Replication Package for "Road Endpoints and City Sizes"

Bruno Barsanetti

This is a replication package of the empirical analysis in "Road Endpoints and City Sizes". The package allows the reproduction of all the empirical results reported in the paper (both in the main body and in the Online Appendix). Before using this package, please read Online Appendix A.5 for a detailed description of the data sources and the data preparation. The do-files are also carefully documented. For any remaining queries, please contact me at bruno.barsanetti@fgv.br

This package contains two Stata do-files. The final versions of these do-files were prepared in Stata 18; they run in previous versions of Stata, at least up to Stata 14. Before running any of the do-files, please unzip the package and set the directory to the folder in which the do-file is.

prepare_data.do: this do-file imports and cleans the raw data in "Data" and prepares the intermediary datasets that are saved in "Working Data". It may take over one hour to run.

analysis.do: this do-file imports the intermediary datasets in "Working Data" and performs all the empirical analyses reported in the paper.

This package contains three folders. See below for a list with the content of each folder:

Working_Data: this folder contains the intermediate files that are prepared in prepare_data.do. If this do-file runs, it overwrites the files in this folder. The files in this folder are sufficient for running the empirical analysis in analysis.do. The folder contains the following files (in alphabetical order):

amc_historic.dta: dataset containing a crosswalk between modern municipality codes and minimum comparable areas (AMCs) for different historic censuses.

amc1970_data.dta: dataset with information at the level of 1970-2010 AMCs.

amc1991_data.dta: dataset with information at the level of 1991-2010 AMCs.

census1872.dta: dataset with information from the 1872 census.

cum_migration.dta: dataset with historic migration flows into Brazil.

dist_data.dta: dataset with information at the district level.

frontier_vars.dta: dataset with results of the 2018 presidential election.

mun_data.dta: dataset at the municipality level (according to 2010 borders).

mun_data_neighboring.dta: dataset with information on the non-rail municipalities that neighbor the sample municipalities.

pastpopdensity.dta: dataset with information on historical population densities.

sectors2010.dta: dataset with information on the number of sectors of economic activity in each municipality, according to the 2010 census.

Maps: this folder contains the QGIS projects that reproduce the 3 maps in the paper (Figures 2, A.3 and A.4). Each project is named for the respective figure. They all use the municipality shapefile "Mun Data/mun_samplemap.shp", which contains selected variables from the main sample (i.e. mun_data.dta), and the shapefiles in "Auxiliary Shapefiles": luz.shp, which contains the point where *Estação da Luz* is; rail_sample.shp, a shapefile cut from the 2017 *Declaração de Rede* which contains the railroad segments in the sample; and sp_borders.shp, a polygon of the state of São Paulo.

Data: this folder contains the raw datasets used in prepare_data.do. Most files are semi-colon delimited text files. The folder contains the following files (in alphabetical order):

area_district.csv: file with the area of each district, calculated based on an equal area Albers projection of the district shapefiles for the 2010 census from IBGE.

censo10.dta: file with the micro data of the 2010 census, unpackaged and labeled with the Stata package from PUC-Rio Datazoom.

censo1872_data.csv: file with the municipality-level tabulations from the 1872 census, digitized by CEDEPLAR.

coord_district.csv: file with the geographic coordinates of each district, calculated from the district shapefiles for the 2010 census from IBGE.

crosswalk_tse.dta: crosswalk between IBGE and TSE municipality codes.

foundation_year.csv: list of the municipality incorporation dates, manually collected from IBGE Cidades.

gdp.csv: estimates of municipality GDP and value added by sector in 2010 from IBGE (*Produto Interno Bruto dos Municípios*, table 5938).

historical_agggdp.csv: estimates of historical Brazilian GDP and population from Araújo, Carpena and Cunha, "Brazilian business cycles and growth from 1850 to 2000", *Estudos Econômicos* 38 (2008), 557-581.

IPEADATA_Municipios_X_AMCs.csv: table of the minimum comparable areas by Reis, Pimentel, Alvarenga and Santos, "Áreas mínimas comparáveis para os períodos intercensitários de 1872 a 2000", *Rio de Janeiro, Ipea/Dimac* 40 (2008).

migration.csv: yearly series of migration from IBGE "Brasil: 500 anos de povoamento". Rio de Janeiro, 2000.

mun_area.csv: municipality areas according to IBGE.

mun_coordinates.csv: coordinates of the municipality centroids, based on municipality shapefiles for 2010 from IBGE.

mun_neighboring.csv: list of non-railroad municipalities that border at least one of the 204 sample municipalities, manually identified on the map.

mun_rivers.csv: list of municipalities that intersect at least one of the main rivers on the map of the Paraná basin from the *Agência Nacional das Águas*.

mun_shapefile.csv: attribute table from the municipality shapefile, including the average of the rasters of the geographic fundamentals (calculated using the zonal statistics function in QGIS).

pop_district.csv: district populations (total and urban) from the 2010 census, Table 202, "População residente, por sexo e situação do domicílio".

pop_mun.csv: municipality populations (total and urban) from the 1970, 1980, 1991, 2000, and 2010 census, Table 202, "População residente, por sexo e situação do domicílio".

pop_1890.csv: municipality population from the 1890 census, manually digitized from the "Quadros Gerais" from the *Synopse do Recenseamento de 1890*.

pop_1900.csv: municipality population from the 1900 census, manually digitized from the "Quadros Gerais" from the *Synopse do Recenseamento de 1900*.

pop_1920.csv: municipality population from the 1920 census, manually digitized from table "População dos municípios de cada um dos Estados, segundo o sexo e o estado civil", from volume IV of the census publication.

pop_1940.csv: municipality population from the 1940 census, manually digitized from table 48, "População total, de fato", from the census publication.

pop_1950.csv: municipality population from the 1950 census manually digitized from table 39, "População presente, por sexo e nacionalidade, segundo as zonas fisiográficas e os municípios", from the census publication.

pop_1960.csv: municipality population from the 1960 census, manually digitized from table 2, "Grupos de idade, por sexo, segundo as zonas fisiográficas e os municípios", from the census publication.

rail_stations.csv: railroad station dataset containing stations coordinates, names, opening years and original railroad company. This dataset is based on the attribute table of the shapefile on railroad stations released by ANTT jointly with the 2017 *Declaração de Rede*, and manually complemented with information from the website *Estações Ferroviárias do Brasil*.

regions.csv: data with the *Regiões Geográficas Imediatas* and *Regiões Geográficas Intermediárias* of each municipality, according to the 2017 division prepared by IBGE

station_to_district.csv: list of the IBGE district to which each railroad station in the sample belongs to, obtained from a join in QGIS between the railroad station and district shapefiles, using the intersection as the geometric predicate.

votacao_candidato_munzona_2018_BR.csv: voting data by candidate and geography from the 2018 general elections in Brazil, prepared by the *Tribunal Superior Eleitoral* (TSE).

Besides the files above, which are read by prepare_data.do to prepare the working datasets, the folder also contains the subfolder "Data/Spatial". This subfolder contains shapefiles and rasters with the raw data; some of the .csv files in "Data" are attribute tables of these shapefiles. This sub-folder contains the following files (I do not list auxiliary files to the shapefiles):

Rasters:

acrisols.tif: raster of acrisols, rasterized from the Embrapa soil shapefile.

coffee_int.tif: raster of coffee potential yield, from GAEZ v.3.

dem_alos.tif: altitude from ALOS Global Surface Model.

latosols.tif: raster of latosols, rasterized from the Embrapa soil shapefile.

malaria_index.tif: raster of the number of days a year that support malaria transmission, averaged between *P. falciparum* and *P. vivax*, from the Malaria Atlas.

maize_int.tif: raster of maize potential yield, from GAEZ v.3.

sugar_int.tif: raster of sugarcane potential yield, from GAEZ v.3.

terraroxa.tif: raster of *terra roxa*, rasterized from the Embrapa soil shapefile.

tri_alos.tif: terrain ruggedness index derived from dem_alos.tif.

Shapefiles:

dbo_MalhaFerroviaria_Estacoes_spatial_point.shp: raw points shapefile with the location and information on the railroad stations from the 2017 *Declaração de Rede*.

districts.shp: shapefile on districts of the states of São Paulo, Minas Gerais, and Paraná.

districts.shp: shapefile on districts of the states of São Paulo, Minas Gerais, and Paraná. The attribute table of this shapefile is in "Data" as coord_district.csv.

districts_albers.shp: shapefile on districts of the states of São Paulo, Minas Gerais, and Paraná, reprojected to projection ESRI: 102033 – South America Albers Equal Area Conic. This projection allows a calculation of district areas. The attribute table of this shapefile is in "Data" as area_district.csv.

rivers.shp: shapefile with the main rivers of the Paraná river basin, as those rivers whose names are highlighted in the map of the basin by ANA. The shapefiles is selected from the *Base Hidrográfica Ottocodificada*, v. 1.3.

solos_br5m_2011_lat_long_wgs84.shp: shapefile on soil types from Embrapa.

station_to_district.shp: shapefile that resulted from a spatial join between stations.shp as a base layer, districts.shp as the join layer, and "intersects" as the geometric predicate. The attribute table of this shapefile is in "Data" as station_to_district.csv.

stations.shp: point shapefile on the railroad stations in the sample. This shapefile was selected from dbo_MalhaFerroviaria_Estacoes_spatial_point.shp, and additional attributes were manually inputed with information from *Estações Ferroviárias do Brasil.*