2023 Update to Longitudinal Cross-National Distance Data

This companion describes several cycles of updates to the cross-national and longitudinal data for the nine distance dimensions (including administrative, cultural, demographic, economic, financial, global connectedness, knowledge and geographic distance) that are described in our paper 2010 Berry, Guillen and Zhou (2010), "An Institutional Approach to Cross National Distance". The most recent update was done in 2023 and captures data up to 2021 for most of our distance dimensions.

We offer two versions for each of our nine dimensions, including what we call **Mahalanobis Pooled** and **Mahalanobis Year**:

- To create the **Mahalanobis Pooled**, we used a pooled covariance matrix to create all distance dimensions – meaning that data from all years were used to calculate the covariance matrix.
- To create the **Mahalanobis Year**, we only used data from the current year to calculate the covariance matrices for each distance dimension.

When considering which of these versions to use, you should consider the other data you have. If you have panel data for your other variables, our Mahalanobis Pooled data takes into account scale and correlation across variables over time. If you have cross sectional data, then you may want to use the distance dimensions that use separate matrices for each year.

Starting with the 2020 update, you can now download the data in two formats: excel and Stata. As they have been in the past, the excel files are separate for each of the nine distance dimensions, with each spreadsheet in the excel file containing one year of country-dyad distance calculations. We started including a Stata version in 2020 and we have included two Stata files in the 2023 update that contain the nine distance dimensions for each country-dyad for every year within one dataset. For years of coverage for each dimension, please see Table 3 below. The file "distances pooled" contains the Mahalanobis distance for each dimension using a pooled covariance matrix and "distancesbyyear" contains the Mahalanobis distance for each dimension using only the current year. (Within these Stata files "." reflects missing data from raw source data and "0" is recorded for each pair where the country is the same in the "Country_a" and "Country_b" dyad.)

Raw Data and Updating:

We use the WDI country codes as country identifiers throughout our data files. As always, our distance dimensions are only as good as the underlying raw data that were used to create them. Where there is missing data across countries and time in the source data, we also suffer from the same missing data points. In addition, there is one distance dimension where we interpolated between two data points to create yearly data – our cultural distance dimension. Appendix 1 (below) describes the items we used in the World Values Surveys to approximate Hofstede's (1980) four cultural dimensions.

We downloaded all data from the sources listed in Tables 2 and 3 in Berry et al. (2010), reproduced and updated below. We calculated our distance dimensions incorporating all of the variables for each individual dimension listed in Tables 2 and 3. In Appendix 2 (below), we explain how the Mahalanobis calculations we did to create these distance measures differ from the more common Euclidean distance calculations.

Some changes that we had to make to be able to do the updating over the last few updating cycles include the following:

Notes from the 2023 updates:

1). For our political distance measure, updates to the underlying polcon database source (one of the dimensions we use to create our political distance dimension) was changed to the Varieties of Democracy (V-DEM) to define the efficacy of the government branches and entities with the 2021 release and our 2023 update uses the recently released version for all years. In previous updates, we started the political distance data in 1960, but did not have data for FDI as a percentage of GDP until 1970. We now just report political distance data from 1970 onwards, but we are happy to share the previous data upon request.

2). For our finance distance measure, two of the three variables we use from the World Bank data were not updated to 2021 (as of June, 2023), including "market capitalization of listed companies" and the "number of domestic listed companies." Because of this, we downloaded the data from the original source (the World Federation of Exchanges - WFE). This data was checked and verified against the data published for earlier years on the World Bank website to confirm it is the same data for years prior to 2021.

3). The economic distance measure in the 2023 update uses the WDI data that was posted in October, 2022. At this time, the GDP per capita (income) measure used a base year of 2015 (instead of the 2000 that was used in our original economic distance data and 2005 that was used in the 2014 update and 2010 in the 2017 and 2020 updates). In addition, the more recent versions of the WDI raw data uses the inflation rate instead of the GDP deflator.

4).For our knowledge distance measure, although the "Calendar Year Patent Statistics" related to geographic origin were not yet updated for 2021 on the USPTO website when we downloaded all data for the 2023 update, we contacted the USPTO and they provided us with the updated numbers.

5.) It is worth noting that the knowledge distance dimension uses comprehensive data on scientific articles and patents. There are occasions when countries have zero scientific articles and patents attributed to them. If two countries both have zero scientific articles and patents, then the distance between them is calculated as zero. Researchers should take this into account when using this data.

6). As with previous cycles, there are occasions when there are data changes to the historical data in the data sources we use. For instance, in the 2017 update, there was data available on stock market capitalization and listed companies for Cote D'Ivoire from 1992 to 2016. In the 2020 World Development Indicators, the World Bank has revised this, and not included any data on these variables for Cote D'Ivoire. When this happens, we assume the World Bank has a valid reason for this omission, and so we do not include the old data in our updated dataset. This means that occasionally, countries that were included in a particular measure in previous versions of the dataset are omitted in newer versions. In the 2023 update, we found that there was more missing data that had previously been reported in earlier years. This is most noticeable in the Political and

Global Connectedness distances. For example, in the 2020 update, for the year 2010, we had coverage for 149 countries for political distance, while we only have coverage for 137 countries in the 2023 update. While we only calculate distance with data that is reported by the World Bank at the time of updating, we are happy to share the previous data for any countries that have been omitted in later rounds upon request (with the caveat that there is usually a reason that this data has been revised and deleted by the World Bank).

Notes from prior updating cycles:

6). From the 2017 update onwards, there have also been changes to the raw data on which our Financial Distance measure is based. The WDI has changed its source for stock market data, meaning that where we previously had coverage of 99 countries in our last year of the spreadsheets (in 2012), we now only have 53 countries in our last year (2021) in the 2023 updated data. (We are happy to send anyone the old data for researchers looking to use a wider set of countries until 2012 – please just e-mail us.) The following is the explanation from WDI for the decrease in country coverage: *Stock market data were previously sourced from Standard & Poor's until they discontinued their "Global Stock Markets Factbook" and database in April 2013. Time series have been replaced in December 2015 with data from the World Federation of Exchanges and may differ from the previous S&P definitions and methodology*

7). From 2017 update: Since 2014, the World Values Survey panel dataset no longer includes data from the European Values Survey (though it did in 2010). Our updated Cultural Distance data includes data items from what is currently available in the World Values Survey and the European Values Surveys for all years. Some of these changes to the underlying raw data mean that the updated data may not be similar to older versions of the data. Please keep this in mind if you are trying to download new years of data only. The data available from the Stata file released by the World Values Survey does not contain all countries available, and so we downloaded the data manually using the "online data analysis" feature on the website of the World Values Survey for the seventh wave. This has increased our coverage for some countries (in 2010 we previously had coverage for 65 countries, and now we have coverage for 92). The cultural distance data is interpolated in between survey years. The last survey year for each country is the last year the country appears in the data. For example, the United States did their survey for the 2017-2019 WVS wave in 2017, and therefore data can only be interpolated until 2017. The last year of WVS surveys in the last wave is 2022, but this only includes 15 countries, and therefore the culture data is only included up to 2018.

8). From 2017 update: Since 2014, we have removed the common language dimension of the administrative distance. Our source data for language is the CIA Factbook and this source is inconsistent in how it reports languages across countries (meaning, sometimes it lists languages and a percentage break-down across those languages, while other times it lists languages with no percentage break-down). Given these inconsistencies, we decided to drop the language component of our Administrative Distance dimension. (Our posted data is highly correlated with what we used in our JIBS publication.)

9). When we updated the global connectedness data in 2020, there were only 90 countries that had updated data for internet users for the last year. This has been rectified in the more recent data releases, and so we now have more coverage. However, generally the global connectedness data release lags other dimensions in data release, and so this dimension is only updated to 2020.

10.) From 2017 update: For the knowledge distance, in 2016, there was a revision to the WDI measure of Scientific Articles, which results in a series break in the data, which only now covers 2000 – 2013. This raw data is only collected by WDI intermittently and interpolated. So, we went to the source of this new data, Thompson Reuters Incites database, which publishes this data yearly. We downloaded this for the longer time period covering 1980 to 2019. The specific categories we collected data for are Physics, Biology & Biochemistry, Chemistry, Mathematics, Clinical Medicine, Engineering, Space Science, Immunology, Microbiology, Materials Science, Molecular Biology & Genetics, Geosciences, Environment/Ecology, consistent with the data collection description in the WDI.

11.) In the 2017 update, we made substantial changes to the political distance dimension. In previous versions, we captured regional trade agreements in numerous variables as inputs into the distance calculation. As regional trade agreements increased, this caused problems in the covariance matrix. We have replaced the regional trade agreement variable with two other variables that represent the political distance dimension: the FDI to GDP ratio, and the number of bilateral trade agreements a country has. In the very early years of the data, the old political distance and new political distance dimension are somewhat correlated, but in later years, when there were more regional trade agreements, they are less so.

Please remember that in using the data posted on this website you agree to cite our paper as the source in all publications, whether printed, digital or otherwise, and in any genre (including papers, reviews, notes, powerpoint presentations, etc.). Please note that we are not responsible for any errors or omissions and we always appreciate hearing any issues you find in the data.

Suggested citation:

Berry, H., Guillen, M and Zhou, N. 2010. An Institutional Approach to Cross-National Distance, *Journal of International Business Studies* 41(9): 1460-1480.

Table 2: Indicator Component Variables Used in the Calculation of Distance Dimensions (for 2020)

Dimension:	Component Variables:
1. Economic Distance	
Income	GDP per capita, 2010 USD
Inflation	GDP deflator (% GDP)
Exports	Exports of goods and services (% GDP)
Imports	Imports of goods and services (% GDP)
2. Financial Distance	
Private Credit	Domestic credit to private sector (% GDP)
Stock Market Cap	Market capitalization of listed companies (% GDP)
Listed Companies	Number of listed companies (per one million population)
3. Political Distance	
Policy Making Uncertainty	Political stability measured by considering independent Institutional actors with veto power
Size of the state	General government final consumption expenditure (% of GDP)
WTO member	Membership in WTO (GATT before 1993)
Bilateral Trade Agreements	Number of Bilateral Trade Agreements
FDI to GDP	Foreign direct investment, net inflows (% of GDP)
4. Administrative Distance	
Colonizer-colonized link	Whether dyad shares a colonial tie
Common religion	% population that share the same religion in the dyad
Legal system	Whether dyad shares the same legal system
5. Cultural Distance	
Power distance	WVS question on obedience and respect for authority
Uncertainty avoidance	WVS questions on trusting people
Individualism	WVS questions on independence and the role of government in providing for its citizens
Masculinity	WVS questions on the importance of family and work
6. Demographic Distance	
Life expectancy	Life expectancy at birth, total (years)
Birth rate	Birth rate, crude (per 1,000 people)
Population under 14	Population ages 0-14 (% of total)
Population under 65	Population ages 65 and above (% of total)
7. Knowledge Distance	
Patents	Number of patents per one million population
Scientific Articles	Number of scientific articles per one million population
8. Global Connectedness Distance	
International Tourism Expend	International tourism, expenditures (% GDP)
International Tourism Receipts	International tourism, receipts (% GDP)
Internet use	Internet users per 1,000 people
9. Geographic Distance	
Great circle distance	Great circle distance between two countries according to the coordinates of the geographic center of the countries

Table 3: Distance Dimensions, Sources, Year Availability, and Country Coverage

Dimension:	**Source**	**Years Available**	**# of Countries (in last year)**
1. Economic Distance			
Income	WDI	1961-2021	191
Inflation	WDI	1961-2021	191
Exports	WDI	1961-2021	164
Imports	WDI	1961-2021	164
2. Financial Distance			
Private credit	WDI	1960-2021	146
Stock market Cap	WDI/WFE	1988-2021	68
Listed companies	WDI/WFE	1988-2021	67
3. Political Distance**			
Policymaking uncertainty	POLCONV	1960-2021	146
Size of the state	WDI	1960-2021	167
World trade agreements	WTO	1960-2021	177
Bilateral trade agreements	WTO	1960-2021	170
FDI to GDP	WDI	1970-2021	184
4. Administrative Distance			
Colonizer-colonized Link	CIA Factbook	constant	198
Common religion	CIA Factbook	constant	198
Legal system	La Porta et al.	constant	198
5. Cultural Distance			
Power distance	WVS	1990-2018	67***
Uncertainty avoidance	WVS	1990-2018	67***
Individualism	WVS	1990-2018	67***
Masculinity	WVS	1990-2018	67***
6. Demographic Distance			
Life expectancy	WDI	1960-2021	203
Birth rate	WDI	1960-2021	207
Population under 14	WDI	1960-2021	211
Population under 65	WDI	1960-2021	211
7. Knowledge Distance			
Patents	USPTO	1975-2021	221
Scientific articles	Thompson Reuters InCites	1980-2021	221
8. Global Connectedness Distance			
International tourism Expend	WDI	1995-2020	132
International tourism Receipts	WDI	1995-2020	121
Internet users	WDI	1995-2020	181
9. Geographic Distance			
Great circle distance	CIA Factbook	constant	196
Common border	CIA Factbook	constant	226

*** Maximum number of countries included in the Cultural Distance Dimension is 91 (in 2005)

Appendix 1: World Values Survey Questions Used in the Calculation of Cultural Distance

In order to replicate Hofstede’s (1980) cultural scores with time-varying measures, we used the World Values Survey (Inglehart, 2004). We downloaded data from five waves of the WVS conducted between 1980 and 2014 for as many as 69 countries. However, not all of our questions were available in the early waves of the surveys and we have focused on those surveys with sufficient data on all questions in what we have posted. Effectively, this means that we have data from 1990-2014, interpolating the years in between individual surveys. We used mean response scores by country as the input data for all calculations.

To measure Hofstede’s power distance, we computed the percentage of respondents who chose “obedience” in response to question a042: “Here is a list of qualities that children can be encouraged to learn at home. Which, if any, do you consider to be especially important? Please choose up to five.” The other categories were leadership, self-control, thrift saving money and things, determination perseverance, religious faith, unselfishness, and loyalty. We also took into account the percentage of people who responded to question e018 that “it would be a good thing”: “I'm going to read out a list of various changes in our way of life that might take place in the near future. Please tell me for each one, if it were to happen, whether you think it would be a good thing, a bad thing, or don't you mind? —Greater respect for authority.”

To measure uncertainty avoidance we computed the percentage of people answering “very careful” to question a165: “Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?”

To measure individualism we computed the percentage of people who chose “independence” in response to question a029: “Here is a list of qualities that children can be encouraged to learn at home. Which, if any, do you consider to be especially important? Please choose up to five.” The other categories were good manners, politeness and neatness, hard work, honesty, feeling of responsibility, patience, and imagination. We also took into account the response to question e037: “Now I'd like you to tell me your views on various issues. How would you place your views on this scale? 1 means you agree completely with the statement on the left; 10 means you agree completely with the statement on the right; and if your views fall somewhere in between, you can choose any number in between: —The government should take more responsibility to ensure that everyone is provided for. —People should take more responsibility to provide for themselves.”

Finally, to measure masculinity we computed the mean response on a scale from 1 to 5 to the question: “For each of the following, indicate how important it is in your life: —a001 family, and_—a005 work.

Appendix 2: Calculation of Euclidean and Mahalanobis Distances

Euclidean Distance

Suppose that for a specific country in any given year we observe x_i for $i = 1...p$, where x_i is a characteristic of the specified dimension (e.g. for the economic dimension x_1=imports, x_2=GDP per capita etc). Let $\bar{x}_i$ be the arithmetic mean of characteristic i across all countries in any given year. Let σ_{x_i} be the sample standard deviation of characteristic i across all countries in any given year. We first standardized each dimension:

$$z_i = \frac{(x_i - \bar{x}_i)}{\sigma_{x_i}} \qquad \text{for each country in the given year.}$$

Then the Euclidean distance between two countries A and B is calculated as

$$d(A,B) = \sqrt{\sum_{i=0}^{p} (z_i(A) - z_i(B))^2}$$

where $z_i(A)$ and $z_i(B)$ are the values of the standardized characteristic i corresponding to countries A and B respectively.

Mahalanobis Distance

Suppose that for two countries in any given year, we observe two vectors $\mathbf{a} = (a_1, a_2, \dots, a_p)$ and $\mathbf{b} = (b_1, b_2, \dots, b_p)$ of p different characteristics. Similarly, suppose there is an *n*-by-*p* matrix **M** with p columns representing characteristics, and n rows containing each country in each year (so the number of rows would be the summation over all years of the number of countries in each year). We define **C**, a covariance matrix for **M**, as a *p*-by-*p* matrix with element $\mathbf{C}_{\mathrm{ij}}$ equal to the sample covariance of columns *i* and *j* in the matrix **M**. Finally, let **I** be the *p*-by-*p* identity matrix. Then the squared Mahalanobis distance between two countries is calculated as:

$$d(a,b)^2 = (\mathbf{a} - \mathbf{b})\mathbf{C}^{-\mathbf{1}}(\mathbf{a} - \mathbf{b})^{\mathbf{T}}$$

We can alternatively rewrite the Euclidean distance above as:

$$d(a,b) = \sqrt{(\mathbf{a} - \mathbf{b})\mathbf{I}(\mathbf{a} - \mathbf{b})^{\mathbf{T}}}$$

References

Berry, H., Guillen, M and Zhou, N. 2010. An Institutional Approach to Cross-National Distance, *Journal of International Business Studies* 41(9): 1460-1480.

Hofstede, G. 1980. *Culture's Consequences: International Differences in Work-related Values*. Beverly Hills, CA: Sage.

Inglehart, R. 2004. *Human beliefs and values*. Madrid: Siglo XXI.