Package 'cepiigeodist'

May 8, 2026

Title CEPII's GeoDist Datasets

Version 0.1

Description Provides data on countries and their main city or agglomeration and the different distance measures and dummy variables indicating whether two countries are contiguous, share a common language or a colonial relationship. The reference article for these datasets is Mayer and Zignago (2011) <http://www.cepii.fr/CEPII/en/publications/wp/abstract.asp?NoDoc=3877>.

License CC0

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Depends R (>= 2.10)

URL https://pacha.dev/cepiigeodist/

BugReports https://github.com/pachamaltese/cepiigeodist/issues/

Suggests gravity

NeedsCompilation no

Author Mauricio Vargas [aut, cre] (ORCID: <https://orcid.org/0000-0003-1017-7574>), Centre d'études prospectives et d'informations internationales (CEPII) [dtc]

Maintainer Mauricio Vargas <mvargas@dcc.uchile.cl>

Repository CRAN

Date/Publication 2020-09-18 12:20:07 UTC

Contents

dist_cepii *Data on pairs of countries including distance measures and dummy variables indicating common attributes*

Description

Provides 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' dataset. 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. 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.

Format

A data frame with 50176 observations on the following 14 variables.

`iso_o` Country of origin as ISO codes in three characters.

`iso_d` Country of destination as ISO codes in three characters.

`contig` Variable coded as 1 when the two countries are next to each other and 0 otherwise.

`comlang_off` Variable coded as 1 when the two countries share the same official language.

`comlang_ethno` Variable coded as 1 when the two countries have at least 9% of their population speaking the same language.

`colony` Variable coded as 1 when the country in 'iso_o' was ever a colony of the country in 'iso_d'.

`comcol` Variable coded as 1 when the two country share the same colonizer after 1945.

`curcol` Variable coded as 1 when the country in 'iso_o' is a colony of the country in 'iso_d'.

`col45` Variable coded as 1 when the country in 'iso_o' is a colony of the country in 'iso_d' after 1945.

`smctry` Variable coded as 1 when the two countries were or are the same country.

`dist` Simple distance (most populated cities, km)

`distcap` Simple distance between capitals (capitals, km)

`distw` Weighted distance (pop-wt, km) with theta=1 (theta measures the sensitivity of trade flows to bilateral distance dkl)

`distwces` Weighted distance (pop-wt, km) theta=-1.

Source

http://www.cepii.fr/CEPII/en/bdd_modele/download.asp?id=6

References

Mayer, T. & Zignago, S. (2011) Notes on CEPII's distances measures: the GeoDist Database CEPII Working Paper 2011-25

Head, K. & Mayer, T. (2002) Illusory Border Effects: Distance Mismeasurement In-flates Estimates of Home Bias in Trade CEPII Working Paper 2002-01

Examples

```
# filter countries that share borders
dist_cepii[dist_cepii$contig == 1, ]
```

---

geo_cepii *Data on countries and their main city or agglomeration*

Description

There are firstly three 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.

Format

A data frame with 238 observations on the following 34 variables.

`iso2` ISO codes in two characters.

`iso3` ISO codes in three characters.

`cnum` ISO codes in three numbers.

`country` Name of country in English.

`pays` Name of country in French.

`area` Country's area in km2.

`dis_int` Internal distance of country i, dii=.67*sqrt(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.

`city_en` Names of capitals or main cities of the country in English.

`city_fr` Names of capitals or main cities of the country in French.

`lat` Latitude of the city.

`lon` Longitude of the city.

`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.

`maincity` Variable coded as 1 when the city is the most populated of the country and as 2 otherwise.

`citynum` Number of cities for each country used to calculate the weighted distances described in Mayer and Zignago, 2011.

`langoff_1` 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) following the same logic than the "open-circuit languages" in Mélitz (2002).

`langoff_2` Same as langoff_1.

`langoff_3` Same as langoff_1.

`lang20_1` Languages (mother tongue, lingua francas or second languages) spoken by at least 20% of the population of the country.

`lang20_2` Same as lang20_1.

`lang20_3` Same as lang20_1.

`lang20_4` Same as lang20_1.

`lang9_1` Languages (mother tongue, lingua francas or second languages) spoken by between 9% amd 20% of the population of the country.

`lang9_2` Same as lang9_1.

`lang9_3` Same as lang9_1.

`lang9_4` Same as lang9_1.

`colonizer1` Colonizers of the country for a relatively long period of time and with asubstantial participation in the governance of the colonized country.

`colonizer2` Same as colonizer1.

`colonizer3` Same as colonizer1.

`colonizer4` Same as colonizer1.

`short_colonizer1` Colonizers of the country for a relatively short period of time orwith only low involvement in the governance of the colonized country.

`short_colonizer2` Same as short_colonizer1.

`short_colonizer3` Same as short_colonizer1.

Source

http://www.cepii.fr/CEPII/en/bdd_modele/download.asp?id=6

References

Mayer, T. & Zignago, S. (2011) Notes on CEPII's distances measures: the GeoDist Database CEPII Working Paper 2011-25

Head, K. & Mayer, T. (2002) Illusory Border Effects: Distance Mismeasurement In-flates Estimates of Home Bias in Trade CEPII Working Paper 2002-01

Examples

```
# filter to avoid multiple records for the same country
geo_cepii[geo_cepii$cap == 1 & geo_cepii$maincity == 1, ]
```

Index