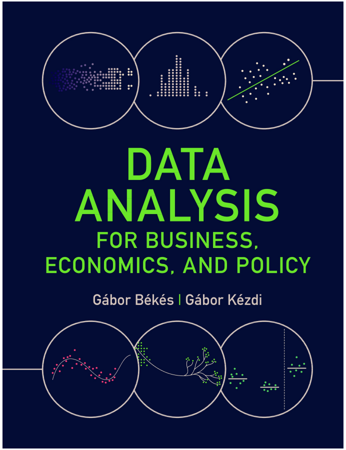
01 Origins of Data

Gábor Békés (CEU)

Data Analysis 1: Exploration

2020

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021
- **gabors-data-analysis.com**
  - Download all data and code `gabors-data-analysis.com/data-and-code/`
- This slideshow is for **Chapter 01**

Motivation

- *Suppose, you want to understand the extent and patterns of differences in online and offline prices. A super project, the Billion Prices Project at MIT did a variety of data collection approaches such as crowd sourcing platforms, mobile phone apps and web scraping methods.*
- *Interested in understanding more about management practices? The World Management Survey is a major effort by academics to survey practices around the world - asking the same questions in many country the same way.*

Roadmap

- What is data
- Data structure
- Data quality
- Data collection:
  - Various methods and data sources
  - Sampling
  - Good practice
  - Big data revolution
  - Ethics
- Variable types

What is data

- Data is most straightforward to analyze if it forms a single data table.
- Format: Data table (matrix)
- A data table consists of *observations* and *variables*.
  - Observations are also known as cases, or rows
  - Variables are sometimes called features or covariates.
- In a data table the rows are the observations, columns are variables.
- Storage: comma separated values .csv (.txt) is simplest. Delimited can be anything: comma(,), semicolon (;) or other (|)
- A dataset is a collection of data tables, typically related / used in a project
  - 10 data tables, same topic for 10 different years

Data structures

- Cross-sectional (xsec) data have information on many units observed at the same time.
- Time series (tseries) data have information on a single unit observed many times.
- Multi-dimensional (panel) data have multiple dimensions.
  - Many cross-sectional units observed many times
  - Units observed in different space

Data structures

A bit more on multi-dimensional - panel (xt) data

- A common type of panel data has many units, each observed multiple times. Such data is sometimes called *longitudinal data*, or cross-section-time-series data, sometimes abbreviated as *xt data*.
- Example: countries observed repeatedly for several years
- In xt data tables observations are identified by two ID variables: one for the cross-sectional units, one for time.
- xt data is *balanced* if all cross-sectional units are observed at the very same time periods. It is called unbalanced if some cross-sectional units are observed more times than others.

Case Study: Finding a good deal among hotels: data collection

- Welcome to Vienna, Austria
- `hotels` dataset
- collected from a price comparison website. Anonymized.
- Vienna, 2017 November weekday, $N = 428$
- For each hotel the data includes information on the location of the hotel, the price on the night in focus in EUR, average customer rating, stars of the hotel, distance to the city center .

Image: en.wikipedia.org/wiki/File:Montage_of_Vienna.jpg

Data structures: Case Study

Table: **List of observations**

hotel_id	accom_type	country	city	city_actual	dist	stars	rating	price
21894	Apartment	Austria	Vienna	Vienna	2.7	4	4.4	81
21897	Hotel	Austria	Vienna	Vienna	1.7	4	3.9	81
21901	Hotel	Austria	Vienna	Vienna	1.4	4	3.7	85
21902	Hotel	Austria	Vienna	Vienna	1.7	3	4	83
21903	Hotel	Austria	Vienna	Vienna	1.2	4	3.9	82

Source: `hotels` dataset. Vienna, for a 2017 November weekday

List of five observations with key variable values:

- ‘accom_type’ is the type of accommodation.
- ‘city’ is the city based on the search, city_actual is the municipality.

Data quality is key

- Data quality is key
- If our data is useless to answer our question the results of our analysis are bound to be useless...
- ... no matter how fancy method we apply to it.

Data quality and your question

Data quality is generally a subjective notion!

- First you have to specify what is your (research) question!
- What do you want to explore or understand?
- If you have a clear answer, then you can decide on your data quality!

However, there are some objective measures to decide if you have your question!

Data quality

1. Content - what is the substance a variable captures. Always check details.
2. Validity - is the content of variable close to intended content. "Durability" vs "Quality"
3. Reliability. If we were to measure the same variable multiple times for the same observation it should give the same result.
4. Comparability in measurement across observations.
5. Coverage. Ideally complete coverage. In practice, they may not include all planned units (incomplete coverage).
6. Unbiased selection. In incomplete coverage, observations that are included should be similar to all observations that were intended to be covered.

Sidenote

- This is not the type of class where you will have to memorize a list.
- But you should be able to judge the quality of variables in work.
- And you should always remember: **GIGO**: garbage in, garbage out.

Data analysts should know their data

.

- How data was born
- All details of measurement that may be relevant for their analysis

To this end, consider having

- README.txt that describes where dataset comes from
- VARIABLES.xls that provides basic information on your variables

Data collection

- Automated data collection
- Survey
- Administrative / Census
- Big Data

Data collection: Digital

Automated data collection

- Application Programming Interface, or API – directly load data into a statistical software.
  - API is a software intermediary, or an interface,
  - It allows programs, or scripts, to talk to each other.
- API widely used in many context.
  - Macro data: FRED - St Louis Fed at `research.stlouisfed.org/docs/api/fred/`, also World Bank, etc.
  - Micro data such as weather at: `openweathermap.org/api`
- Data collection limited to dataset.
- Typically additional info available.

Data collection: Digital

Automated data collection

- Web scraping - collecting data from online platform
- html code includes data, can be found, analyzed and collected
- Need extensive cleaning
- Once a procedure is ready (code, script), can be repeated
- Data collection limited to what is on a site

Data collection: Administrative

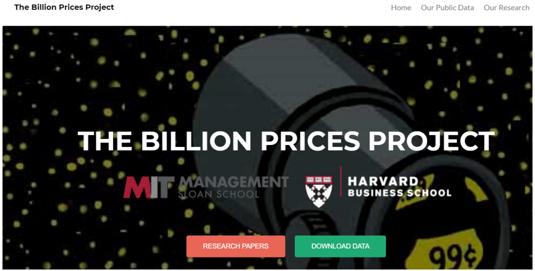
- Business transactions
- Government records, taxes, social security
- Often: census - records on the population
- Many advantages
  - Often great coverage, few missing values, high quality content
  - Many well defined and documented variables
- Some disadvantages
  - Variables defined for business/government purposes. May not fit in analysis plans
  - Often not detailed/specific enough
  - Biggest problem is **very limited** access

Case Study I - Finding a good deal among hotels: data collection

- The dataset on hotels in Vienna was collected from a price comparison website, by web scraping.
- On a specific date
- The purpose of the website is not facilitating data analysis...
- No other potential source
- Good quality, but noise, needed work to make it ready for analysis.
- Coverage is good but not full. Hotels advertising on these websites are not a random sub-sample. Which are the hotels that are left out?

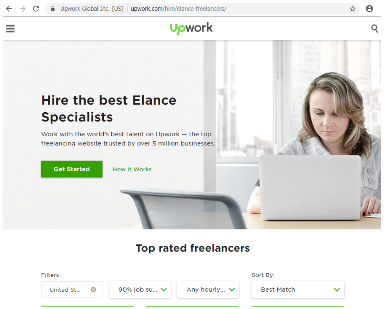
Case Study II - Comparing online and offline prices: data collection

- The Billion Prices Project - academic initiative - product prices collected
- This course: Cavallo (2017, AER)
- 56 large multi-channel retailers in 10 countries.
- price levels identical about 72 percent of the time.
- Price changes are not synchronized but have similar frequencies and average sizes.



Case Study II - Comparing online and offline prices: data collection

- BPP is about measuring prices for the same products sold through different channels
- Mixed methods
- Offline data collectors by Mechanical Turk / Upwork
- Online prices were scraped
- Project managers focusing on collecting info on exactly the same products on approximately during the same time


Upwork Global Inc. [US] | upwork.com/hire/elance-freelancers/

Hire the best Elance Specialists

Work with the world's best talent on Upwork — the top freelancing website trusted by over 5 million businesses.

Top rated freelancers

Filters: United St... 90% job su... Any hourly...

Sort By: Best Match

Data quality - billion prices project data

1. Content - what product, what price
2. Validity - intention is price of target product available at store. What could go wrong?
3. Reliability. Timing is very difficult especially if price change frequently
4. Comparability in measurement- are products *equally* well identified? Laptop vs cheese
5. Coverage. Not universal. Project plan choice.
6. Unbiased selection. Time consuming planning. If electronic goods, need a typical set of TVs, phones etc.

Data collection: Survey

- Surveys collect data by asking people (*respondents*) and recording their answers.
- Answers to a *questionnaire* are short and easily transformed into variables.
- Major advantage: you can ask exactly what you want to know

- There are two major kinds of surveys: self-administered surveys and interviews.
- Web, telephone, in person, mix - computer aided interview.
- Choice of data collection approach matters a great deal.
- Self-administered survey
  - cheap and efficient, can use visual aids.
  - What could go wrong?

Sampling

- In many cases, we can collect data on all the people we care about (= the *population*). Often this is not possible...
- For cost/time reasons, we need to take a sample - this is the process of *sampling*.
- Samples have to represent the population. A sample is *representative* if the distribution of all variables in the sample are the same as, or very close to, their corresponding distribution in the population.
  - The distribution of variables is the frequency of their values, e.g., fraction female, percent with income within a certain range. (*More on this in Class 3.*)

Sample: Representativeness

- The difficulty is: whether a sample is representative is impossible to tell directly.
- There are two ways of assessing whether a sample is representative:
- Evaluating the data collection *process* - subjective with objective elements
- *Benchmarking* the few variables for which we know the distribution in the population.
    - For instance, there may be some national statistics.
    - Or very similar businesses collected data.
    - Reality check always really useful

Sampling: Random samples

- *Random sampling* is the process that most likely leads to representative samples.
- All observations in the population have the same chance of being selected into the sample.
- In practice: randomization rule (e.g. flip a fair coin)
- Any other methods that are not randomly picking observations may yield an unexpected bias thus preventing our sample from being representative.
- Practically just like random sampling include fixed rules that are unrelated to the distribution of variables in the data.
- Examples?

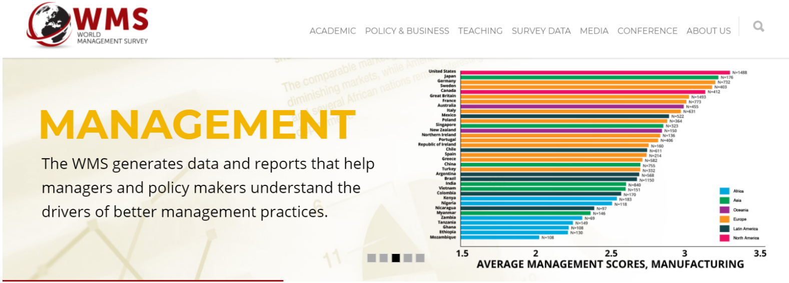
Sampling: Random samples

- In small samples (dozens-few hundred) anything is possible.

- Sample of a several thousand observations may equally well represent populations of fifty thousand or ten millions
- The required sample size depends on details of what you want to measure!
- More on this topic later

Case Study III - Management quality and firm size: data collection

- What causes superior performance of some countries? What causes superior performance of some firms in some countries?
- Many potential arguments: Institutions that lead to competitive markets. Education that helps research - yields new patents
- www.worldmanagementsurvey.org - Massive survey on firm features and management.



WMS WORLD MANAGEMENT SURVEY

ACADEMIC POLICY & BUSINESS TEACHING SURVEY DATA MEDIA CONFERENCE ABOUT US

MANAGEMENT

The WMS generates data and reports that help managers and policy makers understand the drivers of better management practices.

Case Study III - Management quality and firm size: data collection

- Ask 10K+ manufacturing firms (also public sector)
- Developing management questions
  - Scorecard for 18 monitoring, targets and incentives practices
  - Approx 45 minute phone interview of manufacturing plant managers
- Obtaining unbiased comparable responses ("Double-blind")
  - Interviewers do not know the company's performance
  - Managers are not informed (in advance) they are scored
  - Run from London, with same training and country rotation
- Getting firms to participate in the interview
  - Introduced as "Lean-manufacturing" interview, no financials
  - Run by 100+ MBAs (credible with business experience)

Case Study III - Management quality and firm size: data collection

Example question: "how is performance tracked?"

- **(1)**: Measures tracked do not indicate directly if overall business objectives are being met. Certain processes are not tracked at all.
- **(3)**: Most key performance indicators are tracked formally. Tracking is overseen by senior management.
- **(5)**: Performance is continuously tracked and communicated, both formally and informally, to all staff using a range of visual management tools.

Case Study III - Management quality and firm size: data collection

- Survey quality assessment
- Content of each score - based on information gathered in a standardized way translated to scores by the interviewers using standardized rules.
- Validity, reliability and comparability - How to think about assessment?
- What would be an alternative? Pros and Cons?

What is different with Big Data?

- Big Data refers to: (i) massive (very large) datasets that are (ii) often automatically and continuously collected and stored, and (iii) may be of complex nature.

(i) Very large. Billions of observations. (Bigger than what fits into your computer.)

  - Warning: just because sample is large, it is not necessarily representative!!!!

(ii) Automatic collection. Not for your analytic purpose - unlike a survey. Data collected by apps, sensors.

(iii) Complex - text (video, music/noise), network, multidimensional, maps

Sample selection bias

- The sample you collect is different to the population
- This difference is crucial in the story
- Example: Predicting presidential election
  - 1936: Literary Digest. FD Roosevelt vs Landon. 10m people asked. 2m replied. Biggest poll ever. Landon was predicted win 57%
  - What could have gone wrong?

Legal and ethical aspects

- Data collection - ethical and legal constraints
- Especially with sensitive information
- GDPR

Always communicate with the source owner(s) and or with legal professional if you are planning to use seemingly sensitive data!

Data collection: hard, time-consuming, costly.

- Collecting data is tedious task, costly as well.
- Usually it is not as simple as you think...
- Collect your experience with the data collecting assignment!

Summary

How is your data? ?

- Data quality, such as poor coverage (large share of missing observations), will determined what you can do with the data.
- Data may come from existing sources (such as tax authority, World Bank) or you may need to carry out a survey. Surveys may be more befitting but expensive and time consuming.
- Representative sample is essential for any any analysis. Even with big data.
- To respect data confidentiality is a key ethical rule to follow.