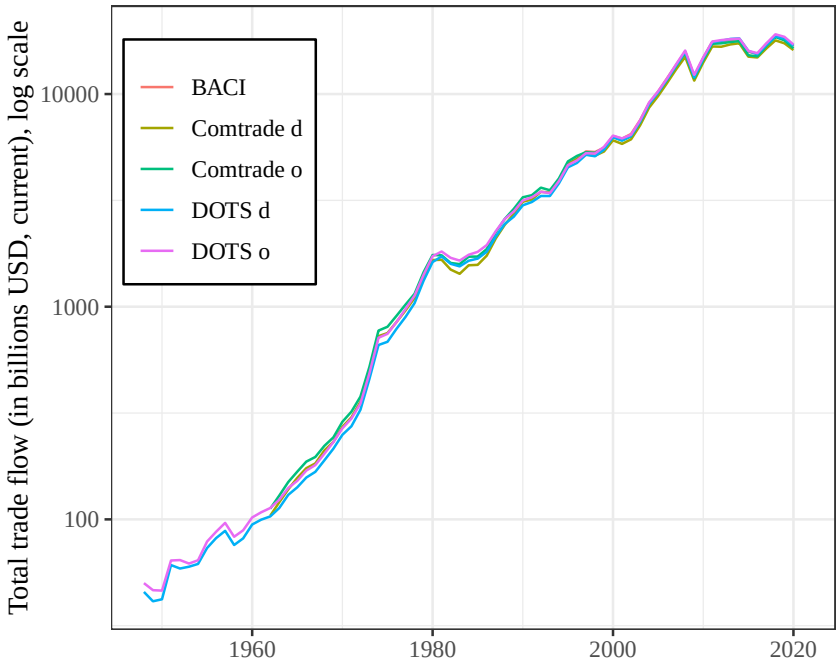


(a) (b)

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 origin = destination. (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 origin = destination.

Figure 9 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 9: 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.

References

Bailey, Michael, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong (2018). "Social Connectedness: Measurement, Determinants, and Effects". *Journal of Economic Perspectives* 3.32, pp. 259–80.

Bailey, Michael, Anton Strezhnev, and Erik Voeten (2017). "Estimating dynamic state preferences from united nations voting data". *Journal of Conflict Resolution* 2.61, pp. 430–56.

Barbieri, Katherine (2005). "The Liberal Illusion: Does Trade Promote Peace?" Ed. by Ann Arbor: University of Michigan Press.

Disdier, Anne-Célia and Thierry Mayer (2007). "Je t'aime, moi non plus: Bilateral opinions and international trade". *European Journal of Political Economy* 23.4, pp. 1140–1159.

Feenstra, R C, R Inklaar, and M P Timmer (2015). "The Next Generation of the Penn World Table". *American Economic Review* 105.10, pp. 3150–3182.

Head, K, T Mayer, and J Ries (2010). "The erosion of colonial trade linkages after independence." *Journal of International Economics* 81.1, pp. 1–14.

Head, Keith and Thierry Mayer (2010). "Illusory Border Effects : Distance Mismeasurement Inflates Estimates of Home Bias in Trade". *The Gravity Model in International Trade*.

Hinz, Julian (2021). *The view from space: Theory-based time-varying distances in the gravity model*. Tech. rep.

LaPorta, R, F Lopez-de-Silanes, and A Shleifer (2008). "The Economic Consequences of Legal Origins." *Journal of Economic Literature* 46.2, pp. 285–332.

LaPorta, R, F Lopez-de-Silanes, A Shleifer, and R Vishny (1999). "The Quality of Government." *Journal of Law, Economics and Organization* 15.1, pp. 222–279.

Mayer, T and S Zignago (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

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

The most populated city of a country may vary over time. The table below gathers information on all such changes.

Country ID	Country	City	First year	Last year
BEN	Benin	Abomey-Calavi	2015	2019
BEN	Benin	Cotonou	1948	2014
BRA	Brazil	Rio de Janeiro	1948	1969
BRA	Brazil	São Paulo	1970	2019
BFA	Burkina Faso	Bobo-Dioulasso	1948	1954
BFA	Burkina Faso	Ouagadougou	1955	2019
CMR	Cameroon	Douala	1948	2014
CMR	Cameroon	Yaoundé	2015	2019

CAN	Canada	Montréal	1948	1979
CAN	Canada	Toronto	1980	2019
GHA	Ghana	Accra	1948	2014
GHA	Ghana	Kumasi	2015	2019
IND	India	Delhi	2005	2019
IND	India	Kolkata (Calcutta)	1948	1979
IND	India	Mumbai (Bombay)	1980	2004
MWI	Malawi	Blantyre-Limbe	1948	2009
MWI	Malawi	Lilongwe	2010	2019
MOZ	Mozambique	Maputo	1948	2014
MOZ	Mozambique	Matola	2015	2019
NLD	Netherlands	Amsterdam	1948	2019
NLD	Netherlands	Rotterdam	1960	1994
NGA	Nigeria	Ibadan	1948	1959
NGA	Nigeria	Lagos	1960	2019
YEM.1	North Yemen	Al-Hudaydah	1948	1964
YEM.1	North Yemen	Sana'a'	1965	2019
QAT	Qatar	Ad-Dawhah (Doha)	1948	2014
QAT	Qatar	Ar-Rayyan	2015	2019
SAU	Saudi Arabia	Ar-Riyadh (Riyadh)	1965	2019
SAU	Saudi Arabia	Makkah (Mecca)	1948	1964
ZAF	South Africa	Cape Town	1980	1999
ZAF	South Africa	Johannesburg	1948	2019
ESP	Spain	Barcelona	1948	1969
ESP	Spain	Madrid	1970	2019
SYR	Syria	Dimashq (Damascus)	1948	2019
SYR	Syria	Halab (Aleppo)	1995	2014
TWN	Taiwan	Taibei	1948	1989
TWN	Taiwan	Xinbei	1990	2019
ZMB	Zambia	Kitwe	1948	1964
ZMB	Zambia	Lusaka	1965	2019

A.5 Capital to capital distances

The *UN WUP Capitals* dataset only provides capitals as of 2018. There are a few instances in which capitals have changed over the years and we update the dataset accordingly using Wikipedia. In particular:

- Brazil: the current capital is Brasília, but Rio de Janeiro was the capital until 1960.
- Côte d'Ivoire: Abidjan was the former capital, from 1934 to 1983. In 1983 Yamoussoukro became the new capital.

- Nigeria: Lagos was the capital until 1991, when the capital was moved to Abuja.
- Tanzania: Zanzibar City was the capital of Zanzibar until 1964, while Dar es Salaam was the capital of Tanganyika. The capital of Tanzania (formed from union of Zanzibar and Tanganyika) from 1964 has been Dodoma. In this case we do not worry as Zanzibar and Tanganyika are not included in the Gravity dataset.
- Burundi: the capital was Bujumbura from 1962 to 2018 and has been moved to Gitega in 2018. While the UN Urban Population dataset does not include Gitega either, we add it manually to our Capitals dataset.
- Malawi: Zomba was the capital until 1975. Since then, the capital has moved to Lilongwe. The UN Urban Population dataset does not include data for this city, hence we add it manually.
- Kazakhstan: the capital was Almaty from 1993 to 1998, when it was moved to Astana. In 2019 the capital Astana was renamed Nur Sultan.
- Oman: Salalah was the capital until 1970, when the capital was moved to Musqat.
- Myanmar: Rangoon (Yangon) was the capital until 2005, when the capital was moved to Naypyidaw.
- Philippines: Baguio was the summer capital from 1946 to 1976, while Quezon City was the "normal" capital. The capital was moved to Manila in 1976, but Quezon City, with Manila, became parts of Metro Manila. Hence, in this case we do not change the capital city.
- China: Nanking was the capital of the Republic of China from 1945 to 1949 (de facto). It was then moved to Taipei after the loss of the mainland which caused the nationalist government to flee to Taiwan and made Taipei the temporary capital of the Republic of China (now capital of Taiwan). Beijing was proclaimed the capital of the People's Republic of China. Since the Gravity dataset begins in 1949, we don't have to worry about this.
- Palau: Koror was the capital from 1994 to 2006, when the capital was moved to Ngerulmud. While the UN Urban Population dataset does not include Ngerulmud either, we add it manually to our Capitals dataset.

When capitals change, we consider the new capital as the capital for the year in which the change occurred, rather than the old capital.

In some cases, the *UN WUP Capitals* dataset lists more than one capital per country. We present below the list of the concerned countries, and the decision we made for each of these countries.

- Benin: the dataset lists Cotonou as the economic capital and Porto-Novo as the constitutional capital, which we choose as capital.

- Bolivia: the dataset lists La Paz as the seat of government and Sucre as the constitutional capital, which we choose as capital.
- Taiwan: Taipei is denoted as "Others" rather than as capital, and we choose this as capital.
- Netherlands has Amsterdam as capital and the Hague as seat of government. We choose the first as capital.
- South Africa has 3 capitals listed: Pretoria (executive), Cape Town (legislative), Bloemfontein (judicial). We choose Pretoria as capital.
- Sri Lanka has Colombo as capital and Sri Jayewardenepura Kotte as legislative capital. We choose Colombo as capital.
- Tokelau has its capital (Tokelau) listed as "Others", and we choose it as capital.
- Falkland Islands (Malvinas) has its capital (Stanley) listed as "Others", and we choose it as capital.

For past territorial configurations, we manually assign the following capitals:

- Bonn was the capital of West Germany from 1949 until 1990, while east Berlin was the capital of East Germany.
- Yugoslavia and Serbia and Montenegro: Belgrade was the capital of both states.
- Czechoslovakia: Prague.
- USSR: Moscow.
- South Vietnam: Saigon (Ho Chi Minh City) was the capital before unification in 1975 when Hanoi became the capital of Vietnam. Hanoi was also the capital of North Vietnam until reunification.
- South Yemen: Aden.
- North Yemen: Sana'a (1918–1948, 1962–1990); Ta'izz (1948–1962).

Lastly, for the Netherlands Antilles and a few other small island territories (British Indian Ocean Territory, Christmas Island, Cocos (Keeling) Islands, Norfolk Island, Pitcairn Islands), we use data from Wikipedia to identify their capital and data from https://latitude.to/ to identify their latitude and longitude.