Replication data and scripts for “Measuring Economic Growth from Outer Space” by J. Vernon Henderson, Adam Storeygard and David N. Weil

The data described below are for replicating the results in "Measuring Economic Growth from Outer Space", forthcoming in the *American Economic Review*. Please cite the paper when using any of these data. The replication can be done starting from 3 different stages in the analysis. Working backwards from the final products, these are referred to as final_tables, full_tabular, and spatial. Final_tables and full_tabular can be carried out with Stata alone. Final_tables reproduces all tables in the paper. Full_tabular is the farthest the user can go back in the process without using GIS software, but it is somewhat less documented than final_tables. Spatial starts essentially from raw data, but additionally requires GIS software, and is also less documented. More details are below. Note that all scripts (.do, .aml, .bat and .pyw) in all sections below contain a line that effectively sets the working directory to “F:\adam\replication” or “F:/adam/replication” These must be changed to the working directory into which the user decompresses all the files listed below.

Final tables (hsw_final_tables_replication.zip)

Enclosed in hsw_final_tables_replication.zip are the following 10 files:

lightspaper_replication.do - this Stata do file replicates all the tables in the paper, except for section 5, using the datafiles global_total_dn_uncal.dta, global_total_dn_uncal_longdiff9206.dta, isocvout.dbf, ginioutu.dbf, and samptab.txt.

global_total_dn_uncal.dta - this Stata datafile is the primary panel dataset used in analysis. The country-year is the unit of analysis.

global_total_dn_uncal_longdiff9206.dta - this Stata datafile is the primary dataset used in the long differences analysis, 1992-2006. The country is the unit of analysis.

isocvout.dbf - this dbase format tabular datafile contains a column for each satellite-year and, for each country, a row for each number of days of coverage (i.e. the number of nights of valid data). Each number in the table then reports the number of satellite-year-pixels corresponding to the satellite-year column and the country-number of days row. Lightspaper_replication.do parses this table and reports an overall mean and standard deviation of days.

ginioutu.dbf - this dbase format tabular datafile contains a column for each satellite-year and, for each country, a row for each digital number (DN). Each number in the table then reports the number of satellite-year-pixels corresponding to the satellite-year column and the country-DN row. Lightspaper_replication.do parses this table and reports selected DN bins for selected countries in Table 1.

samptab.txt - this text format tabular datafile has a row for each approximately 1 square kilometer pixel. Each row contains a radiance-calibrated light value for a limited set of nights in the winter of 1996-97, an uncalibrated digital number (DN) for the most closely corresponding satellite year, F-12, 1997, and a longitude (x) and latitude (y) value. It is restricted to pixels with a non-zero value in the radiance-calibrated dataset. Lightspaper_replication.do creates the online Appendix table using these data.

africa_coastmalariaprimate.do - this Stata do file reports the results found in section 5, using the datafiles ctycs41.dbf, ctypr41.dbf, and ctyml41.dbf, corresponding to coast-hinterland, primate city, and malaria regions, respectively.

ctycs41.dbf - this dbase format tabular datafile contains a column for each satellite-year and, for each country in the African sample, a row for the inland region and a row for the coastal region. Some countries do not contain both. Each number in the table then reports the total digital number (DN) corresponding to the satellite-year column and the country-region row. Africa_coastmalariaprimate.do parses this table and reports aggregate statistics.

ctypr41.dbf - this dbase format tabular datafile contains a column for each satellite-year and, for each country in the African sample, a row for the primate city and a row for all other areas. Each number in the table then reports the total digital number (DN) corresponding to the satellite-year column and the country-region row. Africa_coastmalariaprimate.do parses this table and reports aggregate statistics.

ctyml41.dbf - this dbase format tabular datafile contains a column for each satellite-year and, for each country in the African sample, a row for each continental quartile of the Kiszewski et al (2004) index of the stability of malaria transmission. Some countries do not contain all four quartiles. Each number in the table then reports the total digital number (DN) corresponding to the satellite-year column and the country-region row. Africa_coastmalariaprimate.do parses this table and reports aggregate statistics.

Full tabular (hsw_full_tabular_replication.zip)

Enclosed in hsw_full_tabular_replication.zip are the following files and ginioutu.dbf, described above, which are used to recreate the main analysis files described above (global_total_dn_uncal.dta and global_total_dn_uncal_longdiff9206.dta):

v4ginicalc_uncal.do - starting from ginioutu.dbf, described above, this Stata do file calculates the Gini of lights of each country-year in the data. Its output, ginistata_uncal.dta is combined with the rest of the data in v4lights_stataprep_uncal.do

v4lights_stataprep_uncal.do - this stata do file creates the analysis datasets global_total_dn_uncal.dta and global_total_dn_uncal_longdiff9206.dta, described above, from 8 inputs: wb_dq.xls, imf_dds.xls, dhselect.xls, wdi_limited.dta, ctryoutu.dbf, ctryout2.dbf, ginioutu.dbf (described above) and ginistata_uncal.dta (described above).

wb_dq.xls - this Microsoft Excel file contains all the information from Appendix A of World Bank (2002) in tabular format.

imf_dds.xls - this Microsoft Excel file contains the membership of the IMF's General Data Dissemination Standard (GDDS) and Special Data Dissemination Standard (SDDS), as downloaded and entered from http://dsbb.imf.org/Pages/GDDS/ImportantDates.aspx and http://dsbb.imf.org/Pages/SDDS/DateOfSubscription.aspx on 20 April 2010.

dhselect.xls - This Microsoft Excel file contains data on household electricity access, downloaded and entered from the STATcompiler database for Macro International's Demographic and Health Surveys at http://www.measuredhs.com on 10 October 2010.

wdi_limited.dta – This Stata datafile contains selected indicators downloaded from the World Bank's World Development Indicators (WDI) database on 26 January 2010.

ctryoutu.dbf - this dbase format tabular datafile contains a row for every country and nearly 200 statistics related to population, land area, and (mostly) lights digital number (DN) calculated using GIS software.

ctryout2.dbf - this dbase format tabular datafile contains a row for every country and columns corresponding to several statistics related to missing lights data calculated using GIS software.

Spatial (multiple files)

Running the entire analysis, from the raw data, is a much more involved process. In addition to Stata, ArcGIS Desktop and ArcInfo Workstation GIS software (at the ArcInfo license level) and the open source compression utility 7-Zip are required. Python is also required, though this is typically installed with ArcGIS. Users unfamiliar with ArcGIS, and specifically with the Arc Macro Language (AML) are not encouraged to start from this point. This is not intended as a GIS teaching dataset. *We are providing these data and scripts as a service for users familiar with the relevant software, with no warranty or offer of consulting services or advice in how to use them.* The scripts used have only been tested on the authors' system, which runs Windows 7 Professional, ArcGIS version 9.3 service pack 1, Python 2.5.1, Stata 10.1, and 7-Zip 9.20. At minimum, the user will have to change several pathnames to these programs in the DOS batch files, in addition to those pathnames described above, to correspond to the relevant ones on his or her system.

The initial data to be downloaded use about 10 Gb of disk space. As written, the scripts require an additional ~50 Gb of free space and ~6 hours to run to create the final tables from the input files. Rewriting v4unzip.bat could make it substantially more parsimonious with space.

Replication_database.xls, a Microsoft excel file, provides an overview of the process used to create all the analysis in the paper from the raw input files. The sheet "orig_files" lists all 58 input data and script files, as well as their sources. Please cite the sources noted when using any of the individual datasets.

The sheet “scripts” lists the main 11 scripts used, along with the input and output files associated with each. These scripts are DOS batch files, AML and python scripts for ArcGIS, and stata do files. A 12th script, superbat.bat, is a DOS batch file that calls all the other scripts consecutively. (A final script, shape2cover.aml, written by Stephen Lead and in the public domain, is also included, as it is called by some of the other AML scripts).

Thirty-four separate files must be downloaded to run all analysis from the raw data:

hsw_all_other_files.zip - this contains all the data and script files required, except for those described below. It should be decompressed manually, but decompression of the files within it is carried out by the included DOS batch file v4unzip.bat. It includes replication_database.xls, which has more information on all the other files, including original sources.

F%sy%.v4.tar - each of the 30 values of %sy%, one for each satellite-year, is a separate file. It contains the global lights digital number (DN) grid, as well as the grid containing the number of nights of coverage for each pixel for that satellite-year. Image and data processing by NOAA's National Geophysical Data Center. DMSP data collected by US Air Force Weather Agency. More information is available at http://www.ngdc.noaa.gov/dmsp/

F101992.v4.tar, *F101993.v4.tar*, *F101994.v4.tar*, *F121994.v4.tar*, *F121995.v4.tar*, *F121996.v4.tar*, *F121997.v4.tar*, *F121998.v4.tar*, *F121999.v4.tar*, *F141997.v4.tar*, *F141998.v4.tar*, *F141999.v4.tar*, *F142000.v4.tar*, *F142001.v4.tar*, *F142002.v4.tar*, *F142003.v4.tar*, *F152000.v4.tar*, *F152001.v4.tar*, *F152002.v4.tar*, *F152003.v4.tar*, *F152004.v4.tar*, *F152005.v4.tar*, *F152006.v4.tar*, *F152007.v4.tar*, *F152008.v4.tar*, *F162004.v4.tar*, *F162005.v4.tar*, *F162006.v4.tar*, *F162007.v4.tar*, *F162008.v4.tar*

The final three datasets must be downloaded from http://sedac.ciesin.columbia.edu/gpw/ and placed in the same folder as the others, without decompressing them:

gl_gpwv3_pcount_00_wrk_25.zip - this is a global grid of population for the year 2000, from the Gridded Population of the World (GPW) version 3, at 2.5-minute resolution, in grid format.

gl_grumpv1_area_ascii_30.zip - this is a global grid of land area, from the Global Rural Urban Mapping Project (GRUMP), alpha version, in ASCII format.

af_grumpv1_ppoints_csv.zip - this is a dataset of settlement points for Africa, from the Global Rural Urban Mapping Project (GRUMP), alpha version, in CSV format.