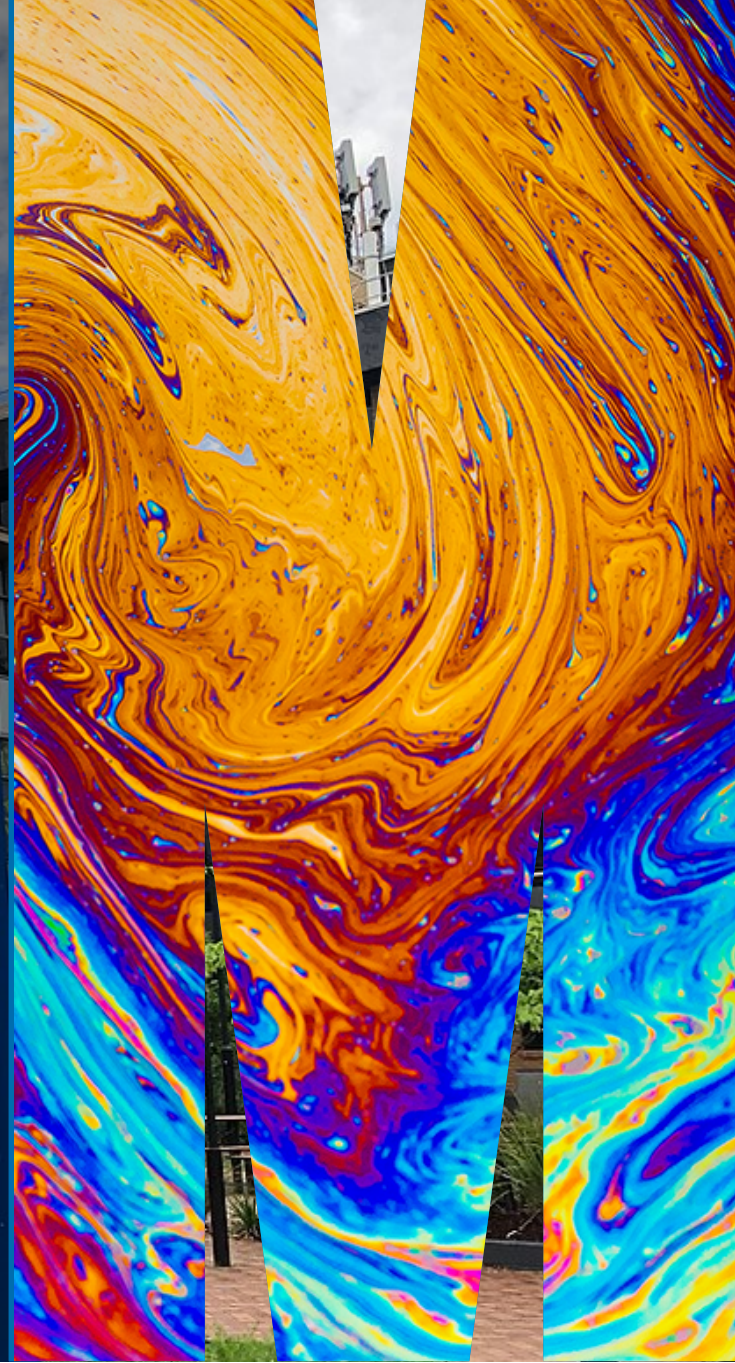
ETC5512: Wild Caught Data

Week 1

Data collection



Lecturer: *Didier Nibbering*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Start with a question?

Start with a question?

What questions do you have..?

- .. about a virus?
  - https://opendatahandbook.org/value-stories/en/open-sourcing-genomes/
- .. about bush fires and floods?
  - https://www.pmc.gov.au/public-data/open-data
- .. about saving the environment?
  - http://save-the-rain.com/SR2/#

Data examples in this unit

- Dr Nibbering:
  - Macroeconomic data
- Dr Menendez:
  - Great Barrier Reef data
- Dr Tanaka:
  - Australian census and election
  - International student assessment
- Professor Cook:
  - Airline traffic
  - Sports statistics

Macroeconomic data

- Macroeconomic data dominates the news
- Everyone affected by interest, exchange, and inflation rates
- Data helps voters and governments understand challenges

Great Barrier Reef data

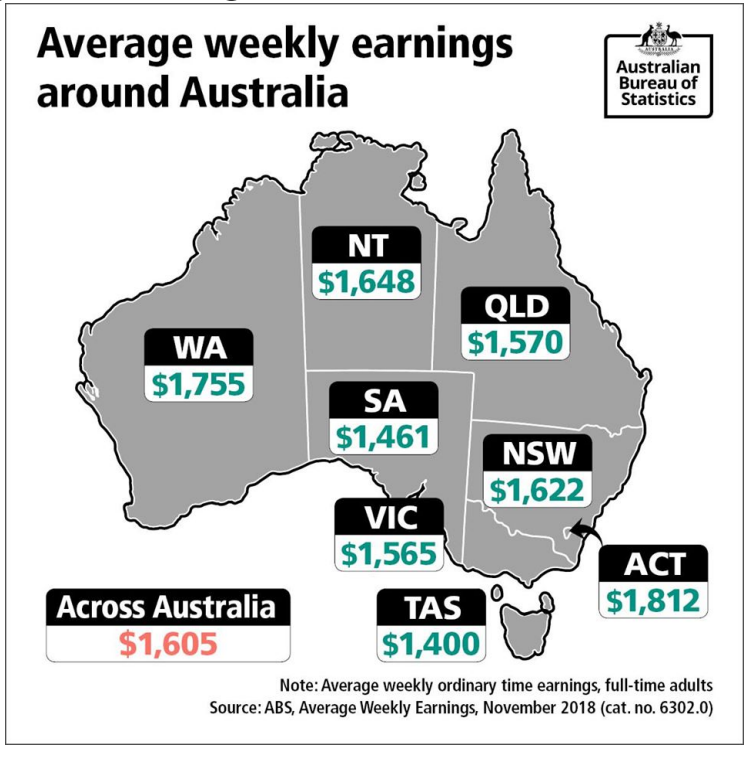
How do government organizations collect and use data?

- investigate the state of the Great Barrier Reef (GBR)
- data collected by the Australian Institute of Marine Science

Australian census and election

We'll delve into "fresh and local" government data to uncover insights about the Aussie demographic.

Why does ACT have the highest weekly earnings?



International student assessment

Girls consistently outperform boys in reading skills – but could this be changing?

February 12, 2020 10.44pm AEDT

Monkey Business Images/shutterstock

Email
Twitter 17
Facebook 55
LinkedIn
Print

Girls consistently outperform boys on reading tests – and have done so for several decades around the world. Lack of motivation, a weak vocabulary, poor reading engagement and lack of role models have all been considered possible reasons for this disparity.

But results from recent global student assessments, known as PISA, suggests the global reading gap is closing. But in most cases, not for the reasons teachers, parents, researchers, or governments might have hoped for. Indeed, it seems gaps closed overall not because boys are necessarily doing massively better, but because the performance of girls has declined.

PISA is the OECD's Programme for International Student Assessment and it tests the reading, mathematics, and science literacy skills of 15-year-old students across a variety of OECD and partner countries. The latest PISA from 2018, which was the seventh since 2000, focused on reading and included about 80 countries. So far three rounds of PISA have focused on reading –- 2000, 2009, and 2018 – giving researchers almost 20 years of trends to analyse.

Authors

Louis Volante
Professor, Faculty of Education, Brock University

Francesca Borgonovi
British Academy Global Professor, UCL

Disclosure statement

Louis Volante receives funding from the Social Sciences and Humanities Research Council of Canada (SSHRC).

Francesca Borgonovi is currently on leave from the OECD, which is responsible for the development of the PISA study. She received funding from the British Academy through its Global Professorship programme. The views expressed are purely those of the authors and may not in any circumstance be regarded as stating an official position of the OECD or the British Academy.

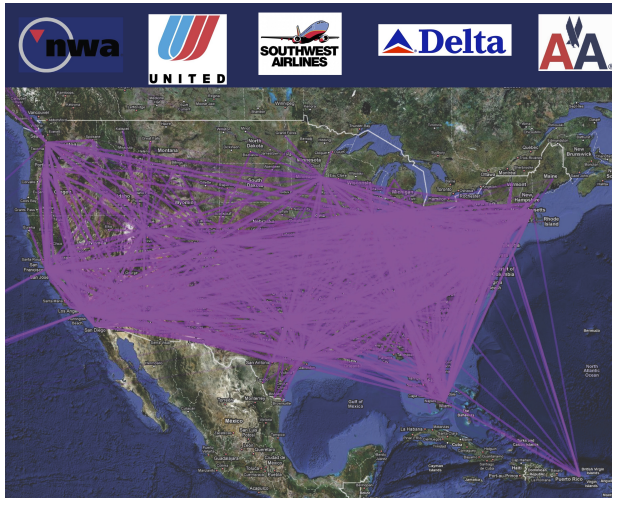
Partners

Source: The Conversation

US Airline traffic

From Professor Di Cook: *Sometimes I start with a data description, and from this questions are generated, and a workflow of operations on the data is designed to extract an answer to the question.* There is *really extensive* ✈️ information about every commercial flight that has flown in the USA since the early 1980s. For each flight the variables are scheduled departure time, actual departure time, carrier, plane id, origin, destination, departure delay, delay reason, ... Many, many questions...

- **What time of day is it more likely to see delays?**
- **What carriers have more efficient performance?**
- **Where my plane come from and go to next?**
- **If I have a choice of airports, which might present a lower risk of delay?**



Sports statistics

From Professor Di Cook: Sports statistics are readily available on many web sites. These can be extracted using web scraping tools. Primarily we come to sports with some idea about the game.

- Tennis:
  - What's the relationship between age and winning matches in grand slams?
  - Is it important to serve fast and hard in order to win matches?
- Cricket:
  - Which team has the best batting statistics?
  - Could we predict the team that will likely win the match?

Now that you have a question...

Data collection methods

- Investigate the relationship between variables
- Explanatory variables explain variation in response variable
- Collect observations on the variables

Data collection methods

- Observational data
  - No manipulation of the subjects' environment
  - Data are observed and collected on each subject

- Experimental data
  - Manipulate the subjects' environment
  - Then measure the response variable

Observational or experimental data?

Description 1:
The Academic Performance Index is computed for all California schools based on standardised testing of students. The data sets contain information and characteristics for 100 schools.

Description 2:
The response is the length of odontoblasts in 60 guinea pigs. Each animal received one of three dose levels of vitamin C by one of two delivery methods.

Description 3:
This data frame contains the responses of 237 Statistics I students at the University of Adelaide to a number of questions.

Observational data

Examples

- Surveys of households or firms
  - Who will win the US Presidential election?
- Government administrative data
  - Where can I find the best schools?
- Data from points of contact between transacting parties
  - Who are buying my products?

Observational data

Who will win the US Presidential election?

- Group of people we want information from
  - Population
- Group of people we get information from
  - Sample

Observational data

Percentage of votes for Republican candidate

- Population
  - Parameter
- Sample
  - Statistic

Observational data

How well represents the sample the population?

- Simple random sampling scheme
  - Every unit same sample probability
- Stratified multistage cluster sampling
  - Large-scale surveys as CPS and PSID

https://www.census.gov/programs-surveys/cps.html
https://psidonline.isr.umich.edu/

Observational data

- Stratified sampling
  - Nonoverlapping subpopulations that exhaust the population
  - States or provinces in a country
- Multistage sampling
  - Draw PSU at random from strata
  - Draw SSU at random from selected PSU
- Cluster sampling
  - Divide population into representative clusters
  - Select a cluster as your sample

Observational data

Different households have different sample probabilities

- Sampling weights
- Inversely proportional to sample probability
- Used for unbiased estimators population parameters

Observational data

Biased samples

- Exogenous sampling
  - Segmenting on socioeconomic factors
  - Biased if factors correlated with outcome
- Response-based sampling
  - Sample probability depends on response
  - Survey transport choice in sample of PT users
- Length-biased sampling
  - Sample the stock vs sample the flow
  - Longer duration of employment in stock sample

Observational data

Quality Survey data

- Nonresponse
- Missing data
- Mismeasured data
- Sample attrition

Observational data

Different formats

- Cross-section data
- Repeated cross-section data
  - Case-control studies
- Panel or longitudinal data
  - Cohort studies

Observational data

about student performance

Experimental data

Experimental data

- Vary causal variable of interest..
- while holding other covariates at controlled settings..
- to observe a response variable

Experimental data

- Treatment and control group
- Groups randomly selected
- Matching treatment and control groups

Experimental data

- Placebo effect
- Double-blind experiments
- Confounding variables

Experimental data

from lab experiments

Experimental data

Wild-caught experiments?

- Standard (laboratory) experiments
  - Willing recipients of randomly assigned treatment and passive administrators of a standard protocol
- Social experiments
  - human subjects and treatment administrators are active and forward looking individuals with personal preferences

Experimental data

Social experiments

- Health insurance with varying copayment rate
- Tax plans with alternative income guarantees
- Job search assistance programs

Experimental data

Limitations social experiments

- Cooperation participants
- Ethical objections
- Substitution bias
- Sample attrition
- Hawthorne effect

Social experiments

with job training

Experimental data

Natural experiments

- Subset of population is subjected to an exogenous variation in a variable, that would ordinarily be subject to endogenous variation

- Generate treatment and control groups in inexpensively and in real-world setting

Experimental data

Good natural experiments if

- Genuinely exogenous
- Impact sufficiently large
- Good treatment and control groups

Experimental data

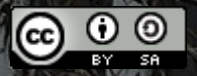
Natural experiments

- Administrative rules
- Unanticipated legislation
- Natural events

Natural experiments

with twins

That's it!



Lecturer: Didier Nibbering

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu