README

The World Bank Enterprise Surveys (WBES) are available at: https://login.enterprisesurveys.org/content/sites/financeandprivatesector/en/signup.html. The nightlight data is available at: https://doi.org/10.6084/m9.figshare.9828827.v2

Data cleaning and generating variables

Step 1. Number of cities in each country to be included in the sample: folder "Structural break test"

Based on the Zipf's law and the law of the Primate City, we include different numbers of cities in different categories of countries in the sample. Details are in "structural break test.pdf". Data are uploaded as "data_eviews - simp.dta".

Step 2. Identify cities and their boundaries: folder "identifying cities"

(1) Download maps on the boundaries of administrative subdivisions from the Database of Global Administrative Areas (https://gadm.org/download_country_v3.html). Save into one folder in your computer. In our code, we name this data folder as "city_map".

(2) Combine maps from all countries: Run code "merge_all_city.py". Get output "All.shp". Save its attribute table as "All.csv".

(3) Assign city information to each administrative subdivision. That is, identify each city and its boundary based on urban and well-developed suburban administrative subdivisions: Run code "city_definition_map.do". Get output "all_city_define.csv".

Step 3. Matching firms in WBES with city names: folder "identifying cities"

(1) Get GPS coordinates data in WBES (available to researchers, but a research proposal is required for approval). Create two shape files to plot the location of firms: GPS1.shp, GPS2.shp. (Note: we created two files because excel files could not handle so many observations in one file).

(2) Match GPS coordinates with the map of administrative subdivisions: Run code "match_city_GPS_output.py". Get output "matched1.xls" and "matched2.xls". To simplify the data, only keep idstd, join_count, country name, and names of administrative subdivisions (maximal 5 levels of subdivisions). Get output "matched1_to_stata.csv" and "matched2_to_stata.csv".

(3) Assign city information to each firm: Run code "city_definition_WBES.do". Get output "firm_city.dta".

Note: "city_definition_WBES.do" and "city_definition_map.do" are both necessary. The former is at the firm level and the latter is at the administrative subdivision level. The latter will be used to calculate

average nightlight luminosity, and then matched to firms in WBES using the city information in the former.

Step 4. Calculate average nightlight luminosity and make sure it can be matched to firm-level WBES: folder “Nightlight”

(1) Remove gas flare. Calculate average nightlight luminosity in each city in each year. Input data: “All.shp”, "All_flare.shp" (available in the folder), “all_city_define.csv”. Run code “nl_process.py”.

(2) Match the nightlight information with firms in WBES: Run code “combine_with_nightlight.do”. Output: “firm_all_cityname.dta”.

Step 5. Define war zones to be excluded from the sample: folder “Excluding war zone”

(1) Data can be found at the Uppsala Conflict Data Programme (UCDP): https://ucdp.uu.se/downloads/index.html#ged_global

Choose the dataset “UCDP/PRIO Armed Conflict Dataset version 20.1” and download “ucdp-prio-acd-201.dta” (Note: when we downloaded the data, it was 20.1. Now it is updated to version 21.1)

(2) Select the conflicts with >1000 deaths per year and match them with the city information: Run code “ucdp-prio-data.do” and “match_GPS.py”.

Step 6. Match everything above to get the data for regression: folder “Results”

(1) Using idstd, match the GPS coordinates data to the full set of the WBES data. This way a firm’s GPS coordinates are linked to all other information about it in WBES. Output: “2010-2020-match-gps.dta”

(2) Match all information together: Run code “match_all_data”. Output: “matched_WBES_regression.dta".

(3) Download data on GDP, region dummies and city population. Merge them to “matched_WBES_regression.dta":

https://databank.worldbank.org/source/world-development-indicators/Type/TABLE/preview/on

https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

Data on city population size is manually obtained from the most recent official statistics data for each country, and only counts the population in the city proper or urban areas. The data is uploaded as “city population data.dta” in the folder.

Step 7. Generate variables for regression: folder “Results”

Run code “Generating variables” (Note: the code also excludes developed countries and cities with <20 firms).

Regression tables and graphs: folder “Results”

1. For maps: Figure 1 and Figure 4

(1) Download a map of the world from: https://gadm.org/download_country_v3.html. Save the attribute table to “country_name.csv”.

(2) Run code “Code for maps - fig1 and 4.do”. Tables generated for figure 1 and 4 are in the folder “Results – data for figures”.

2. For regression: Figure 2, Figure 3 and all regression tables

Run code “Code for regression Tables and Graphs”. Input data is “matched_WBES_regression.dta" except:

(1) Figure 2: input data is "innovation intention" (panel a and b) or "innovation intensity" (panel c and d). Use this site to produce 3D graphs: https://imclient.herokuapp.com/chartsbuilder/?charttype=3d

(2) Figure 3: input data is “innovation intention large city or not.dta” (panel a) and “innovation intensity large city or not.dta” (panel b)