GeoDist: the CEPII's distances and geographical database

Thierry Mayer and Soledad Zignago

CEPII

3. May 2006

GeoDist: the CEPII's distances and geographical database

Thierry Mayer *
Soledad Zignago †

November 6, 2010‡

Abstract. This document details GeoDist's, the CEPII's distances and geographical database. We have calculated and made available different measures of bilateral distances available for most countries across the world (225 countries in the current version of the database). For most of them, different calculations of "*intra*-national distances" are also available. The GeoDist webpage provides two distinct files: a country-specific one (geo_cepii) and a dyadic one (dist_cepii) including the set of different distance and common dummy variables used in gravity equations to identify particular links between countries such as colonial past, common languages, contiguity. We try to improve upon the existing similar datasets in terms of geographical coverage, measurement and number of variables provided.

Keywords: Distances, Database, Gravity, International Trade.

JEL classification: F10, C80.

*Sciences-Po, CEPII and CEPR (thierry.mayer@sciences-po.fr).
†Banque de France (soledad.zignago@gmail.com).
‡We thank Guillaume Gaulier to his participation at earliest stages of this work.

1 Introduction

We have build and made available two datasets providing useful data for empirical economic research including geographical elements and variables. A common use of these files is the estimation by trade economists of gravity equations describing bilateral patterns of trade flows. Other datasets have been proposed and provide geographical and distance data, notably those developed by Jon Haveman, Vernon Henderson and Andrew Rose. We try to improve upon the existing sets of variables in terms of geographical coverage, measurement and the number of variables provided. Covariates such as bilateral distance, contiguity, or colonial historical links have also been used in other fields than international trade: for the study of bilateral flows of foreign direct investment for instance, but also by researchers interested in explaining migration patterns, international flows of tourists, of telephone traffic, etc. Even outside economics, several researchers in different social sciences use these types of variables. Political scientists, for instance, use distance and contiguity (among other determinants) to explain why some pairs of countries have a higher probability than others of going to war.

Our first dataset (geo_cepii), incorporates country-specific geographical variables for 225 countries in the world, including the geographical coordinates of their capital cities, the languages spoken in the country under different definitions, a variable indicating whether the country is landlocked, etc. The second dataset (dist_cepii) is dyadic, in the sense that it includes variables valid for pairs of countries. Distance is the most common example of such a variable, and the file includes different measures of bilateral distances (in kilometers) available for most countries across the world (again 225 countries in the current version of the database).

2 The country-specific files: `geo_cepii.xls` and `geo_cepii.dta`

The `geo_cepii` files provide data on countries and their main city or agglomeration. There are first 3 identification codes of the country according to the ISO classification, the country's area in square kilometers, used to calculate in particular its internal distance. Variables indicating whether the country is landlocked and which continent it is part of are also included.

There are several language variables that can be used to create different indexes of language proximity or dummy variables for common language in dyadic applications like gravity equations. The sources for all language information are the web site www.ethnologue.org and the CIA World Factbook. For each country, we report the official languages (up to three), as well as the languages spoken by at least 20% of the population and the languages spoken by between 9 and 20% of the population (up to four languages in each of those cases). Colonial linkage variables are also often used by economists to proxy for similarities in cultural, political or legal institutions. Our dataset provides several variables (based on the CIA World Factbook, and the Correlates of War Project run by political scientists, available at cow2.la.psu.edu) that identify for each country, up to 4 long-term and

up to 3 short-term colonizers in the whole history of the country.

The distance calculation described in the next section requires information on geographical coordinates of at least one city in each of the country. The simplest measure of geodesic distance considers only the main city of the country, reported here with the English and French names, latitude and longitude. In most cases, the main city is the capital of the country. However, for 13 out of the 225 countries, we considered that the capital was not populated enough to represent the 'economic center' of the country. For these countries, we propose the distances data calculated for both the capital city and the economic center. Consequently, there are 238 (225+13) observations in the geo_cepii.xls file. Also included is a variable providing the number of cities for each country (available in the www.world-gazetteer.com dataset) used to calculate our weighted distances described in the next section.

2.1 Country-level variables

- `iso2`, `iso3`, `cnum`: ISO codes in two and three characters, and in three numbers respectively.[1]
- `country`, `pays`: Name of country in English and French respectively.[2]
- `area`: Country's area in km^2.
- `dis_int`: Internal distance of country i, $d_{ii} = .67\sqrt{\text{area}/\pi}$ (an often used measure of average distance between producers and consumers in a country, see Head and Mayer, 2002 for more on this topic).
- `landlocked`: Dummy variable set equal to 1 for landlocked countries.
- `continent`: Continent to which the country is belonging
- `langoff_i`: Official or national languages and languages spoken by at least 20% of the population of the country (and spoken in another country of the world[3]) following the same logic than the "open-circuit languages" in Mélitz (2002).
- `lang20_i`: Languages (mother tongue, lingua francas or second languages) spoken by at least 20% of the population of the country.

[1]The numeric codes are the United Nations Standard Countries/Area codes used in trade data. Consequently, the code for Belgium is not 056 but 058, the Belgium-Luxembourg code.

[2]Countries and capital names in French follow "Pays et capitales du Monde. Pays indépendants au 1.01.2001" of the Institut Géographique National. English names follow "The World Factbook" of the CIA.

[3]Because their similarity, we consider the Papiamento as Spanish.

- `lang9_i`: Languages (mother tongue, lingua francas or second languages) spoken by between 9% and 20% of the population of the country. [4].
- `colonizeri`: Colonizers of the country for a relatively long period of time and with a substantial participation in the governance of the colonized country.
- `short_colonizeri`: Colonizers of the country for a relatively short period of time or with only low involvement in the governance of the colonized country [5]

2.2 Cities variables used in the computation of distances

The following (country-specific also) variables describe the city used to calculate simple distances, i.e. the ones where only one city (or "agglomeration", which usually corresponds to an enlarged definition of the city: "Essen" is for instance the biggest agglomeration of Germany in our sample) by country is considered. In most cases, the main city is the capital of the country. However, for 13 out of the 225 countries, we considered that the capital was not populated enough to represent the "economic center" of the country. For these countries, we propose the distances data calculated for both the capital city and the economic center. Consequently, there are 238 (225+13) observations in the `geo_cepii.xls` file.[6]

- `city_en`, `city_fr`: Names of capitals or main cities of the country in English and French.
- `lat`, `lon`: Latitude and longitude of the city.[7]
- `cap`: Variable equals to 1 if the city is the capital of the country, to 0 if the city is the most populated city (`maincity` equals to 1) but not the capital, and to 2 in the cases of two capitals, if the city is the most populated but the "second" capital or the previous capital[8].

[4]The first source of the language variables is the web site http://www.ethnologue.com/which allows us to calculate the share of the population of each country speaking any languages but mainly as a mother tongue. Hence, to have precise idea about the lingua francas and second languages spoken in each country, we used two other valuable sources : the CIA world factbook and Jacques Leclerc web page "L'aménagement linguistique dans le monde."

[5]The main sources to create this variables were TheFreeDictionary.com, the Correlates of War Project and the CIA World Factbook.

[6]Those cases where the economic center differs from the capital are: South Africa (The Cap), Germany (Essen), Australia (Sydney), Benin (Cotonou), Bolivia (La Paz), Brazil (São Paulo), Canada (Toronto), Côte d'Ivoire (Abidjan), United States (New York), Kazakstan (Almaty), Nigeria (Lagos), Tanzania (Dar Es Salam) and Turkey (Istambul).

[7]The source of these geographic coordinates is generally the PcGlobe software. In 14% of the cases different web sources were additionally consulted (http://www.world-gazetteer.com most notably).

[8]In general we have considered only one capital by country. Some countries have however a second official capital city more important in size like South Africa with Cape Town, Benin with Cotonou or Bolivia with La Paz. Similarly, recent capitals are generally smaller than old capitals

- `maincity`: Variable coded as 1 when the city is the most populated of the country and as 2 otherwise[9].
- `citynum`: Number of cities for each country used to calculate our weighted distances described in the next section.

3 The bilateral files: `dist_cepii.xls` and `dist_cepii.dta`

The `dist_cepii` files provide the bilateral data: the different distance measures and dummy variables indicating whether the two countries are contiguous, share a common language or a colonial relationship.

There are two kinds of distance measures: simple distances, for which only one city is necessary to calculate international distances; and weighted distances, for which we need data on principal cities in each country. The simple distances are calculated following the great circle formula, which uses latitudes and longitudes of the most important city (in terms of population) or of its official capital. These two variables incorporate internal distances based on areas provided in the geo_cepii.xls file. The two weighted distance measures use city-level data to assess the geographic distribution of population inside each nation. The idea is to calculate distance between two countries based on bilateral distances between the largest cities of those two countries, those inter-city distances being weighted by the share of the city in the overall country's population. This procedure can be used in a totally consistent way for both internal and international distances. We use latitudes, longitudes and populations data of main agglomerations of all countries available in www.world-gazetteer.com. The distance formula used is a generalized mean of city-to-city bilateral distances developed by Head and Mayer (2002), which takes the arithmetic mean and the harmonic means as special cases. We provide the two variables corresponding to those cases.

3.1 Simple distances: `dist` and `distcap`

Geodesic distances are calculated following the great circle formula, which uses latitudes and longitudes of the most important cities/agglomerations (in terms of population) for the `dist` variable and the geographic coordinates of the capital cities for the `distcap` variable. These two variables incorporate internal distances based on areas and also provided in the `geo_cepii.xls` file (see description above).

as for instance Dodoma, which is planned as the new Tanzania capital, or Abuja in Nigeria, or Astana in Kazakhstan.

[9]Keeping where `maincity` equals to 1, the sample has 225 observations with one city per country country.

3.2 Weighted distances: `distw` and `distwces`

We compute two distance measures using city-level data to assess the geographic distribution of population (in 2004) inside each nation. The basic idea is to calculate distance between two countries based on bilateral distances between the biggest cities of those two countries, those inter-city distances being weighted by the share of the city in the overall country's population.[10] This procedure can be used in a totally consistent way for both internal and international distances. We use data of the World Gazetteer web site, which provides current population figures and geographic coordinates for cities, towns and places of all countries[11]. The general formula developed by Head and Mayer (2002) and used for calculating distances between country i and j is

$$d_{ij} = \left(\sum_{k \in i} (\text{pop}_k / \text{pop}_i) \sum_{\ell \in j} (\text{pop}_\ell / \text{pop}_j) d_{k\ell}^{\theta} \right)^{1/\theta}, \quad (1)$$

where pop_k designates the population of agglomeration k belonging to country i. The parameter θ measures the sensitivity of trade flows to bilateral distance $d_{k\ell}$. For the `distw` calculation, θ is set equal to 1. The `distwces` calculation sets θ equal to -1, which corresponds to the usual coefficient estimated from gravity models of bilateral trade flows.[12]

3.3 Other gravity variables

Finally the `dist_cepii.xls` file provides also dummy variables indicating wheter the two countries are contiguous (`contig`), share a common language, have had a common colonizer after 1945 (`comcol`), have ever had a colonial link (`colony`), have had a colonial relationship after 1945 (`col45`), are currently in a colonial relationship (`curcol`)[13] or were/are the same country (`smctry`)[14].

[10] See Head and Mayer (2002) for more details about international and intra-national distance calculations.

[11] More precisely, we use the popdata.zip file available at http://www.world-gazetteer.com and take the 25 more populated cities by country.

[12] For the 10 countries that have only one city counted in the dataset, the weighted distances, distw and distwces, have been replaced by the simple distance, dist, since, otherwise, it would equal 32.186 kilometer (20 miles) which is the convention for the inner-city distance (the distance of a city to itself.)

[13] These colonial variables are widely influenced by the data of Andrew Rose.

[14] This variable complements the `comcol` variable setting to one if countries were or are the same state or the same administrative entity for a long period (25-50 years in the twentieth century, 75 year in the ninetieth and 100 years before). This definition covers countries have been belong to the same empire (Austro-Hungarian, Persian, Turkish), countries have been divided (Czechoslovakia, Yugoslavia) and countries have been belong to the same administrative colonial area. For instance, Spanish colonies are distinguished following their administrative divisions in the colonial period (viceroyalties). According to this definition, Argentina, Bolivia, Paraguay and Uruguay were thus a single country. Similarly, the Philippines were subordinated to the New Spain viceroyalty and thus `smctry` equals to one with Mexico. Sources for this variable came from http://www.worldstatesmen.org/.

There are two common languages dummies, the first one based on the fact that two countries share a common official language, and the other one set to one if a language is spoken by at least 9% of the population in both countries. Trying to give a precise definition of a colonial relationship is obviously a difficult task. Colonization is here a fairly general term that we use to describe a relationship between two countries, independently of their level of development, in which one has governed the other over a long period of time and contributed to the current state of its institutions.

4 References

- **K. Head and T. Mayer (2002)**, ''Illusory Border Effects: Distance Mismeasurement Inflates Estimates of Home Bias in Trade", *CEPII Working Paper* 2002-01.
- **Mélitz, J. (2002)**, ''Language and Foreign Trade", *CEPR Discussion Paper* 3590.