UNIVERSITY
OF OSLO

Survey analysis and data bases

What data exist and how can we use them?

Martin Moland

ARENA, Centre for European Studies

September 10 2024

Overview

❶ Introduction

❷ What is out there?

❸ Merging macro and micro data

❹ Putting it all together

❺ Working with time series

❻ Good data habits and open science

About my background

- Postdoc researcher at ARENA, Centre for European Studies
- Study public opinion (what people believe about political issues)
- Lectures will hopefully be useful for everyone, but will probably be biased towards my own interests in attitudes. . .

Why use existing data sources?

- It is generally **freely available**
- Most of these surveys have been periodically **fielded for a long time:**
  - Allows you to construct time series

What are the drawbacks of existing data?

- Data are **not customized** for your particular case:
  - May make it more difficult to measure precisely what you are after.

- The **lack of harmonization** could present a problem:
  - Sometimes data are messy in ways you can’t really anticipate.

Global data I

- **Biggest cross-national data: SDR**
  - https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YOCX0M

- SDR has a **geographical reach** (Europe, Africa, US etc…) that is unparallelled, but comes with some caveats:
  - Has a very **limited number of variables** (mostly voting behavior and very general attitude questions)
  - Some **variations in the coverage** of the surveys
  - May not be for you if you are interested in **studying attitudes**

Global data II

- **World Values Survey**
  - https://www.worldvaluessurvey.org/WVSContents.jsp

- Very **broad coverage,** also of countries that are typically not covered in most surveys (such as Algeria).

- However, like all non-customized data it comes with some caveats:
  - Survey waves are fielded over **many years**.
  - Makes it incredibly important to **control for time**.

- Includes **extensive data on attitudes**

European data I

- **European Social Survey:**
  - https://www.europeansocialsurvey.org/data-portal
  - Includes both **trends** (like democratic satisfaction) and **special modules** that are not frequently repeated.

- Surveys are fielded over multiple years (like WVS)
  - Last round was fielded in all countries between 2023 – 2024.

- Has one clear benefit: The data are harmonized over time, so very little cleaning needed.

European data II

- **Eurobarometer**
  - https://www.gesis.org/en/eurobarometer-data-service

- If you are interested in **EU issues**, this is the only game in town.
  - Has harmonized data from 1973 – 2002 and 2004 – 2021.

- **One large problem**:
  - Questions tend to drop out of surveys from time to time, creating gaps in the time series.

Scandinavian data

- **Swedish Citizens' Panel (SCP):**
  - https://www.gu.se/en/som-institute/the-swedish-citizen-panel/swedish-citizen-panel-for-researchers/panel-waves-and-technical-reports
  - Panel data, so allows for within-person comparisons.

- **Norwegian Citizens' Panel (NCP):**
  - https://www.uib.no/en/digsscore/122111/norwegian-citizen-panel

- Latter includes **KODEM**, which asks identical questions of citizens, politicians and bureaucrats:
  - https://www.uib.no/en/digsscore/169533/kodem-data-and-conditions-use

Other data

- **More specialized surveys:**
  - Eurostudent survey, which surveys European students, is one example

- **Surveys done by individual researchers**
  - Occasionally difficult to access
  - These data are increasingly put on Dataverses (more on those later). If you can’t find any link to it in the paper, send an e-mail.

A problem with data belonging to specific papers is that they typically aren’t repeated over time.

Your ideas

- What kind of questions are you hoping to answer in your theses/semester papers?

- What kind of geographic areas are you interested in?

Combining context and individuals I

- Sometimes you want to study the **interplay between individuals and contextual factors** (like economic growth).

- Most **international organizations** have databases with the statistics they collect. These can be merged with survey data.

- **Only one problem:** The data may be at a different timescale (monthly rather than yearly, for instance).

Combining context and individuals II

- **One large provider is the World Bank:**
  - You can access the World Bank API through the *R* package *wbstats*.
  - Contains most of the statistics you are likely to need, whether on GDP, population or other economic indicators.

- **Another possibility is the International Monetary Fund:**
  - Accessible through *R* with the *imfr* package
  - Mainly relevant for those of you who need international finance statistics, and data related to trade.

Combining context and individuals III

- **For European data Eurostat is your best source:**
  - Relatively standardized data on everything from trade and GDP to birth rates.
  - Only contains data on the countries that are in Europe (logically), but includes both EU and non-EU countries.
  - Accessible through the *R* package *eurostat*.

- **Many European countries have statistics agencies with accessible data:**
  - Best bet is to look around. Some may have data accessible through R, others require some legwork.
  - In Norway this is SSB. Their data can be accessed either through R or by downloading Excel files from Statistikkbanken.

One practical example

Moving from the abstract to the concrete: Let’s try to put all of this together.

- Hypothetical example:
  - You want to test how **election quality** shapes **trust in politicians**

One practical example

Moving from the abstract to the concrete: Let’s try to put all of this together.

- Hypothetical example:
  - You want to test how **election quality** shapes **trust in politicians**
- We first need measures of trust in politicians
- Then a measure for election quality

One practical example

Moving from the abstract to the concrete: Let’s try to put all of this together.

- Hypothetical example:
  - You want to test how **election quality** shapes **trust in politicians**
- We first need measures of trust in politicians
- Then a measure for election quality
- However, one problem remains:
  - A lot of macro-factors can confound the relationship.

One practical example

Moving from the abstract to the concrete: Let's try to put all of this together.

- Hypothetical example:
  - You want to test how **election quality** shapes **trust in politicians**
- We first need measures of trust in politicians
- Then a measure for election quality
- However, one problem remains:
  - A lot of macro-factors can confound the relationship.

Measures of trust

- We use ESS' **trstplt** (trust in politicians) as DV.
- 0 means no trust, 10 means full trust
- "Using this card, please tell me on a score of 0-10 how much you personally trust each of the institutions I read out. [...] Firstly... ...politicians?"

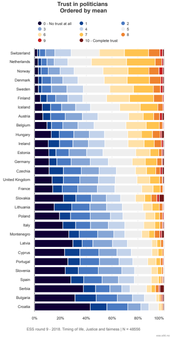
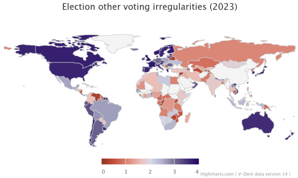


Figure: Trust in politicians

Democracy indicator

- We use election and voting irregularities as measured by **V-DEM** as our democracy indicator.
  - The variable is scaled between 0-4.
  - Country names are full names (Sweden, Norway etc.)


Election other voting irregularities (2023)

Highcharts.com | V-Dem data version 14 |

Figure: Election and other voting irregularities

Important control variables

- **GDP variation**:
  - We can get this from Eurostat and most other international organizations.
  - One possible measure is GDP per capita. Operates on a € scale.
  - Eurostat uses two-letter country codes to indicate countries (NO for Norway etc.)

- **Quality of governance**:
  - We can get this from the Quality of Governance dataset.
  - The Freedom House quality of governance index gives us what we are after (but is only one of many options). It is scaled between 0-12.
  - Uses full names for country variables.

The Problem of Harmonization

- **Common Issues:**
  - *Unit Names:* Inconsistent coding (e.g., NOR, Norway, NO), whether related to countries or other geographic units.
  - *Variable Scales:*
    - Different scales (e.g., Likert scale (1-5), dollars, etc.).

Everything just said also applies to merging data within countries (county names are likely to be just as chaotically coded as country names)

If using contextual data from more than one organization, make sure they use the same definitions (for example of what counts as higher education or not).

How Do You Merge It All?

- **Suggested Approach:**
  - Start with the survey as the base.
  - When merging, make sure the ***N* is the same** before and after.

- **If creating time series:** Make sure to merge on both the unit and time.

- **For cross-sections:** You only need to merge on the unit.

How to merge datasets in R

Say we have a dataset called “trust_gov” (our survey), “control_data” (our country-year controls). We want to merge the two. Our unit- and time variables are simply “country” and “year”.

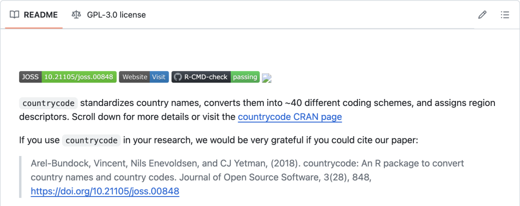
For country-year formats:

```
library(tidyverse)
complete_data = left_join(trust_gov, control_data,
by = c("year", "country"))
```

If we only have one time period:

```
library(tidyverse)
complete_data = left_join(trust_gov, control_data,
by = c("country"))
```

Handling unit name inconsistencies



`countrycode` standardizes country names, converts them into ~40 different coding schemes, and assigns region descriptors. Scroll down for more details or visit the countrycode CRAN page

If you use `countrycode` in your research, we would be very grateful if you could cite our paper:

> Arel-Bundock, Vincent, Nils Enevoldsen, and CJ Yetman, (2018). countrycode: An R package to convert country names and country codes. Journal of Open Source Software, 3(28), 848, https://doi.org/10.21105/joss.00848

Figure: https://github.com/vincentarelbundock/countrycode

- Biggest hassle is often handling **different country codes**
- *R* package ***countrycode*** can speed up the process:
  - Can convert country codes to and from 40+ different coding schemes

Example code

We have V-DEM data for our main independent variable. The unit variable in V-DEM is **country_name**. It uses the full names of variables. To merge this with, for instance, GDP data from **Eurostat** we need to recode V-DEM country names to Eurostat codes. Here’s one way to do it via *countrycode*:

```
library(countrycode)
library(tidyverse)

V_DEM = V_DEM %>%
  mutate(country_new = countrycode(country_name,
  "country.name", "eurostat"))
```

Standardizing variables

- Standardization means converting all variables to a **common scale**
  - Makes it easier to compare the effect sizes
  - Particularly important if your data includes, for instance, stuff like both GDP and income.
- When standardizing variables are **centered and divided by standard deviation**
  - This does not change the inference (i.e. the p-value will be identical)

The code in R is very simple (I use GDP as an example):

```
library(tidyverse)
complete_data = complete_data %>%
    mutate(gdp_scaled = scale(gdp))
```

What to do if you want to test changes over time?

- **This is where existing data really shine**:
  - Most large survey projects have data from a number of years.

- **Plenty of examples**:
  1. Eurobarometer: 1972 – 2024
  2. World Values Survey: 1982 – 2022
  3. Afrobarometer: 1999 – 2024

- **Usually requires tough choices**:
  1. What are you interested in testing? Do the data actually exist?
  2. Will you be able to actually test this in the same data over time?

Measuring what you want across years

- **Construct validity is a challenge in time-series**:
  - What people mean by different things will change over time

- **Also an issue with indices**:
  - Not a given that the same things will correlate to an underlying construct over a long period.

- **One important rule**:
  - Run tests of construct validity (factor analysis etc.) for different periods of your time series.

Good data habits (I)

- **First off, why should you care?**
  - It saves you time in the long run.
  - Remembering what you did two months ago is usually impossible.
  - Maybe most importantly: You need to submit R (or other software) files with your thesis. It’s helpful to know where everything is.

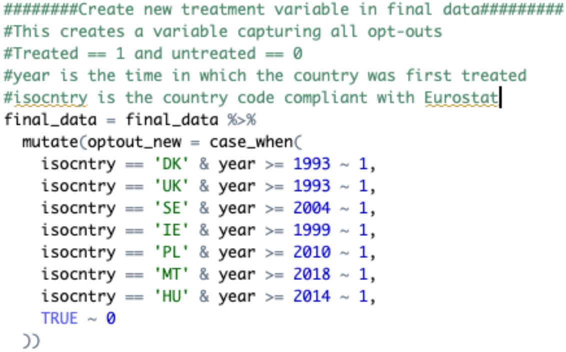
Good data habits (II)

- **Have different folders:**
  - Figures
  - Datasets
  - Scripts

- **Use separate scripts for each procedure:**
  - Makes it much easier to figure out what you did at the beginning of a project.

- **Add (a lot of) comments to your code**:
  - Best practice is to use # to tell yourself (and others) precisely why you coded things the way you did
  - RStudio also allows you to use ### as a way to create subheadings (really makes life easier)

A practical example (I)



```
########Create new treatment variable in final data#########
#This creates a variable capturing all opt-outs
#Treated == 1 and untreated == 0
#year is the time in which the country was first treated
#isocntry is the country code compliant with Eurostat
final_data = final_data %>%
  mutate(optout_new = case_when(
    isocntry == 'DK' & year >= 1993 ~ 1,
    isocntry == 'UK' & year >= 1993 ~ 1,
    isocntry == 'SE' & year >= 2004 ~ 1,
    isocntry == 'IE' & year >= 1999 ~ 1,
    isocntry == 'PL' & year >= 2010 ~ 1,
    isocntry == 'MT' & year >= 2018 ~ 1,
    isocntry == 'HU' & year >= 2014 ~ 1,
    TRUE ~ 0
  ))
```

Figure: Example R code with comments

A practical example (II)

- **Here I do three things**:
  - First, show what the variable is
  - Second, show what the variable values actually mean
  - Third, show what each variable means in substantive terms.

Helps people who don’t know the data figure out exactly what’s happening.

Open and reproducible science

What is open and reproducible science:

- **Open:** Data and code transparency
- **Reproducible:** Every researcher should be able to reproduce your work

Transparency is important for a bunch of reasons:

- Lets other people “check your work”
- Lets others use the data you have (painstakingly) cleaned
- Shows that you have nothing to hide (always a good thing for a researcher)

Want to know more?

The Plain
Person's Guide

~/>_

to Plain Text
Social Science

Kieran Healy

Figure: Kieran Healy's The Plain Person's Guide to Plain Text Social Science

If you want to learn more, Kieran Healy's book is very useful as a guide.

How to make it open and reproducible?

- **Make it “one click”:**
  - Ideally anyone should be able to run your script (using one click) and get exactly the same results as you.

- **Use seeds**:
  - Computers are random number generators (meaning many results will be slightly different even if you do the same thing)
  - Seeds (that you set using the set.seed command in R) tell the computer to choose a solution that gives exactly the same result every time you run a command.

How to use seeds

You use the set.seed command in R to choose the seed. For our hypothetical example it would look like this:

```
set.seed(1234)
trstplt_reg = lm(trstplt ~ election_fairness + gdp +
fh_qog, data = complete_data)
```

Now the result will be **identical** every time you run the code!

Data transparency

- **First, check the requirements:**
  - UiO might have rules for whether and how you should publish data you use for your theses.

- **If you want to publish data, use a “Dataverse”:**
  - The biggest one is Harvard Dataverse (https://dataverse.harvard.edu/, but a bunch of others also exist)