CEPII
CENTRE
D'ÉTUDES PROSPECTIVES
ET D'INFORMATIONS
INTERNATIONALES

N° 2011 – 25

December

NOTES ON CEPII'S DISTANCES MEASURES: THE GEODIST DATABASE

Thierry Mayer & Soledad Zignago

NON-TECHNICAL SUMMARY

GeoDist makes available the exhaustive set of gravity variables developed in Mayer and Zignago (2005) to analyze market access difficulties in global and regional trade flows. GeoDist provides useful data online (http://www.cepii.fr/anglaisgraph/bdd/distances.htm) 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. 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. Other datasets have been proposed in the literature 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.

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, and their colonial links. 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.

The main contribution of GeoDist is to compute internal (or intra-national) and international bilateral distances in a totally consistent way. How define internal distances of countries? How make those constructed internal distances consistent with 'traditional' international distances

CEPII
CENTRE
D'ÉTUDES PROSPECTIVES
ET D'INFORMATIONS
INTERNATIONALES

N° 2011 – 25

Novembre

NOTES SUR LA BASE DE DONNÉES DE DISTANCES DU CEPII (*GEODIST*)

Thierry Mayer & Soledad Zignago

RÉSUMÉ NON TECHNIQUE

GeoDist fournit l'ensemble des données développées par Mayer and Zignago (2005) pour mesurer les difficultés d'accès aux marchés mondiaux. GeoDist, ou base de données de distances du CEPII, propose en ligne (http://www.cepii.fr/anglaisgraph/bdd/distances.htm) des données géographiques utiles à la recherche empirique, en particulier pour l'estimation des équations de gravité dans le domaine du commerce international. Par rapport aux séries élaborées par Jon Haveman, Vernon Henderson et Andrew Rose, nous avons étendu la couverture géographique, affiné les mesures et développé le nombre des variables. Au-delà de l'analyse du commerce, la distance entre deux pays, leur contigüité, les liens historiques sont autant de variables utilisées dans d'autres champs de recherche, comme ceux des investissements directs, des flux migratoires ou touristiques, du trafic téléphonique, etc. Les chercheurs en sciences sociales recourent également à des variables ; en sciences politiques par exemple, distance et contigüité sont prises en compte dans le calcul des probabilités de conflit.

Une première série de données rassemble les variables caractérisant chacun des 225 pays. Le fichier geo_cepii (geo_cepii.xls ou geo_cepii.dta) contient les variables géographiques des pays et de leur principale ville ou agglomération : l'identification du pays (codes ISO) ; la superficie (en km2), utilisée en particulier pour le calcul des distances internes, les coordonnées géographiques de la (ou des) capitale(s), l'éventuel enclavement, le continent, etc. Cette série de données comporte aussi plusieurs variables de langue permettant de déterminer les proximités linguistiques. Pour chaque pays, on peut avoir jusqu'à trois langues officielles ; la base distingue les langues parlées par plus de 20 % de la population et celles parlées par un tranche de 9 à 20 % de la population. Les relations coloniales passées constituent une autre information souvent utilisée par les économistes pour approximer les similitudes culturelles politiques ou institutionnelles.

Une seconde série de données est dyadique, au sens ou les variables sont calculées par couple de pays : la distance (km) entre deux pays est l'exemple type de ce genre de variables bilatérales. Le fichier dist_cepii (dist_cepii.xls ou dist_cepii.dta) contient les variables bilatérales : les différentes mesures de distances et les variables muettes indiquant la contigüité, la communauté de langue, ou de liens coloniaux. On mesure deux types de distances : simple, pour laquelle on recourt à une seule ville ; pondérée, qui considère plusieurs villes par pays afin de prendre en compte la répartition géographique de l'activité économique.

Ces distances pondérées sont la principale contribution de GeoDist. Pour pouvoir comparer les flux internationaux aux flux de commerce “intra-nationaux”, ce que nous faisions dans Mayer et Zignago (2005) en estimant des effets frontière sur l’ensemble des pays du monde, il fallait construire une bonne approximation des distances moyennes parcourues par les biens à l’intérieur de chaque pays. En effet, une sous-estimation des distances relatives biaise mécaniquement à la hausse l’effet frontière estimé. Pour éviter cela, nous tenons compte de la répartition géographique de l’activité économique à l’intérieur des nations en utilisant les populations et coordonnées des principales villes de chaque pays dans le calcul de la matrice des distances. L’idée, inspirée de Head and Mayer (2002) est de calculer les distances entre deux pays comme une moyenne des distances entre leurs principales villes pondérée par leur population. Cette méthodologie permet de calculer des distances internes aux pays de manière cohérente avec le calcul des distances internationales.

Classification J.E.L. **:** F10, C80

Mots clés : Commerce international, Bases de données, Coûts au commerce, Distances, Géographie, Effets Frontière, Gravité.

calculations? The latter question is in fact crucial for obtaining a correct estimate of trade impediments. Any overestimate of the internal / external distance ratio will yield to a mechanic upward bias in the border effect estimate. We have computed these distances using city-level data to assess the geographic distribution of population (in 2004) inside each nation. The basic idea, inspired by Head and Mayer (2002), 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.

*J.E.L. Classification***:** F10, F12; F13, F14, F15, C80.

Keywords : Distances, International Trade, Databases, Gravity, Trade Costs, Border Effects.