'Tis but thy name that is my enemy: On the construction of macro panel datasets in conflict and peace economics

Vanessa A. Boese and Katrin Kamin

Vanessa A. Boese is at the School of Business and Economics, Humboldt University, Berlin, Germany. She may be reached at boesevan@hu-berlin.de. Katrin Kamin is at the Department of Economics, Christian Albrechts University, Kiel, Germany. She may be reached at k.kamin@economics.uni-kiel.de.

Abstract

The empirical analysis of datasets covering a large number of countries and time periods has become an integral part of conflict and peace economics. As such, numerous studies examine relationships between and among macroeconomic, political, and conflict variables and this often involves the merging of disparate datasets to combine relevant variables for which the country unit of analysis, however, is not necessarily the same. This article highlights difficulties in the data merging process and, by way of example, presents detailed country coding unit comparison for two economic (UN Comtrade and World Development Indicators), two democracy (Polity IV and V-Dem), and two conflict datasets (UCDP/PRIO Armed Conflict Dataset and COW Militarized Interstate Disputes Dataset). We find that merging datasets can result in the elimination of very large numbers of observations due to unmergeable records and that dropped observations often include the very countries or territorial entities most of interest in conflict and peace economics.

In conflict and peace economics, the construction of large panel datasets nowadays forms the basis for the majority of empirical cross-country studies. Originating from different sources, such panel datasets contain measures on variables such as international trade, economic growth, GDP, armed conflict, democratization, and government effectiveness.[1] But bringing these variables together, that is merging them into a single dataset, hinges on the exact identification of the country unit under study. To permit reasonable statistical inference, the country unit for which, for example, the trade value is calculated, should respond to the same entity for which all other variables in the dataset are coded. Unfortunately, the names, and even the physical borders, with which countries are coded vary considerably across different data sources.[2]

At the core of the coding differences lies the question "What's in a (country) name"? We argue that there are two complementary parts to the answer. The first regards the entity under observation, the unit of analysis: What is a country? The answer depends on the research framework. For example, the purpose of the Russett, Singer, and Small (1968) state list as well as of the original Gleditsch and Ward (1999) state list was to capture recognized states in the international system. This particular definition of a country is of utmost relevance in analyses of authority structures. Nevertheless, one cannot blindly assume that the unit of analysis, that is, the country, is defined along the same criteria in economic or political datasets. Unfortunately, the burden of comparing the unit of analysis underlying different macro panel datasets lies with the scholar(s) attempting to merge them. As a consequence, we emphasize the importance of discussing the merging process in empirical studies in conflict and peace economics.

The second part to the "What's in a (country) name?" question concerns the entity's label: Numerous scholars have presented ways to adjust for differences in country labels. For example, Paul Hensel (2016) provides a thorough list of alternative historical state names and Heather Ba has created Stata files allowing for the mapping of country names, Correlates of War (COW) codes, and World Bank codes.[3]

That inconsistent country names across different data sources pose a problem is widely known among scholars working with macro panel datasets. Major attempts to standardize worldwide country coding already were undertaken half a century ago by Russett, Singer and Small (1968) and almost twenty years ago by Gleditsch and Ward (1999). Nevertheless, several problems remain unresolved and, unfortunately—with the emergence of readily available software packages and codes—a discussion of "what is the (country) unit of analysis" has become almost unfashionable. In spite of its tediousness and complexity, the country merging process is generally not discussed in academic papers (or in their supplementary materials).

The contribution of this article is hence twofold: First and foremost, it shows that in spite of all country coding scheme standardization efforts and relevant software packages or

The Economics of Peace and Security Journal — ISSN 1749-852X — https://www.EPSJournal.org.uk
© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

codes, the problem of inconsistent country coding in macro panel datasets persist. We therefore want to re-raise awareness of this problem and encourage a discussion of it in empirical cross-country studies in conflict and peace economics. Second, by way of illustration, in the Appendix to this article we provide overview tables of some of the gravest discrepancies in country coding across datasets which facilitate quick cross-dataset comparisons of country units.

A typology of inconsistencies

Inconsistent country names are the tip of the merging iceberg. Not only do names differ, but so does for example the period of existence for some countries. And worse, the documentation on the country coding schemes provided by the data projects is often sparse and contains errors.[4]

The following three types of inconsistencies between country units in different data sources and coding schemes are frequently observed and examined in this article.

Inconsistency type 1: *a state name exists in one dataset but not in the other*. There are several reasons for this, shown here in schematic fashion:

Reason i: Different years (time series do not match and some states do not exist anymore/yet).
Example: When merging PolityIV with Comtrade data the Orange Free State cannot be merged as it ceases to exist before coding of Comtrade data starts.
Result: Country is unmergeable and drops out of analysis because it does not exist in one dataset.

Reason ii: Different definition of statehood.
Example: Some datasets do not code Palestine as they do not consider it to meet formal requirements of statehood.
Result: Country is unmergeable and drops out of analysis because it does not exist in one dataset.

Reason iii: Different state names (labels) or entities/territories (see the third inconsistency described below).
Example: Yugoslavia and its successors are coded in vastly different ways in terms of names and years across datasets. How should these countries or observations be aggregated to make them comparable across datasets and to not loose conflict observations?
Result: Country may drop out of analysis if no action is taken.

The contribution of this article is twofold. First, it shows that in spite of all country coding scheme standardization efforts and relevant software packages and codes, the problem of inconsistent country coding in merging diverse macro panel datasets persists. This can lead to substantial numbers of "missing" values in merged datasets and possibly affect the reliability of inferences drawn from statistical analysis. This is of particular concern in empirical analysis in conflict and peace economics as inconsistent country coding often affects countries in conflict. Second, by way of illustration, we provide overview tables of some of the gravest discrepancies in country coding across datasets.

Inconsistency type 2: *a country is coded under the same name, but for different years in two datasets (time series for given country are not identical in both datasets)*. Again, in schematic fashion:

Reason i: Missing observations within time series.
Example: In V-Dem, Germany, 1945–1948, is not coded since the institutional framework of Germany during those years does not meet the formal criteria for the definition of their democracy indices.
Solution: Depends on application and on underlying assumptions made about reason for missingness, possibly interpolation.

Reason ii: Country starts or ceases to exist and first/last year is not coded consistently across datasets.
Example: PolityIV codes the former East Germany between 1945–1990, whereas V-Dem codes it from 1949–1990.
Solution: Depends on application, possibly extrapolation.

Inconsistency type 3: *a country is coded under different names either (a) for the same years in two datasets or (b) for different years in two datasets*.

Reason i: It is clearly the same state, only the label is different. This is often the case for 3(a), or for 3(b) in combination with inconsistency type 2, reason ii.
Example: "St." versus "Saint" or official versus colloquial state names ("Plurinational State of Bolivia" and "Bolivia").
Solution: Use Stata and R packages for renaming.

Reason ii: The different names might refer to different underlying entities/ territories.
Example: We provide detailed overviews of these cases

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

in Table A3 (Democracy Datasets) and Table A6 (Economic Datasets) of the Appendix.
Solution: The 3(b) case is by far the most difficult case as the years coded do not provide additional evidence on the actual entity captured. The question of how these entities could be compared in a meaningful way across datasets has no straightforward answer; rather, the answer is case dependent.

Inconsistent country coding of types 1 to 3 lead to missing values in the final, merged dataset.[5] In this article we show that the extent of these "missing values" (they are not really missing, just missing due to inconsistencies) is vast and of particular relevance to empirical research in conflict and peace economics. Most country coding schemes differ in the naming and dating of a specific set of countries: Countries which have experienced armed conflict are less democratic and less trade open than the consistently coded ones. As a result, a merged dataset can contain a comparatively high share of missing values for this set of countries. Thus, it can no longer be considered a random sample. To minimize "missings," and to avoid losing valuable information, the process of creating large panel datasets should therefore be done with utmost care.

In general, there are three approaches to code countries in macro panel data: By (string) country names, by numeric code, or by alphabetic code. The most common schemes include (but are not limited to) the COW country list, the Gleditsch/Ward state list, and the ISO 3166 list of country codes.[6] In theory, numeric and alphabetic codes should facilitate the merging process. Unfortunately, several numeric and alphabetic codes schemes exist and often they are neither implemented consistently nor are the country codes easily translatable to each other. In R the package "countrycode" and in Stata the package "kountry" help with these issues.[7] These packages map country names and codes from one kind of macro country codes to another. They come with a slight disadvantage, though, as "[t]he mapping between the available dataset_names [types of country coding schemes] is not always perfect."[8] This is especially dire when using a comparatively new dataset such as V-Dem which does not follow any of the coded country schemes exactly. In addition, this assumes that each source dataset correctly applies the country coding scheme it is based on. In the following sections we show that this is not the case for several datasets. By letting Stata or R packages adjust the country names, the renaming—and subsequently the merging process—is put into a black box, inherently making it more vulnerable to mistakes.

We aim to take this data merging process out of its black box and use actual country names to prevent merging mistakes. In what follows we provide a detailed comparison of six datasets covering the indicators trade, democracy, and conflict. For each dataset a table with actual country names and years in the data is provided (see Boese and Kamin, 2018a, 2018b). These tables present an overview of the gravest discrepancies in country coding and allow for quick cross-dataset comparisons of country units. In addition, this article gives an overview of the extent of the country coding problem by comparing structural properties of the set of inconsistently coded countries to those of the uniformly coded ones and by discussing missing data as well as differences in annual coding.

On the one hand, this article provides assistance to scholars merging several source datasets. On the other, it highlights naming inconsistencies between data documentation, such as code books, and actual observations in the data. Such inconsistencies potentially lead to merging problems when blindly using the Stata or R packages (and the country coding scheme specified in the documentation) discussed above. We have the highest respect for all the data projects discussed in this article. We therefore hope that the lists of these inconsistencies are also of assistance to the data projects in aligning their documentation to their respective datasets.

The following three sections respectively provide thorough comparisons of two democracy, two trade, and two conflict datasets, including detailed tables comparing the country coding units. The article closes with a discussion of the results.

Democracy data

This section compares the country coding units of two democracy datasets: V-Dem version 8 and the PolityIV dataset 2016. The tables referenced in this section can be found in the Appendix as well as in Boese and Kamin (2018a).

We first discuss the countries listed in V-Dem version 8, then discuss the countries in the PolityIV dataset 2016, and then compare characteristics of the observations listed in both datasets with those listed in only one of the datasets.

V-Dem Data version 8

The V-Dem dataset used for this article is V-Dem data version 8, in country year format. The variable of interest is the Electoral Democracy Index, v2x_polyarchy. V-Dem identifies the countries either by name, alphabetical country id, or numerical country id.[9] These country identifiers do not correspond to any of the prevailing country schemes implemented in the Stata or R packages mentioned above. To facilitate the merging process, we therefore provide a detailed list of county coding units in the data[10] and compare it to the country list in the V-Dem code book (Coppedge, *et al*., 2018a).

V-Dem excels in terms of transparency and provides a

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

supplementary article on "V-Dem Country Coding Units v8" which lists and discusses all polities and countries and the respective years for which they are coded as well as a detailed explanation of the country borders used in the coding.[11] It also provides detailed information on years in which a country is not coded (with the variables gapstart and gapend). However, there are several observations for which v2x_polyarchy is missing. Worksheet "Overview" in Boese and Kamin (2018a) shows the number of years for which each country is coded in V-Dem version 8, as well as its gaps (by coding decision) and its additional missing values.

For ten countries the names in dataset and documentation do not match.[12] These name mismatches are by no means a purely alphabetical problem. Take, for example, Vietnam. While there is no country named Vietnam, North or South, in the V-Dem dataset there is a "Republic of Vietnam" (coded from 1802–1975) and a "Democratic Republic of Vietnam" (coded from 1945–2017). The V-Dem Country Coding Units document, however, provides a detailed overview of the polities forming part of:

> "Vietnam, South (35)
>
> > Coded: 1802–1975. History: (...) Republic of Vietnam (also known as South Vietnam) (1955–1975)" and
>
> "Vietnam, North (34)
>
> > Coded: 1945– History: Democratic Republic of Vietnam (i.e. North Vietnam) [declared] (1945); Democratic Republic of Vietnam (1945–1949); Democratic Republic of Vietnam [independent state] (1949–). Note: From 1976, the polity also includes areas formerly belonging to Republic of Vietnam (South Vietnam)."[13]

Take another example. In the documentation the numerical country id (365) is coded for two countries: Oldenburg, 1789–1867, and Saxe-Weimar-Eisenach, 1809–1867. In the dataset, however, only Saxe-Weimar-Eisenach is assigned country_id 365 while Oldenburg is assigned code 364.

Table 1: Number of (un)mergeable countries in a merge of the PolityIV dataset wih the COW country list

Merging by	*Country name*	*Numeric code*	*Alphabetic code*
Unmergeable no. of countries in PolityIV	26	11	19
Mergeable no. of countries in PolityIV and COW	169	183	177

Table 2: Description of democracy datasets

Dataset	*A: V-Dem*	*B: PolityIV*
Total no. of obs	26,537	17,228
Total no. of nonmissing obs	24,115	16,992
No. of countries	201	195
Years covered	1789–2017	1800–2016

Table 3: Merging V-Dem and PolityIV data

Merging observations	*A: V-Dem*	*B: PolityIV*
Unmergeable only in A	10,929	n/a
Unmergeable only in B	n/a	1,619
Mergeable in both	15,609	
Nonmissing only in A	9,380	n/a
Nonmissing only in B	n/a	1,571
Nonmissing, mergeable in both	14,376	15,421

Table 4: Two sample *t*-tests of average level of democracy

Dataset	*A: V-Dem*	*B: PolityIV*
Unmergeable group	0.1377	–1.5493
Mergeable group	0.3428	–0.4495
Difference	0.2051***	1.0998***

Note: *** Statistically significant at the 1% level.

PolityIV

A second dataset, capturing political authority patterns worldwide and over long periods of time, is the PolityIV project's dataset on "Political Regime Characteristics and Transitions, 1800–2016" (for short, the PolityIV dataset).[14] In the dataset countries are identified by their name, an alphabetic country code, or a numeric code.[15] These identifiers supposedly follow the COW country coding scheme.[16] Table 1 displays the results from merging the PolityIV data with the COW country list, finding that 13 percent of the countries are unmergeable when merging by country name, 6 percent when merging by numeric code, and 10 percent when merging by alphabetic code.[17] The unmergeable groups largely consist of countries of particular interest in conflict and peace economics such as the Koreas, Congos, Germanies, and Serbias. As a consequence, when merging the PolityIV data using a software package taking the dataset to be in "COW coding scheme" these countries may not be properly dealt with. It is worth noting that

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

country names and alphabetic and numeric codes are not coded consistently over time within the PolityIV dataset, i.e., there are 195 different country names, but only 194 different alphabetic and numeric codes. This is not due to a single country having different names and only one code, but to a number of countries and several code/label constellations. Examples include Yugoslavia (either *ccode* 345 and *scode* YUG or *ccode* 347 and *scode* YGS; the fact that 347 and YGS also are used for Serbia and Montenegro in the dataset further complicates matters), Ethiopia (either *ccode* 529 and *scode* ETI or *ccode* 530 and *scode* ETH), Pakistan (either *ccode* 769 and *scode* PKS or *ccode* 770 and *scode* PAK). Further, *ccode* 860 and *scode* ETM is used for East Timor and Timor Leste, and *ccode* 255 and *scode* GMY is used for Germany and Prussia.

Additionally, in the PolityIV dataset we note duplicate observations for Yugoslavia in 1991 and for Ethiopia in 1993. This further complicates the merging process as the scholar is forced to decide how to proceed with these duplicates.

Comparison of the democracy data

Table 2 describes both democracy datasets. The variable of interest in each dataset is a democracy index: *v2x_polyarchy* for the V-Dem data and *polity2* for the PolityIV data.[18] The total number of nonmissing observations refers to the number of observations for which the respective variable of interest contains nonmissing values.

When merging the datasets by country name and year, observations of inconsistency types 1 to 3 cannot be merged. Table 3 shows the number of mergeable and unmergeable observations by source dataset. As discussed, even though an observation might be listed, the variable of interest can contain a missing value. Hence the lower half of Table 3 proves the same information for all observations with nonmissing values. To make the number of observations comparable across datasets in Table 3, only observations from the time period covered by both datasets are considered (that is, V-Dem observations before 1800 as well as the year 2017 were left out to match the PolityIV time series). Around 41 percent of the V-Dem and around 9 percent of the PolityIV observations cannot be merged. To assess whether the unmergeable observations are systematically different from the mergeable ones we calculated the average levels of democracy for each group. Table 4 shows the results of two *t*-tests, one for V-Dem, one for PolityIV. In both datasets, the unmergeable group had a significantly lower average level of democracy. (To be clear, the *t*-tests were carried out only on the nonmissing observations noted in Table 3.)

Economic data

UN Comtrade and the World Bank's World Development Indicators (WDI) contain economic data. We first discuss the countries listed in the UN Comtrade data, then those in the WDI, and then compare the country coding schemes of both datasets. The tables and worksheets referenced to in this section can be found in the Appendix as well as in Boese and Kamin (2018b).

UN Comtrade

The indicator taken from UN Comtrade is total exports in current U.S. dollars from each country to the rest of the world. The Comtrade dataset is an unbalanced panel as it only contains years for which countries have reported trade. Hence, time series differ from country to country. The first year for which some countries reported trade is 1962, the last year is 2017 (few observations are available for the start and end years of the time series). Comtrade offers data coded according to two different systems for international trade statistics: The Harmonized System (HS), introduced in 1988, and the Standard International Trade Classification (SITC), introduced in 1962, with the latter being less detailed than the former. To obtain the longest possible time series, we concatenated SITC classification export data, 1962–1987, with HS classification export data, 1988–2017.

In addition to gaps in the time series caused by missing observations (as discussed above) the export variable contains missing values for several observations. Missing information primarily indicates that trade was not reported and is not to be equated with zero trade flows.[19] This is crucial concerning the tackling of zero trade flows and appropriate model choice.[20]

The country name abbreviations of the official UN country list[21] correspond to the country names used in the Comtrade data with the exception of Côte d'Ivoire and Réunion, which contain spelling errors in the downloaded Comtrade dataset ("C√¥te d'Ivoire" and "R√©union").

World Development Indicators

The economic indicator taken from the World Bank's WDI is trade openness, defined as the percentage share of trade of each country's GDP, that is, (imports+exports)/GDP. Starting in 1960, the time series runs to 2016. The distinction between zero trade and missing data in the WDI is equivalent to the one in UN Comtrade. In contrast to Comtrade, however, the WDI data is a balanced panel with one observation for each country and year. Nevertheless, trade openness contains missing values for several observations due to missing information on GDP, exports, or imports. In addition to countries, WDI provides aggregated information on country groups (such as "Europe &

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

Central Asia" or "Low & Middle Income"). These where taken out of the list to facilitate reading (the full list of country groups removed is available in Boese and Kamin, 2018b, worksheet "Disregarded Country Groups").

To our knowledge, the World Bank does not provide an explicit country coding scheme upon which WDI data are based. However, the World Bank does provide a list of countries upon which the World Integrated Trade Solution (WITS) data are based.[22] It is unclear whether this list also forms the basis of the WDI dataset. Of 15,048 observations in the WDI dataset used in this article, 30 percent (4,560 observations) do not match the WITS list. Several of them are due to naming inconsistencies such as, for example, "Bahamas, The" versus "Bahamas".

Comparing the economic data

In a comparison of the economic datasets[23] the sheer number of naming inconsistencies[24] and single appearances of countries (that is, they appear in one, but not in the other dataset)[25] stands out. Additional cases, difficult to handle when merging datasets, are countries that started and ceased to exist, yielding different country names for different or the same territories and for different years (inconsistency type 3). While WDI refers to each country under one name continuously for the entire time series, this is not the case for the UN Comtrade data. In Comtrade, countries are coded by different names and years. Table A6[26] displays the cases where this kind of inconsistency is in place. The table shows that Comtrade distinguishes the underlying country entities in much more detail. There is, for example, only one "Germany" in the WDI data as opposed to "Germany", "Fmr Fed. Rep. of Germany" and "Fmr Dem. Rep. of Germany" in the UN Comtrade data.

Assuming that the ending of one state and the beginning of a new one are coded in detail through the year variable by WDI, can the country coding units be supposed to be the same across the two datasets? The sparsity of country coding unit documentation renders it impossible to answer this question. There is no information on whether territories changed, and on whether or how much this change was incorporated in the coding. This becomes a severe drawback to the data when complementary variables for the analysis of trade flows, such as country size, GDP, measures of distance and—most importantly—borders are taken into account.[27]

Table 5: Description of trade datasets

Dataset	*A: Comtrade*	*B: WDI*
Total no. of obs	12,768	15,048
Total no. of nonmissing obs*	6,790	10,643
No. of countries	228	264
Years covered	1962–2017	1960–2016

Note: *The total number of nonmissing observations refers to the number of observations for which the respective variable of interest contains nonmissing values.

Table 6: Merging Comtrade and WDI data

Merging observations	*A: Comtrade*	*B: WDI*
Unmergeable only in A*	3,803	n/a
Unmergeable only in B	n/a	6,083
Mergeable in both	8,965	
Nonmissing only in A	1,449	n/a
Nonmissing only in B	n/a	3,765
Nonmissing, mergeable in both	5,341	6,878

Note: *When merging both datasets by country name and year those observations of inconsistencies types 1 to 3 are unmergeable.

Table 7: Two sample *t*-tests of average level of trade and trade openness

Dataset	*A: Comtrade**	*B: WDI***
Unmergeable group	2.72 x 10^{14}	66.16
Mergeable group	3.98 x 10^{13}	76.14
Difference	–2.32 x 10^{14}***	9.98***

Note: *The trade variable in Comtrade is total exports (*TradeValueUS*), range: USD37,310–2.34x10^{16}. **The trade variable in WDI is trade openness (*tradeop*), range: 0–860.8 (in %). *** Statistically significant at the 1% level. The *t*-tests were carried out on the nonmissing observations in Table 5.

The case of Sudan (see Table A6)[28] illustrates the problem: WDI codes "South Sudan" and "Sudan". For the latter, the measure of trade openness is available for the whole time series (1960–2016). For "South Sudan", the indicator is available from 2008–2015. UN Comtrade codes "Sudan" (2012–2015) and "Former Sudan" (1963–2011, with gaps). Hence, WDI takes 2008 as the year of birth for "South Sudan", while Comtrade (implicitly, because it does not code "South Sudan" as a country)[29] codes a new state "Sudan" from 2012 onward. Similar cases are Serbia (with or without data for Kosovo or Montenegro) and China (with or without data for Hong Kong, Macao, and Taiwan).[30]

The country name by itself does not allow for an exact indication of the territory coded. In a statistical analysis only of

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

trade, it might not matter whether Sudan or South Sudan is included. In conflict and peace economics, however, where relationships among conflict, politics, and economics are of high interest, such lack of accuracy effectively becomes an impediment to an appropriate econometric analysis.

Table 5 describes both trade datasets. For Comtrade, the variable of interest is total exports in current U.S. dollars (*TradeValueUS*); for the WDI data, it is trade openness as a percentage of GDP (*tradeop*). Table 6 shows the number of mergeable and unmergeable observations by source dataset. As discussed, even though an observation might be listed the variable of interest can contain a missing value. Hence the bottom half of Table 6 provides the same information for all observations with nonmissing values. To make the number of observations comparable across datasets in Table 6 only observations from the time period covered by both datasets are considered, i.e., 1962–2016. About 30 percent of the Comtrade observations, and about 40 percent of the WDI observations, cannot be merged.[31] To assess whether the unmergeable observations are systematically different from the mergeable ones, we calculated average levels of total exports and trade openness for each group. Table 7 shows the results of two sample *t*-tests: For Comtrade, the average export level is statistically significantly *higher* (given the exponent) in the unmergeable than in the mergeable group. For WDI, the unmergeable country group had a significantly *lower* level of average trade openness. Looking at the naming inconsistencies (Table A4) confirms this "higher-lower" difference: The high levels of export values in the unmergeable group in Table 7 are driven by observations from the U.S., Germany, Macao, and Hong Kong.[32] Table 7 hence provides a good intuition to the effects of inconsistent country coding: Either the cases of high export levels or of low trade openness are lost due to merging problems. Either one is problematic in terms of statistics and, depending on the analytic aim, might lead to biased estimates.

Conflict data

In theory, the datasets for economic and political variables code each variable for all years during which a country exists. The conflict datasets, however, are fundamentally different: By design, they only code conflict variables for years in which a conflict occurred in a given country and which surpassed some conflict criteria (for example, 25 battle-related deaths). Consequently, time series and cross-section data dimensions contain gaps for country-years without armed conflict.

The UCDP Armed Conflict dataset version 18.1 (Pettersson and Eck, 2018; also see Gleditsch, *et al*., 2002; UCDP, 2018) studies armed conflict above a yearly threshold of 25 battle-related deaths. The Militarized Interstate Disputes (MID) B dataset version 4.2 (Palmer, *et al*., 2015) captures militarized interstate disputes which can involve, for example, a display of force without incurring any battle deaths. Therefore, the gaps in the datasets will be very different, and merging them by country and years coded does not provide insights on, or a comparison of, country coding units. Nevertheless, both datasets acknowledge the importance of defining country coding units. In the remainder of this section, we show that even within each of these datasets there are inconsistencies between the country coding units as defined by the respective data project and the actual observations in the data. As a result, these observations are either dropped, potentially falsely matched, or have to be manually adjusted when using Stata or R commands for merging countries.

UCDP/PRIO Armed Conflict dataset version 18.1

The UCDP/PRIO Armed Conflict dataset acknowledges the importance of a precise description of country coding units[33] and dedicates an entire section of its code book[34] to the exact definition of country coding units. It includes a country table with numerical and alphabetical country codes, state names, and start and end years for the countries that form part of the international system of states.

Table A7 lists the countries coded in the actual data and compares them to the system membership table from the UCDP/PRIO code book. The system membership table must include more observations since, by definition, it also includes countries without armed conflict. But Table A7 shows that even when restricted to countries with armed conflict there are inconsistencies in the country names (for example "Burkina Faso" and "Burkina Faso (Upper Volta)", "DR Congo (Zaire)" and "Congo, Democratic Republic of (Zaire)", and "Ivory Coast" and "Cote D'Ivoire").

MID B version 4.2

The MID B version 4.2 dataset includes one observation per participant to a militarized dispute, 1816 –2010, with countries taken from the Correlates of War (COW) list. The MID B dataset itself does not contain (string) state names. Instead, countries are coded with a three-digit numerical code (*ccode*) and with an alphabetical code (*stabb*). Before joining variables from the MID B dataset with any other macro panel data, such as WDI, a first step therefore is to merge MID B with COW, but four countries cannot in fact be merged (Table 8). The three-digit alphabetic codes for these countries are RUM, USR, VTM, and ZAI. This is a perfect example of the difficulties associated with merging by country as it is hardly possible to determine with certainty which underlying entity (territory) is exactly covered, for example, by USR or VTM. This also

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

illustrates why, for this article, we chose to employ merging by country (string) names, not codes. VTM could stand for (Democratic) Republic of Vietnam, Vietnam North, Vietnam South, or Vietnam. While the exact entity coded remains unclear, it is very clear that this case contains information relevant for studies of conflict.

That the MID B dataset states that it follows the COW country list convention when in fact it does not, makes it effectively impossible to determine for some observations which actual underlying entity is considered a country during which period of time.

Table 8: Number of unmergeable countries in a merge of the MID B dataset with the COW country list

Merging by	*Numeric code*	*Alphabetic code*
Unmergeable countries in MID B	4	4
Mergeable countries in MID B and COW	191	191

Discussion and conclusion

Large-scale cross-country datasets are frequently merged in quantitative studies in conflict and peace economics. We find that the coding of country units overlaps across datasets only for a relatively small proportion of countries. Discrepancies in country naming or other forms of country identification such as numerical or alphabetical country IDs are frequent among countries splitting up or (re)uniting during the time period studied. Examples include Yugoslavia, Germany, Vietnam, and Sudan. If the names are not adjusted, these inconsistencies render such observations unmergeable and, when joining variables from several data sources, ultimately result in missing values. When these missing values then are dropped from an analysis, important information is lost. This loss of information is of particular severity in conflict and peace economics as countries which split up or reunite often do so accompanied by armed conflict and thus contain valuable information.

The dataset comparisons made in this article demonstrate that inconsistencies in country coding across macro panel datasets remain a relevant challenge in cross-national studies. They show that for economic datasets as well as democracy datasets the unmergeable group is of a large size (up to about 40 percent of all observations) and significantly differs from the group of mergeable observations. In particular, the group of unmergeable countries is on average less democratic than the mergeable group. Depending on the economic measure analyzed (and, with it, the country naming scheme applied), a group of countries with high exports or another group of countries with low trade openness cannot be merged.

These discrepancies can be attributed, in part, to differences in country labels. Several projects, such as Hensel (2016) and the aforementioned software codes and packages can help adjust them. However, another part of the inconsistent country coding is due to different perceptions and definitions of the unit of analysis. The exercises carried out for this article show that the actual entity captured can differ by source dataset. While this makes creating merged panel datasets consisting of economic, political, or armed conflict factors challenging in its own right, proper merging might be a necessary condition for analysis. For an armed conflict dataset, relevant state units might differ significantly from datasets on democracy or trade flows (the coding of Palestine, Hong Kong, or Macao are examples). As a result, the burden of discussing the unit of analysis studied and of ensuring that countries correspond to the same entity across merged datasets, lies with the individual scholar or team. This article encourages scholars to discuss the merging process in their academic papers (or supplementary materials) and to not take the problem of inconsistent country names lightly. This is particularly the case in conflict and peace economics, where relevant information is systematically lost when unmergeable observations are discarded.

Furthermore, it is worth noting that country names are not the only dimension of macro panels to be carefully compared across datasets before merging. It goes well beyond the scope of this article to additionally compare the actual time periods covered. However, we point out that the time dimension underlying the calendar year coding of macro panels does not necessarily coincide with the actual calendar year. To quote from the World Bank: “In most economies the fiscal year is concurrent with the calendar year ... Most economies report their national accounts and balance of payments data using calendar years, but some use fiscal years.” Time inconsistencies, then, are another potential source of erroneous inference, in particular when studying the effect of conflict on the economy or the political system, or vice versa.[35]

Last, but not least, we pay tribute to the creators of the datasets discussed in this article: Assembling and maintaining these datasets is a Herculean task. The challenges associated with inconsistent country names and units across datasets can, however, lead to serious consequences in conflict and peace economics. Unfortunately, while an easy solution to the noted problems is not likely to exist, given the different purposes each of the source datasets is created for, we hope that our comments here increase broader awareness and discussion of these problems and that our tables in the Appendix (and online)

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

facilitate quick cross-dataset comparisons of country coding.

Notes

Acknowledgments: Vanessa A. Boese is funded by the Cusanuswerk. Both authors thank Scott Gates, Philipp Großkurth, and an anonymous reviewer for excellent comments and suggestions. All remaining errors are our own. Contributing authors are listed alphabetically. All work was shared equally.

[1] Examples of studies using such merged datasets include Hegre, *et al*. (2001), Blomberg and Hess (2006), Gates, *et al*. (2006), Martin, Mayer, and Thoenig (2008), Glick and Taylor (2010), Acemoglu, *et al*. (2019), Dunne and Tian (2015), and d'Agostino, Dunne, and Pieroni (2018).

[2] Hence the title of this article. 'Tis but they name that is my enemy (Romeo and Juliet, Act II, Scene ii, Shakespeare, 2003).

[3] See http://heatherba.web.unc.edu/data-code/.

[4] For discussion, see the sections on democracy, economic, and conflict data in this article.

[5] Note the difference between *missing values* and *missing observations*. For example, on the one hand, in the V-Dem dataset version 8 there are no observations for Germany between 1945 and 1948, leaving the panel unbalanced. In the World Development Indicators, on the other hand, the panel provided is balanced, that is, there is one observation for each country in each year. However, for a number of years the variable of interest contains a missing value. Ultimately, when merging two such sources and using the final dataset for statistical analysis, missing values and missing observations come down to the same thing: *missing information*. For most regressions or other analyses, software like Stata disregards observations whenever they contain missing values.

[6] COW: A country coding scheme employed by several of the macro panel datasets studied in this article. Data can be obtained from http://www.correlatesofwar.org/data-sets/cow-country-codes. There are three variables: numeric and alphabetic country codes and statename. The dataset covers 217 countries. The country list includes 26 duplicate observations. Gleditsch/Ward: The Gleditsch and Ward (1999) state list builds on and revises the COW country list. First published in 1999, a current version is available at http://ksgleditsch.com/statelist.html. ISO: See https://www.iso.org/iso-3166-country-codes.html.

[7] R: See https://cran.r-project.org/web/packages/countrycode/countrycode.pdf. Stata: See Raciborski (2008).

[8] Quote: Raciborski (2008, p. 392). Raciborski (2008) continues with a short overview of the most striking inconsistencies.

[9] Alphabetical country_text_id: "Abbreviated country names," V-Dem Codebook v8, p. 36. Numerical country_id: "Unique country ID designated for each country. A list of countries and their corresponding IDs used in the V-Dem dataset can be found in the country table in the codebook, as well as in the V-Dem Country Coding Units document." V-Dem Codebook v8, p. 36. The *codebook* itself is Coppedge, *et al*. (2018a). The *country coding units document* is Coppedge, *et al*. (2018b).

[10] See Boese and Kamin (2018a), worksheet "V-Dem Codebook vs. Data".

[11] See Coppedge, *et al*. (2018b).

[12] These are: Democratic Republic of Congo, Democratic Republic of Vietnam, German Democratic Republic, Mecklenburg Schwerin, North Korea, Republic of Vietnam, Republic of the Congo, South Korea, São Tomé and Príncipe, and Timor-Leste.

[13] Coppedge, *et al*. (2018b, p. 27).

[14] See Marshall, Gurr, and Jaggers (2017b).

[15] Alphabetic: The variable *scode* ("Alpha Country Code: Each country in the Polity IV dataset is defined by a three-letter alpha code, derived from the Correlates of War's listing of members of the interstate system" (Marshall, Gurr, and Jaggers, 2017a, p. 12). Numeric: *ccode* (numerical, "Numeric Country Code: Each country in the Polity IV dataset is defined by a three-digit numeric code, derived from the Correlates of War's listing of members of the interstate system" (Marshall, Gurr, and Jaggers, 2017a, p. 11).

[16] Supposedly: See Marshall, Gurr, and Jaggers (2017a, p. 11).

[17] To be clear, the share of unmergeable countries is calculated as: number of unmergeable countries/ total number of countries in PolityIV (i.e., 26/195~13.3%, 11/194~5.7%, and 19/194~9.8%. Note that the rows are labeled correctly although one could in fact omit "and COW" from the second row since, if countries are mergeable in a merge between COW and PolityIV, they must exist in both datasets. In the first row, however, are unmergeable countries only, i.e., those which exist only in the PolityIV dataset.

[18] V-Dem's *v2x_polyarchy*: Range 0 to 1 (most democratic). PolityIV's *polity2*: Range –10 to +10 (most democratic).

[19] For a discussion of missings in trade data see, for example, Keshk, Reuveny, and Pollins (2010, Section 3.3, p. 10), Barbieri, Keshk, and Pollins (2009, p. 476), and Boehmer, Jungblut, and Stoll (2011).

[20] See, for example, Santos Silva and Tenreyro (2006).

[21] The UN provides a list of country codes and names at https://unstats.un.org/unsd/tradekb/Knowledgebase/50377/Comtrade-Country-Code-and-Name.

[22] https://wits.worldbank.org/wits/wits/witshelp/content/codes/country_codes.htm.

[23] See Boese and Kamin (2018b), worksheet "Overview".

[24] See Table A4 or Boese and Kamin (2018b), worksheet "naming inconsistencies" for inconsistency type 3, reason i (one country coded with different names but for the same year and years).

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

[25] See Table A5 or Boese and Kamin (2018b), worksheet "existence asymmetry" for inconsistency types 1 and 3.

[26] Also see Boese and Kamin (2018b), worksheet "inconsistency type 3".

[27] Anderson and van Wincoop (2003), for example, demonstrated that national borders are a highly important impediment to trade.

[28] Boese and Kamin (2018b), worksheet "inconsistency 2.0", rows 36–38.

[29] The fact that no "South Sudan" is included in the UN Comtrade data is itself somewhat astonishing since trade data is available (otherwise WDI would not be able to code it).

[30] See World Bank (2017a, p. XVII).

[31] Again, to be clear: 3,803/(3,803+8,965)~29.7% and 6,083/(6,083+8,965)~40.4%.

[32] This is shown in Boese and Kamin (2018b), worksheet "Unmergeable Outliers Comtrade". It contains all unmergeable Comtrade observations sorted by export values (highest first) to show the outliers driving the results.

[33] "The definition of a state is crucial to the UCDP/PRIO conflict list" (UCDP/PRIO Armed Conflict Dataset Codebook, 2018, p.13).

[34] See Section 4: "System Membership Description" (UCDP/PRIO Armed Conflict Dataset Codebook, 2018, p. 13).

[35] Quote from World Bank (2017b, p. 117).

References

Acemoglu, D., S. Naidu, P. Restrepo, and J.A. Robinson. 2019. "Democracy Does Cause Growth." Journal of Political Economy. Vol. 127, No. 1, pp. 47–100. https://doi.org/10.1086/700936

Anderson, J. and E. Van Wincoop. 2003. "Gravity with Gravitas: A Solution to the Border Puzzle." *American Economic Review*. Vol. 93, No. 1, pp. 170–192. https://doi.org/10.1257/000282803321455214

Barbieri, K., O.M.G. Keshk, and B.M. Pollins. 2009. "Trading Data: Evaluating our Assumptions and Coding Rules." *Conflict Management and Peace Science*. Vol. 26, No. 5, pp. 471–491. https://doi.org/10.1177/0738894209343887

Blomberg, S.B. and G.D. Hess. 2006. "How Much Does Violence Tax Trade?" *Review of Economics and Statistics*. Vol. 88, No. 4, pp. 599–612. https://doi.org/10.1162/rest.88.4.599

Boehmer, C.R., B.M.E. Jungblut, and R.J. Stoll. 2011. "Tradeoffs in Trade Data: Do our Assumptions Affect our Results?" *Conflict Management and Peace Science*. Vol. 28, No. 2, pp. 145–167. https://doi.org/10.1177/0738894210396630

Boese, V.A., and K. Kamin. 2018a. "Democracy Datasets.xslx." https://www.dropbox.com/s/3bm674tjk9 iqha4/Democracy%20Datasets.xlsx?dl=0.

Boese, V.A., and K. Kamin. 2018b. "Economic Datasets.xslx." https://www.dropbox.com/s/qtaol2upv0uns30/Economic %20Datasets.xlsx?dl=0.

Coppedge, M., J. Gerring, C.H. Knutsen, S.I. Lindberg, S.-E. Skaaning, J. Teorell, D. Altman, M. Bernhard, A. Cornell, M.S. Fish, H. Gjerløw, A. Glynn, A. Hicken, J. Krusell, A. Lührmann, K.L. Marquardt, K. McMann, V. Mechkova, M. Olin, P. Paxton, D. Pemstein, B. Seim, R. Sigman, J. Staton, A. Sundström, E. Tzelgov, L. Uberti, Y. Wang, T. Wig, and D. Ziblatt. 2018a. "V-Dem Codebook v8". Varieties of Democracy (V-Dem) Project. https://doi.org/10.23696/vdemcy18

Coppedge, M., J. Gerring, C.H. Knutsen, S.I. Lindberg, S. Skaaning, J. Teorell, V. Ciobanu, and M. Olin. 2018b. "V-Dem Country Coding Units V8." V-Dem Working Paper. Forthcoming. http://dx.doi.org/10.2139/ssrn.3172795

d'Agostino, G., J.P. Dunne, and L. Pieroni. 2018. "Military Expenditure, Endogeneity and Economic Growth." *Defense and Peace Economics*. [Published online 18 January 2018. https://doi.org/10.1080/10242694.2017.1422314

Dunne, J.P. and N. Tian. 2015. "Military Expenditure, Economic Growth and Heterogeneity." *Defense and Peace Economics*. Vol. 26, No. 1, pp. 15–31. https://doi.org/10.1080/10242694.2013.848575

Gates, S., H. Hegre, M.P. Jones, and H. Strand. 2006. "Institutional Inconsistency and Political Instability: Polity Duration, 1800–2000." *American Journal of Political Science*. Vol. 50, No. 4, pp. 893–908. https://doi.org/10.1111/j.1540-5907.2006.00222.x

Gleditsch, K.S. and M. D. Ward. 1999. "A Revised List of Independent States since the Congress of Vienna." *International Interactions*. Vol. 25, No. 4, pp. 393–413. https://doi.org/10.1080/03050629908434958

Gleditsch, N.P., P. Wallensteen, M. Eriksson, M. Sollenberg and H. Strand. 2002. "Armed Conflict 1946-2001: A New Dataset." *Journal of Peace Research*. Vol. 39, No. 5, pp. 615–637. https://doi.org/10.1177/0022343302039005007

Glick, R. and A.M. Taylor. 2010. "Collateral Damage: Trade Disruption and the Economic Impact of War." *Review of Economics and Statistics*. Vol. 92, No. 1, pp. 102–127. https://doi.org/10.1162/rest.2009.12023

Hegre, H., T. Ellingsen, S. Gates, and N.P. Gleditsch. 2001. "Toward a Democratic Civil Peace? Democracy, Political Change, and Civil War, 1816–1992." *American Political Science Review*. Vol. 95, No. 1, pp. 33–48.

Hensel, P. 2016. "ICOW Historical State Names Data Set, Version 1.2." Issue Correlates of War (ICOW) Project. http://www.paulhensel.org/icownames.html.

Keshk, O.M.G., R. Reuveny, and B.M. Pollins. 2010. "Trade and Conflict: Proximity, Country Size, and Measures." *Conflict Management and Peace Science*. Vol. 27, No. 1, pp. 3–27. https://doi.org/10.1177/0265659009352137

Marshall, M.G., T.R. Gurr, and K. Jaggers. 2017a. "Polity IV Project: Dataset Users' Manual—Political Regime

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

Characteristics and Transitions, 1800–2016." Available at http://www.systemicpeace.org/inscrdata.html.

Marshall, M.G., T.R. Gurr, and K. Jaggers. 2017b. "Polity IV Project: Political Regime Characteristics and Transitions, 1800–2016." http://www.systemicpeace.org/inscrdata.html.

Martin, P., T. Mayer, and M. Thoenig. 2008. "Make Trade not War." *Review of Economic Studies*. Vol. 75, No. 3, pp. 865–900. https://doi.org/10.1111/j.1467-937X.2008.00492.x

Palmer, G., V. D'Orazio, M. Kenwick, and M. Lane. 2015. "The Mid4 Dataset, 2002–2010: Procedures, Coding Rules and Description." *Conflict Management and Peace Science*. Vol. 32, No. 2, pp. 222–242. https://doi.org/10.1111/j.1467-937X.2008.00492.x

Pettersson, T. and K. Eck. 2018. "Organized Violence, 1989–2017." *Journal of Peace Research*. Vol. 55, No. 4, pp. 535–547. https://doi.org/10.1177/0022343318784101

Raciborski, R. 2008. "kountry: A Stata Utility for Merging Cross-Country Data from Multiple Sources." *Stata Journal*. Vol. 8, No. 3, pp. 390–400. https://doi.org/10.1177/1536867X0800800305

Russett, B.M., J.D. Singer, and M. Small. 1968. "National Political Units in the Twentieth Century: A Standardized List." *American Political Science Review*. Vol. 62, No. 3, pp. 932–951. https://doi.org/10.2307/1953441

Santos Silva, J.M.C. and S. Tenreyro. 2006. "The Log of Gravity." *Review of Economics and Statistics*. Vol. 88, No. 4, pp. 641–658. https://doi.org/10.1162/rest.88.4.641

Shakespeare, W. 2003. *Romeo and Juliet*. New York: Cambridge University Press.

[UCDP] Uppsala Conflict Data Program, International Peace Research Institute, Oslo. 2018. "UCDP/PRIO Armed Conflict Dataset Codebook". Version 18.1. http://ucdp.uu.se/downloads/#d3.

[UN Comtrade]. DESA/UNSD. 2018. *United Nations Comtrade Database*. https://comtrade.un.org/.

World Bank. 2017a. *World Development Indicators 2017*. Washington, D.C. https://data.worldbank.org/products/wdi.

World Bank. 2017b. *World Development Indicators. Annual Report*. Washington, D.C. Available at https://openknowledge.worldbank.org/handle/10986/26447

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

Appendix Tables A1, A2, and A3

Democracy datasets comparison

See Boese and Kamin (2018a) for a very detailed listing of all countries and their respective time series covered. Countries for which only the names/labels differ are listed in Table A1 (that is, countries of inconsistency type 3, reason i.) In the worksheet "Overview" (Boese and Kamin, 2018a), these countries are highlighted in grey.

Countries for which the underlying entity has no perfect match in the other dataset are listed in Table A2. A "perfect match" refers to a counterpart in terms of names and years (and potentially borders). This includes countries of inconsistency types 1 and 3. Countries representing the same or similar historical units are grouped.

Countries unmergable due to name and time inconsistencies are listed in Table A3. This includes countries of inconsistency type 3. Note: # obs=number of observations; N=total number of available observations in data; missing=number of missing years/observations for given country between its first and last year.

Table A1: Countries for which only the names/labels differ (democracy datasets)

V-Dem Version 8	*Polity IV, Version 2016*
Bosnia and Herzegovina	Bosnia
Burma/Myanmar	Myanmar (Burma)
Democratic Republic of Congo	Congo Kinshasa
German Democratic Republic	Germany East
North Korea	Korea North
Piedmont-Sardinia	Sardinia
Republic of Vietnam	Vietnam South
Republic of the Congo	Congo Brazzaville
Slovakia	Slovak Republic
South Korea	Korea South
South Yemen	Yemen South
United Arab Emirates	UAE
United States of America	United States
Würtemberg	Wuerttemburg

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

Table A2: Countries for which the underlying entity has no perfect match in the other dataset (democracy datasets)

V-Dem Version 8	*Polity IV, Version 2016*
Barbados	
	Yugoslavia
Brunswick	
Colombia	Colombia Gran Colombia
Czech Republic	Czech Republic Czechoslovakia
Democratic Republic of Vietnam	Vietnam North Vietnam
German Democratic Republic	Germany East
Germany	Germany Prussia Germany West
Guatemala	United Province of CA (Central America)
Hamburg	
Hanover	
Hesse-Darmstadt	
Hesse-Kassel	
Hong Kong	
Iceland	
Ivory Coast	Ivory Coast Cote D'Ivoire
Maldives	
Mecklenburg Schwerin	
Nassau	
Oldenburg	
	Orange Free State
Palestine/British Mandate	
Palestine/Gaza	
Palestine/West Bank	
Russia	USSR
Saxe-Weimar-Eisenach	
Serbia	Serbia Serbia and Montenegro
Seychelles	
Somaliland	
South Korea	Korea South Korea
South Sudan	South Sudan
Sudan	Sudan Sudan-North
São Tomé and Príncipe	
Timor-Leste	Timor Leste East Timor
Vanuatu	
Yemen	Yemen Yemen North
Zanzibar	

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

Table A3: Countries unmergeable due to name and time inconsistencies (democracy datasets)

V-Dem Version 8, 201 countries					*Polity IV, Version 2016, 195 countries*				
	First	*Last*	*# obs in data*			*First*	*Last*	*# obs in data*	
Country	*Year*	*Year*	*N*	*Missing*	*Country*	*Year*	*Year*	*N*	*Missing*
Bosnia and Herzegovina	1992	2017	26	0	Bosnia	1992	2016	25	0
					Yugoslavia	1921	2002	83	-1
Colombia	1789	2017	229	0	Colombia	1832	2016	185	0
					Gran Colombia	1821	1832	12	0
Czech Republic	1918	2017	100	0	Czech Republic	1993	2016	24	0
					Czechoslovakia	1918	1992	75	0
Democratic Republic of Vietnam	1945	2017	73	0	Vietnam North	1954	1976	23	0
					Vietnam	1976	2016	41	0
Germany	1789	2017	225	4	Germany	1868	2016	105	44
					Prussia	1800	1867	68	0
					Germany West	1945	1990	46	0
Ivory Coast	1900	2017	118	0	Ivory Coast	1960	2015	56	0
					Cote D'Ivoire	2016	2016	1	0
Russia	1789	2017	229	0	Russia	1800	2016	148	69
					USSR	1922	1991	70	0
Serbia	1804	2017	213	1	Serbia	1830	2016	102	85
					Serbia and Montenegro	2003	2006	4	0
South Korea	1789	2017	229	0	Korea South	1948	2016	69	0
					Korea	1800	1910	111	0
Sudan	1900	2017	118	0	Sudan	1956	2011	56	0
					Sudan-North	2011	2016	6	0
South Yemen	1900	1990	91	0	Yemen South	1967	1990	24	0
Yemen	1789	2017	162	67	Yemen	1990	2016	27	0
					Yemen North	1918	1990	73	0
Timor-Leste	1900	2017	118	0	Timor Leste	2016	2016	1	0
					East Timor	2002	2015	14	0

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

Appendix Tables A4, A5, and A6

Economic datasets comparison

Table A4 is a listing of unmergeable names/labels in the UN Comtrade and WDI datasets, due to inconsistency type 3, and shows a large share of countries with high export levels (Boese and Kamin, 2018b, contains the list sorted by total exports; worksheet “Unmergable Outliers Comtrade”. The spreadsheet also provides a list of country groups/regions which were not included in the comparison; worksheet “Disregarded Country Groups”).

Table A5 shows countries for which the underlying entity has no perfect match in the other dataset. A “perfect match” refers to a counterpart in terms of names and years (and potentially borders). This includes countries of inconsistency types 1 and 3. Countries representing the same or similar historical units are grouped.

Table A6 show countries unmergable due to name and time inconsistencies. This includes countries of inconsistency type 3 (N=total number of available observations in data).

Table A4: Countries for which the names/labels differ (economic datasets)

UN Comtrade exports	*WDI trade openness*
Bolivia (Plurinational State of)	Bolivia
Bosnia Herzegovina	Bosnia and Herzegovina
Cabo Verde	Cape Verde
Cayman Isds	Cayman Islands
Central African Rep.	Central African Republic
China, Hong Kong SAR	Hong Kong
China, Macao SAR	Macao SAR, China
Congo	Republic of the Congo
Czechia	Czech Republic
Côte d’Ivoire	Ivory Coast
Dem. Rep. of the Congo	Democratic Republic of Congo
Dominican Rep.	Dominican Republic
FS Micronesia	Micronesia, Fed. Sts.
Faeroe Isds	Faroe Islands
Gambia	The Gambia
Lao People's Dem. Rep.	Laos
Myanmar	Burma/Myanmar
Rep. of Korea	South Korea
Rep. of Moldova	Moldova
Russian Federation	Russia
Saint Kitts and Nevis	St. Kitts and Nevis
Saint Lucia	St. Lucia
Saint Vincent and the Grenadines	St. Vincent and the Grenadines
Sao Tome and Principe	São Tomé and Príncipe
Solomon Isds	Solomon Islands
TFYR of Macedonia	Macedonia
Turks and Caicos Isds	Turks and Caicos Islands
US Virgin Isds	Virgin Islands (U.S.)
USA	United States of America
United Rep. of Tanzania	Tanzania
Viet Nam	Vietnam
Yemen	Yemen, Rep.

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

Table A5: Countries for which the underlying entity has no perfect match in the other dataset (economic datasets)

UN Comtrade exports	*WDI trade openess*
	American Samoa
Belgium	Belgium
Belgium-Luxembourg	
	British Virgin Islands
	Channel Islands
Cook Isds	
	Curacao
Czechia Czechoslovakia	Czech Republic
East and West Pakistan	
	Equatorial Guinea
Ethiopia Frm Ethiopia	Ethiopia
Frm Tanganyika	
Fmr Yugoslavia	
French Guiana	
Germany Fmr Dem. Rep. of Germany Fmr Fed. Rep. of Germany	Germany
	Gibraltar
Guadeloupe	
	Guam
India India, excl. Sikkim	India
	Isle of Man
	Kosovo
	Liechtenstein
	Marshall Islands
Mayotte	
	Monaco
Montserrat	
	Nauru
Neth. Antilles Neth. Antilles and Aruba	
Niue	
	North Korea
	Northern Mariana Islands
Panama Fmr Panama, excl. Canal Zone	Panama
Pensinsula Malayia	
	Puerto Rico
Réunion	
Sabah	
Saint Kitts, Nevis and Anguilla	
Saint Pierre and Miquelon	
	San Marino
Serbia and Montenegro	
	Sint Maarten (Dutch part)
State of Palestine	
	West Bank and Gaza
	St. Martin (French part)
Sudan Fmr Sudan	Sudan
	South Sudan
USA USA (before 1981)	United States of America
	Uzbekistan
Viet Nam Frm Rep. of Vietnam	Vietnam
Yemen Frm Arab Rep. Of Yemen	Yemen, Rep.

Table A6: Countries unmergeable due to name and time inconsistencies (economic datasets)

UN Comtrade exports years available (coded and nonmissing)				*WDI tradeopenness years available (coded and nonmissing)*			
Country	*First Year*	*Last Year*	*N*	*Country*	*First Year*	*Last Year*	*N*
Belgium	1999	2017	19	Belgium	1960	2016	57
Belgium-Luxembourg	1962	1998	30				
Bosnia Herzegovina	2003	2017	15	Bosnia and Herzegovina	1994	2016	23
Czechia	1993	2017	24	Czech Republic	1990	2016	27
Czechoslovakia	1968	1987	20				
Pakistan	1972	2017	31	Pakistan	1967	2016	50
East and West Pakistan	1962	1971	10				
Ethiopia	1995	2016	21	Ethiopia	2011	2016	6
Fmr Ethiopia	1962	1987	21				
Fmr Yugoslavia	1962	1987	26				
Germany	1991	2017	27	Germany	1970	2016	47
Fmr Dem. Rep. of Germany	1985	1987	3				
Fmr Fed. Rep. of Germany	1962	1990	29				
India	1975	2017	43	India	1960	2016	57
India, excl. Sikkim	1962	1974	13				
Panama	1978	2016	32	Panama	1960	2016	57
Fmr Panama, excl.Canal Zone	1962	1977	16				
Serbia	2005	2017	13	Serbia	1995	2016	22
Serbia and Montenegro	1992	2004	9				
State of Palestine	2007	2016	10	West Bank and Gaza	1994	2016	23
Sudan	2012	2015	2	Sudan	1960	2016	57
Fmr Sudan	1963	2011	37				
				South Sudan	2008	2015	8
Viet Nam	2000	2016	17	Vietnam	1986	2016	31
Fmr Rep. of Vietnam	1963	1973	11				
Yemen	2004	2015	12	Yemen, Rep.	1990	2016	27
Fmr Arab Rep. of Yemen	1975	1981	6				

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

Appendix Table A7

Conflict dataset and codebook comparison

Table A7 is a comparison of country coding units in the UCDP/PRIO Armed Conflict dataset 18.1 to the coding units supplied in the code book. Countries with inconsistent labels are highlighted in **blue**; countries which only exist in the dataset but not in code book are highlighted in **red**.

Table A7: Comparison of country coding units in UCDP/PRIO Armed Conflict dataset 18.1 and the coding units supplied in the respective code book

Countries coded as state actors in side A or B of the UCDP/PRIO Armed Conflict dataset 18.1				*System membership table (Table 3) UCDP/PRIO Armed Conflict Dataset Codebook (pp.15–20)*		
Country	*First Year*	*Last Year*	*# obs*	*State Name*	*First Year*	*Last Year*
Afghanistan	1978	2017	47	Afghanistan	1946	2012
Albania	1946	1946	2	Albania	1946	2012
Algeria	1963	2017	30	Algeria	1962	2012
Angola	1975	2017	36	Angola	1975	2012
Argentina	1955	1982	8	Argentina	1946	2012
				Armenia	1991	2012
Australia	2003	2003	2	Australia	1946	2012
				Austria	1946	2012
Azerbaijan	1991	2017	15	Azerbaijan	1991	2012
				Bahamas	1973	2012
				Bahrain	1971	2012
Bangladesh	1975	2017	21	Bangladesh	1971	2012
				Barbados	1966	2012
				Belarus (Byelorussia)	1991	2012
				Belgium	1946	2012
				Belize	1981	2012
				Benin	1960	2012
				Bhutan	1949	2012
Bolivia	1946	1967	3	Bolivia	1946	2012
Bosnia-Herzegovina	1992	1995	9	Bosnia-Herzegovina	1992	2012
				Botswana	1966	2012
				Brazil	1946	2012
				Brunei	1984	2012
				Bulgaria	1946	2012
Burkina Faso	**1985**	**1987**	**3**	**Burkina Faso (Upper Volta)**	**1960**	**2012**
Burundi	1965	2015	19	Burundi	1962	2012
Cambodia (Kampuchea)	1967	2011	42	Cambodia (Kampuchea)	1953	2012
Cameroon	1960	2017	10	Cameroon	1960	2012
				Canada	1946	2012
				Cape Verde	1975	2012
Central African Republic	2001	2013	8	Central African Republic	1960	2012
Chad	1966	2017	43	Chad	1960	2012
Chile	1973	1973	1	Chile	1946	2012
China	1946	2008	45	China	1946	2012
Colombia	1964	2016	53	Colombia	1946	2012
Comoros	1989	1997	2	Comoros	1975	2012
Congo	1993	2016	6	Congo	1960	2012
DR Congo (Zaire)	**1960**	**2017**	**30**	**Congo, Democratic Republic of (Zaire)**	**1960**	**2012**
Costa Rica	1948	1948	1	Costa Rica	1946	2012
Ivory Coast	**2002**	**2011**	**4**	**Cote D'Ivoire**	**1960**	**2012**
Croatia	1992	1995	3	Croatia	1991	2012
Cuba	1953	1961	5	Cuba	1946	2012
Cyprus	1974	1974	2	Cyprus	1960	2012
				Czech Republic	1993	2012
				Czechoslovakia	1946	1992

Table A7 (continued)

Countries coded as state actors in side A or B of the UCDP/PRIO Armed Conflict dataset 18.1				*System membership table (Table 3) UCDP/PRIO Armed Conflict Dataset Codebook (pp.15–20)*		
	First	*Last*			*First*	*Last*
Country	*Year*	*Year*	*# obs*	*State Name*	*Year*	*Year*
				Denmark	1946	2012
Djibouti	1991	2008	7	Djibouti	1977	2012
Dominican Republic	1965	1965	1	Dominican Republic	1946	2012
				East Timor	2002	2012
Ecuador	1995	1995	2	Ecuador	1946	2012
Egypt	1948	2017	29	Egypt	1946	2012
El Salvador	1969	1991	16	El Salvador	1946	2012
				Equatorial Guinea	1968	2012
Eritrea	1997	2016	12	Eritrea	1993	2012
				Estonia	1991	2012
Ethiopia	1960	2016	131	Ethiopia	1946	2012
				Fiji	1970	2012
				Finland	1946	2012
France	1946	1962	55	France	1946	2012
Gabon	1964	1964	1	Gabon	1960	2012
Gambia	1981	1981	1	Gambia	1965	2012
Georgia	1991	2008	8	Georgia	1991	2012
				German Democratic Republic	1949	1990
				German Federal Republic	1949	2012
Ghana	1966	1983	3	Ghana	1957	2012
Greece	1946	1949	4	Greece	1946	2012
Grenada	**1983**	**1983**	**2**			
Guatemala	1949	1995	34	Guatemala	1946	2012
Guinea	2000	2001	2	Guinea	1958	2012
Guinea-Bissau	1998	1999	2	Guinea-Bissau	1974	2012
				Guyana	1966	2012
Haiti	1989	2004	3	Haiti	1946	2012
Honduras	1957	1969	3	Honduras	1946	2012
Hungary	1956	1956	2	Hungary	1946	2012
Hyderabad	**1947**	**1948**	**4**			
				Iceland	1946	2012
India	1948	2017	220	India	1947	2012
Indonesia	1950	2005	52	Indonesia	1946	2012
Iran	**1946**	**2017**	**62**	**Iran (Persia)**	**1946**	**2012**
Iraq	1948	2017	78	Iraq	1946	2012
				Ireland	1946	2012
Israel	1948	2014	86	Israel	1948	2012
				Italy/Sardinia	1946	2012
				Jamaica	1962	2012
				Japan	1946	2012
Jordan	1948	2016	6	Jordan	1946	2012
				Kazakhstan	1991	2012
Kenya	1982	2017	4	Kenya	1963	2012
				Kosovo	2008	2012
Kuwait	1990	1991	2	Kuwait	1961	2012
				Kyrgyz Republic	1991	2012

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

Table A7 (continued)

Countries coded as state actors in side A or B of the UCDP/PRIO Armed Conflict dataset 18.1				*System membership table (Table 3) UCDP/PRIO Armed Conflict Dataset Codebook (pp.15–20)*		
	First	*Last*			*First*	*Last*
Country	*Year*	*Year*	*# obs*	*State Name*	*Year*	*Year*
Laos	1959	1990	22	Laos	1954	2012
				Latvia	1991	2012
Lebanon	1948	2017	17	Lebanon	1946	2012
Lesotho	1998	1998	1	Lesotho	1966	2012
Liberia	1980	2003	7	Liberia	1946	2012
Libya	1987	2017	8	Libya	1951	2012
				Lithuania	1991	2012
				Luxembourg	1946	2012
Macedonia, FYR	**2001**	**2001**	**1**	**Macedonia (FRY)**	**1991**	**2012**
Madagascar	**1971**	**1971**	**1**	**Madagascar (Malagasy)**	**1960**	**2012**
				Malawi	1964	2012
Malaysia	1958	2013	15	Malaysia	1957	2012
				Maldives	1965	2012
Mali	1985	2017	18	Mali	1960	2012
				Malta	1964	2012
Mauritania	1975	2011	6	Mauritania	1960	2012
				Mauritius	1968	2012
Mexico	1994	1996	2	Mexico	1946	2012
Moldova	1992	1992	1	Moldova	1991	2012
				Mongolia	1946	2012
				Montenegro	2006	2012
Morocco	1963	1989	17	Morocco	1956	2012
Mozambique	1977	2016	18	Mozambique	1975	2012
Myanmar (Burma)	1948	2017	275	Myanmar (Burma)	1948	2012
				Namibia	1990	2012
Nepal	1960	2006	14	Nepal	1946	2012
Netherlands	1946	1962	5	Netherlands	1946	2012
				New Zealand	1946	2012
Nicaragua	1957	1990	13	Nicaragua	1946	2012
Niger	1991	2017	10	Niger	1960	2012
Nigeria	1966	2017	20	Nigeria	1960	2012
North Korea	1949	1953	10	North Korea	1948	2012
				Norway	1946	2012
Oman	1957	1975	8	Oman	1946	2012
Pakistan	1948	2017	55	Pakistan	1947	2012
Panama	1989	1989	3	Panama	1946	2012
Papua New Guinea	1990	1996	6	Papua New Guinea	1975	2012
Paraguay	1947	1989	3	Paraguay	1946	2012
Peru	1965	2010	24	Peru	1946	2012
Philippines	1946	2017	104	Philippines	1946	2012
				Poland	1946	2012
Portugal	1961	1974	36	Portugal	1946	2012
				Qatar	1971	2012
Rumania	1989	1989	1	Rumania	1946	2012
Russia (Soviet Union)	1946	2017	44	Russia (Soviet Union)	1946	2012
Rwanda	1990	2016	17	Rwanda	1962	2012

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

Table A7 (continued)

Countries coded as state actors in side A or B of the UCDP/PRIO Armed Conflict dataset 18.1				*System membership table (Table 3) UCDP/PRIO Armed Conflict Dataset Codebook (pp.15–20)*		
	First	*Last*			*First*	*Last*
Country	*Year*	*Year*	*# obs*	*State Name*	*Year*	*Year*
Saudi Arabia	1979	1979	1	Saudi Arabia	1946	2012
Senegal	1990	2011	10	Senegal	1960	2012
Serbia (Yugoslavia)	**1991**	**1999**	**5**	**Serbia**	**2006**	**2012**
				Yugoslavia (Serbia)	**1946**	**2006**
Sierra Leone	1991	2001	11	Sierra Leone	1961	2012
				Singapore	1965	2012
				Slovakia	1993	2012
				Slovenia	1992	2012
				Solomon Islands	1978	2012
Somalia	1964	2017	32	Somalia	1960	2012
South Africa	1966	1988	30	South Africa	1946	2012
South Korea	1949	1953	5	South Korea	1948	2012
South Sudan	2011	2017	9	South Sudan	2011	2012
Spain	1957	1991	11	Spain	1946	2012
Sri Lanka	1971	2009	27	Sri Lanka	1948	2012
Sudan	1963	2017	49	Sudan	1956	2012
Suriname	**1987**	**1987**	**1**	**Surinam**	**1975**	**2012**
				Swaziland	1968	2012
				Sweden	1946	2012
				Switzerland	1946	2012
Syria	1948	2017	27	Syria	1946	2012
Taiwan	1949	1958	4	Taiwan	1949	2012
Tajikistan	1992	2011	10	Tajikistan	1991	2012
Tanzania	**1978**	**1978**	**2**	**Tanzania/Tanganyika**	**1961**	**2012**
Thailand	1946	2017	32	Thailand	1946	2012
				Tibet	1946	1950
Togo	1986	1986	1	Togo	1960	2012
Trinidad and Tobago	1990	1990	1	Trinidad and Tobago	1962	2012
Tunisia	1961	2016	3	Tunisia	1956	2012
Turkey	**1974**	**2017**	**41**	**Turkey/Ottoman Empire**	**1946**	**2012**
				Turkmenistan	1991	2012
Uganda	1971	2017	41	Uganda	1962	2012
Ukraine	2014	2017	7	Ukraine	1991	2012
				United Arab Emirates	1971	2012
United Kingdom	1946	2003	56	United Kingdom	1946	2012
United States of America	1950	2017	23	United States of America	1946	2012
Uruguay	1972	1972	1	Uruguay	1946	2012
Uzbekistan	1999	2004	3	Uzbekistan	1991	2012
Venezuela	1962	1992	3	Venezuela	1946	2012
Vietnam (North Vietnam)	**1965**	**1988**	**24**	**Vietnam, Democratic Republic of**	**1954**	**2012**
South Vietnam	**1955**	**1975**	**32**	**Vietnam, Republic of**	**1954**	**1975**
Yemen (North Yemen)	**1948**	**2017**	**27**	**Yemen (Arab Republic of Yemen)**	**1946**	**2012**
South Yemen	**1972**	**1986**	**5**	**Yemen, People's Republic of**	**1967**	**1990**
				Zambia	1964	2012
				Zanzibar	1963	1964
Zimbabwe (Rhodesia)	1967	1979	9	Zimbabwe (Rhodesia)	1965	2012

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk

References

Boese, V.A., and K. Kamin. 2018a. "Democracy Datasets.xslx." https://www.dropbox.com/s/3bm674tjk9iqha4/Democracy%20Datasets.xlsx?dl=0.

Boese, V.A., and K. Kamin. 2018b. "Economic Datasets.xslx." https://www.dropbox.com/s/qtaol2upv0uns30/Economic%20Datasets.xlsx?dl=0.

© EPS Publishing, 2019. All rights reserved. For permissions, email: ManagingEditor@EPSJournal.org.uk