Class 6: Computation & Economics – Data and Visualization

Damian Clarke

Friday March 5, 2020

Research Methods II
MRes. in Economics

Today's Plan

- Data and Data Management
  - Introduction and Background
  - Data Storage and Data Wrangling
  - Databases and Data Management
  - Geographic Data
- Visualization
- An Example

Data and Data Management

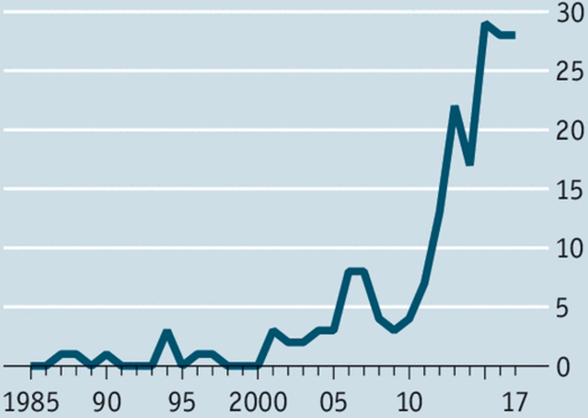
Data in Economics

Many papers in economics rely on some type of data (some graphs on next slides).

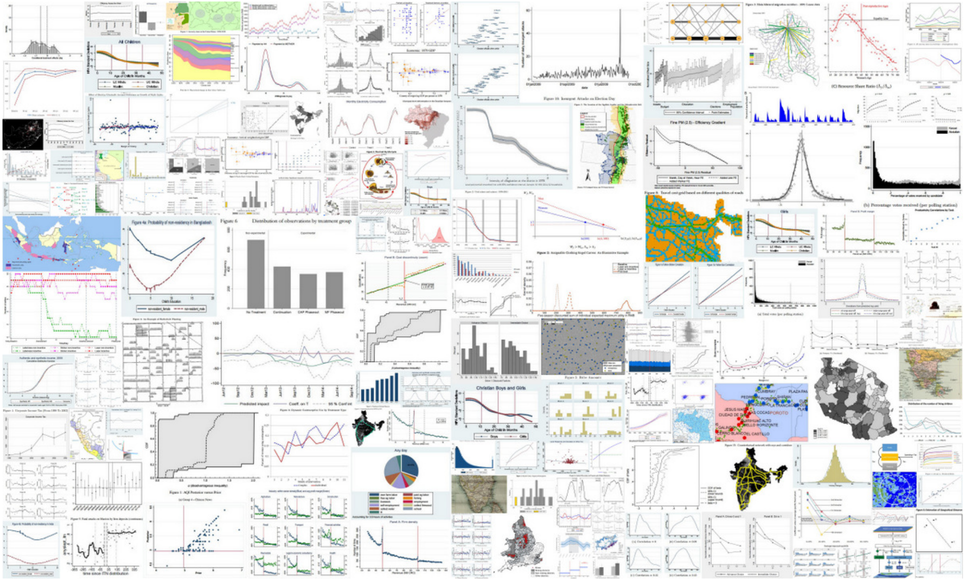
- This data is very diverse in its forms
  - "Economic indicators"
  - Administrative data
  - Text data
  - Scraped data
  - Geographic data
  - "Big data"
  - Network structures
- In general, format of data will dictate the types of methods you want to use
- But today we'll try to discuss some general principles for data use & management

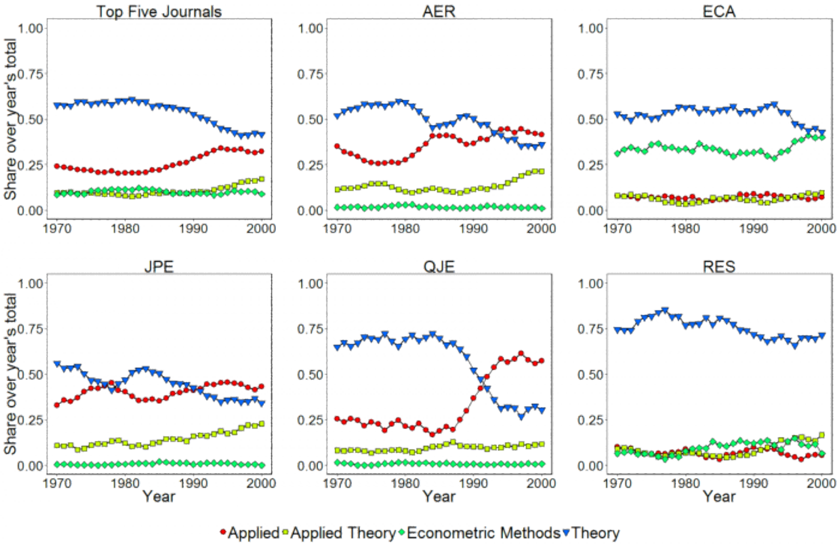
Administrative work

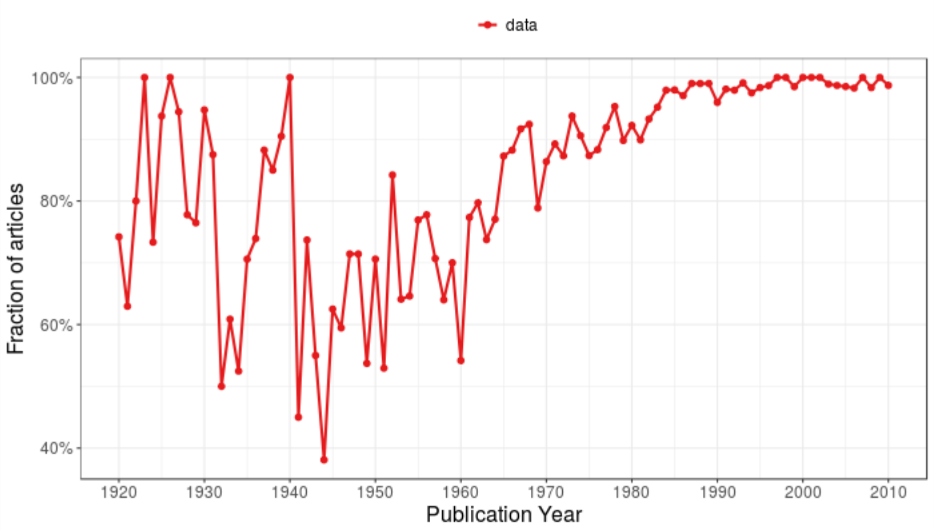
Number of NBER working papers published with "administrative data" in their abstracts



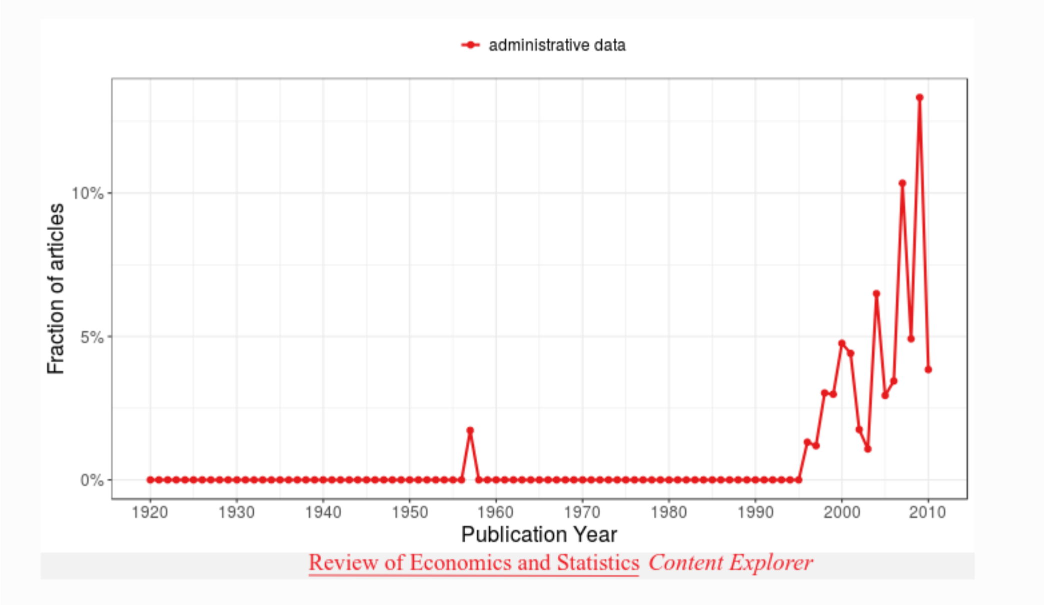
Sources: NBER; *The Economist*


All Children
Age of Child in Months
Figure 4a: Probability of non-residency in Bangladesh
Figure 6 Distribution of observations by treatment group
No Treatment
Continuation
CAP Phaseout
MP Phaseout
Predicted impact
Coeff. on T
95 % Conf int
Figure 10: Insurgent Attacks on Election Day
Fine PM (2.5) - Efficiency Gradient
(C) Resource Share Ratio
(b) Percentage votes received (per polling station)
(a) Total votes (per polling station)
Girls
Christian Boys and Girls
Boys
Panel A: Firm density
Any day
Figure 1: AQI Prediction versus Price
(a) Group 4 - Chinese News
Figure 2: Corporate Income Tax
time since ITN distribution
Price


Top Five Journals
AER
ECA
JPE
QJE
RES
Share over year's total
1.00
0.75
0.50
0.25
0.00
1970
1980
1990
2000
Year
● Applied
▪ Applied Theory
◆ Econometric Methods
▼ Theory



Review of Economics and Statistics *Content Explorer*



Review of Economics and Statistics *Content Explorer*

Explore here: `https://rest-wordcount.shinyapps.io/explorer/`

Data Storage

Often, the format of our data will be out of our control. We will get what we are given from the web, government agencies, survey teams, etc.

- This may be propriety formats, for example pdf, xlsx, Rdata, dta, ...
- They may also be stored in some markup language, eg html, xml.
- We have to deal with this.
- Do not assume stability over time, backwards compatability, or ongoing availability.

Data Storage

When accessing previously created data such as surveys, read any and all documentation available.

- Missing values are indicated in different ways (eg 99, 999, `NA`, ,`NaN`, `""`).
- Incorrect interaction with missing variables can be catastrophic
- Similarly, documentation should hopefully indicate any details about top-coding, levels in categorical variables, etc.
- Important to consider the nature of delimiters, ends of lines and ends of files

Data Programs

There are many programs for working with data in a way convenient for economic research. This includes (among many others):

- Python (+ Pandas)
- R
- Stata
- SAS
- FORTRAN
- Julia (+ DataFrames)

Data Programs

Your selection of programs for working with data will depend a lot on what you are doing.

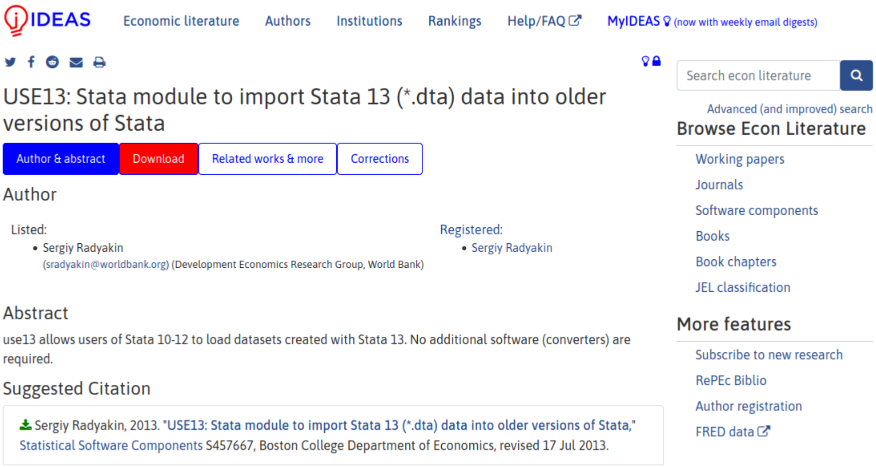
- For example, if it is mostly manipulation of strings, perhaps Python
- If it is working with well-known estimators perhaps Stata, R
- If the data is very large or you require computationally very demanding procedures, perhaps SAS or FORTRAN

Data Storage

When you are dealing with your own data storage, you do have control of how this is done and can do it well.

- Use plain text formats, eg `csv`, `tsv`, `txt`.
- Be consistent across time within a project.
- Avoid proprietry data types.
- As always, document excessively. For example, what is your delimiter and why.
- Aim to be as cross-platform as possible, for example, prefer ZIP over RAR for compressed data
- There are perhaps some exceptions to this if you are *sure* the data will only be on your OS...

The Dangers of Proprietry Data Types



USE13: Stata module to import Stata 13 (*.dta) data into older versions of Stata

Author

Listed:

- Sergiy Radyakin (sradyakin@worldbank.org) (Development Economics Research Group, World Bank)

Registered:

- Sergiy Radyakin

Abstract

use13 allows users of Stata 10-12 to load datasets created with Stata 13. No additional software (converters) are required.

Suggested Citation

Sergiy Radyakin, 2013. "USE13: Stata module to import Stata 13 (*.dta) data into older versions of Stata," Statistical Software Components S457667, Boston College Department of Economics, revised 17 Jul 2013.

Data Wrangling

Data wrangling refers to the process of dealing with raw input data and converting it into a format more amenable for future analyses.

- This includes importing data, cleaning data, converting to the required format, exporting data
- This can include the use of tools like regular expressions (discussed more next class)
- There are many good language-specific tutorials online (eg searching for “data wrangling + `language`”)
- Likely data wrangling will be the majority of time spent in data analysis in an empirical paper

Data Wrangling

Some general processes that are likely to be necessary when moving from raw data to analysis include:

1. Collapsing (or summarizing) a dataset
2. Merging or joining multiple datasets
3. Appending or stacking multiple datasets
4. Reshaping datasets (from wide to long or vice versa)

We will look into these in the extended example in this class.

Data Management

There are also many ways to deal with relational databases. Database software allows us to efficiently deal with relationships where there are various possible links of interst, and possible redundancies when stored in a single file.

- A simple example: health records
  - Data on patients including their personal characteristics, diagnoses, pharmaceutical usages, hospital visits, etc.
  - Information can be stored separately: Hospitals; Diagonsis codes; Drug codes; Patient information
  - All can be linked using some type of unique key
- The database consists of the data itself, plus the metadata on what it contains
- Administrative data is often stored in this way

Data Management and Databases

There are many Data Base Management Systems for dealing with database queries. These are based on an underlying query language, and often used with a graphical user interface (GUI).

- DBMS (among many others):
  - MySQL
  - SQLite
  - Oracle
  - Microsoft Access
- Query Languages
  - Most common is SQL
- GUIs:
  - Microsoft Access

Information on the popularity of different DBMS is available here: `https://db-engines.com/en/`

Data Management and Databases

Note that even if you are working with database files (eg `.mdb`, `.accdb`) one option is to work with the data from directly within your statistical programming language...

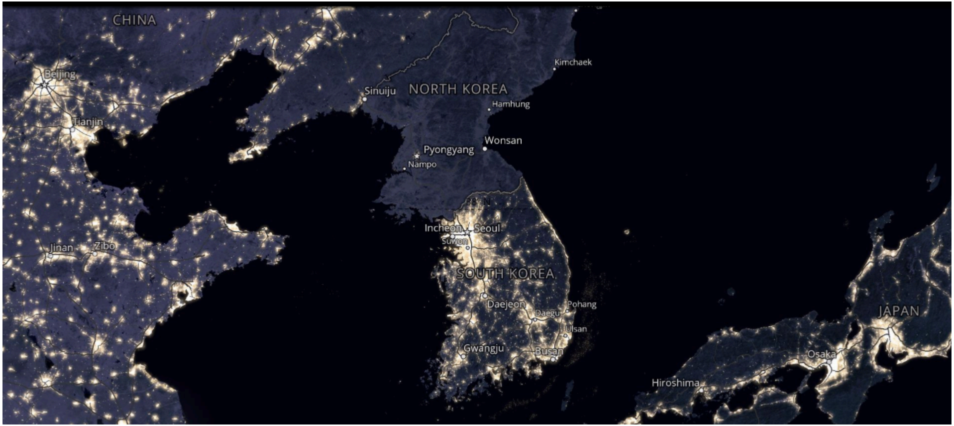
- dplyr+dbplyr or RSQLite in R
- odbc in Stata
- MySQLdb in Python

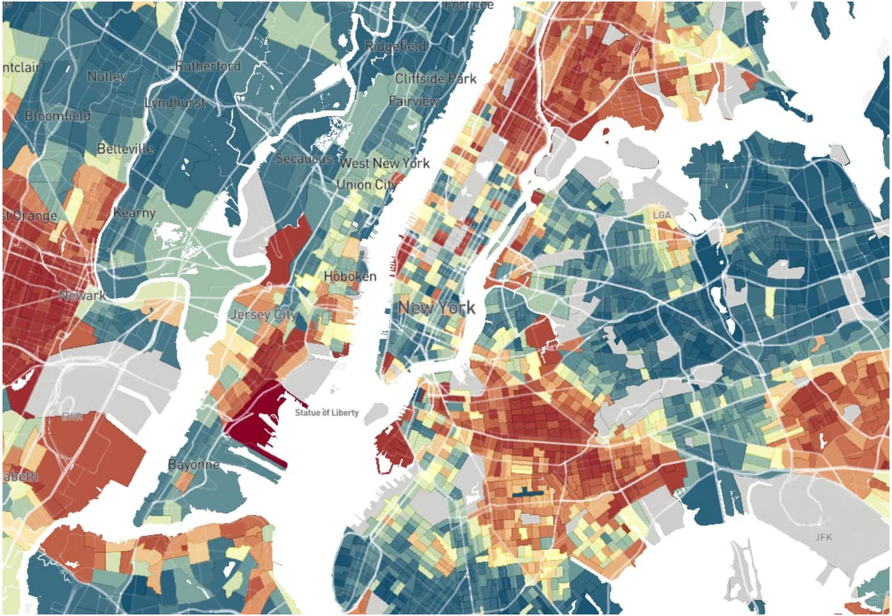
Geographic Data

Certain types of data are tricky (though not impossible) to deal with in our standard statistics packages. This includes network data and geographic data.

- In these cases it may be worth exploring other options, for example qGIS or ArcGIS for mapping
- Geographic data can allow for both deep insights, and highly local results to capture heterogeneity
- Some examples...

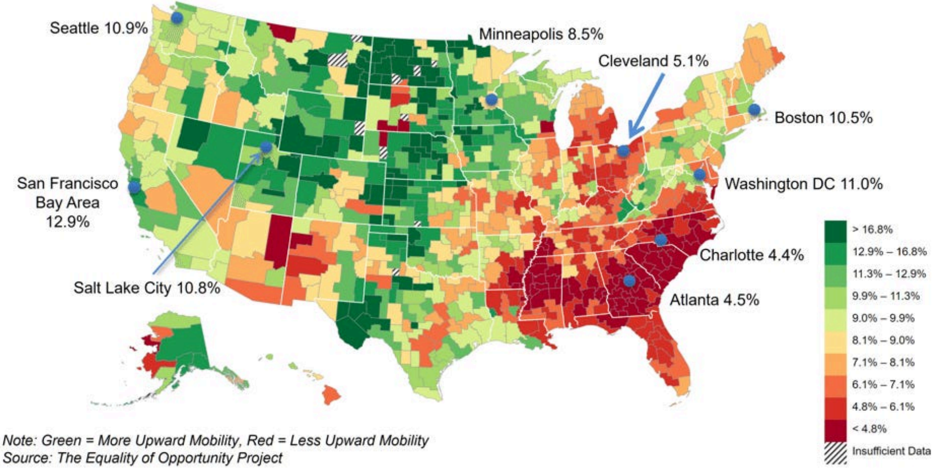
Night lights




Fort Lee
ntclair
Nutley
Rutherford
Ridgefield
Cliffside Park
Fairview
Lyndhurst
Bloomfield
Belleville
Secaucus
West New York
Union City
Kearny
st Orange
LGA
Hoboken
Newark
New York
Jersey City
Statue of Liberty
EWR
Bayonne
abeth
JFK

The Geography of Upward Mobility in the United States

Chances of Reaching the Top Fifth Starting from the Bottom Fifth by Metro Area



Note: Green = More Upward Mobility, Red = Less Upward Mobility
Source: The Equality of Opportunity Project

Visualization

Visualization

Effectively presenting data in a visual way is a very useful to explore relationships, and allow your audience to understand what you are studying.

- In general, you should invest a lot of time in examining data with tabulations, cross tabulations, and one way and two way plots
- This additionally acts as a check for possible issues in data (eg strange missings, outliers, etc.)

Visualization

There are many resources online documenting different plot types (in a language specific way). For example:

- Julia: `https://plot.ly/julia/`
- MATLAB: `https://la.mathworks.com/help/matlab/creating_plots/types-of-matlab-plots.html?lang=en`
- R: `https://www.r-graph-gallery.com/`
- Stata: `https://www.stata.com/support/faqs/graphics/gph/stata-graphs/`
- Python: `https://matplotlib.org/3.1.3/tutorials/introductory/sample_plots.html`

Visualization

For further general information in your future work it is likely worth checking out:

- For economics: “An Economist’s Guide to Data Visualization” (Schwabish, Journal of Economic Perspecitves, 2014). Free at `https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.28.1.209`
- More generally: Leland (2012) “The Grammar of Graphics”, Kirk (2016) “Data Visualisation: A Handbook for Data Driven Design”

An Example

An Example: Visualizing Geographic Data

To try a few of these things, I want to look at a ‘real’ example.

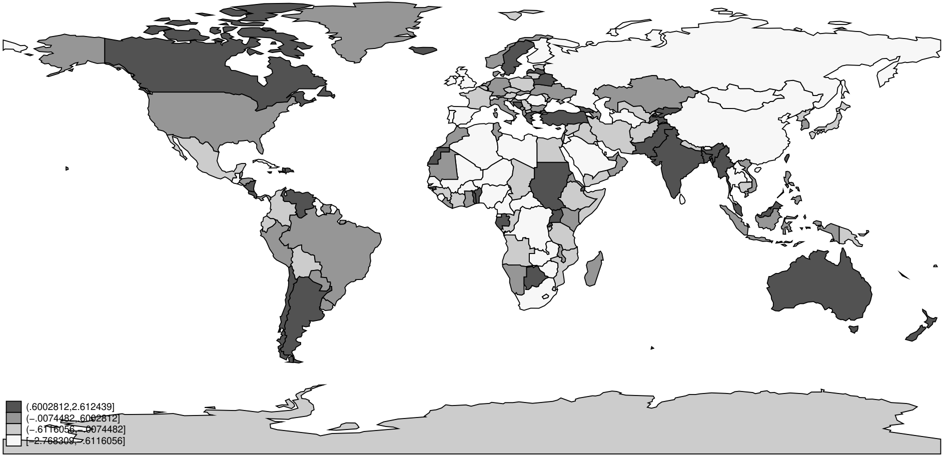
- This is ‘real’ in the sense that we will download recent data off the web
- We will also download all source files necessary and convert them to required formats
- We will do all data wrangling
- We will look at the difference some reasonable stylistic choices can make

An Example: Visualizing Geographic Data

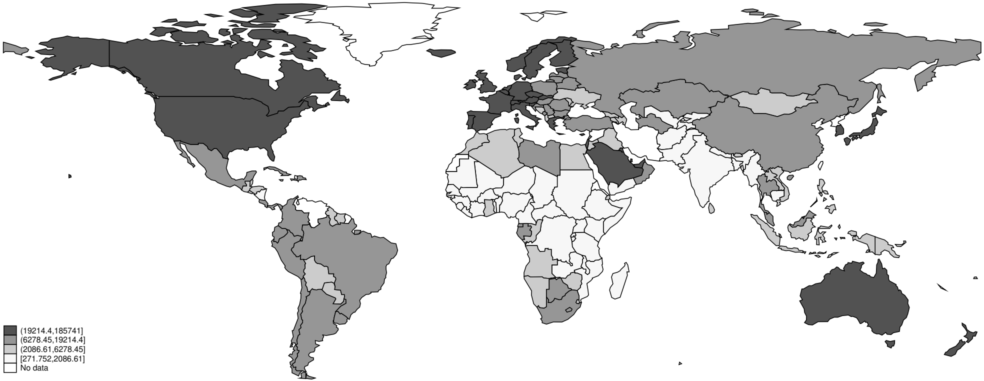
We will do this in Stata, though you could certainly do it in many other languages. Let's look at the "do file" mapping.do. We will require data or tools on the following sites:

- `https://acleddata.com/#/dashboard` (Armed conflict data)
- `http://thematicmapping.org/downloads/world_borders.php` (World map)
- `https://mapshaper.org/` (Auxiliary tool)
- World Bank data bank

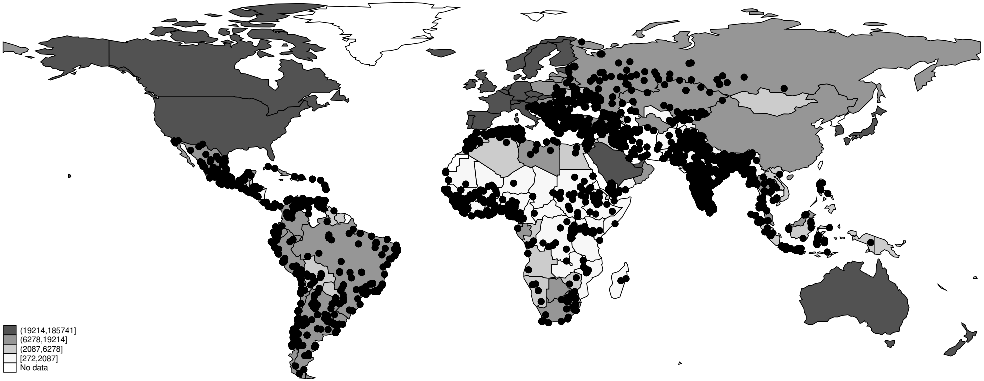
Building Up Progressively (1)



Building Up Progressively (2)

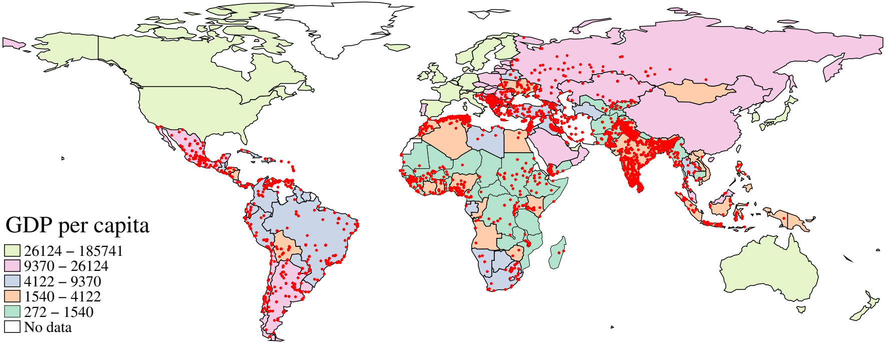


Building Up Progressively (3)



Building Up Progressively (4)

GDP per Capita and Protests (Jan 2020)



Source: World Bank (GDP data) and The Armed Conflict Location & Event Data Project.