THE VALUE OF OFFICIAL STATISTICS AS A PUBLIC GOOD

OFFICIAL AND UNOFFICIAL STATISTICS FOR THE PUBLIC GOOD?

DAVID J. HAND

EUROPEAN STATISTICS DAY
20.10.2017
BETTER DATA.
BETTER LIVES.
Lisbon, PORTUGAL

Instituto Nacional de Estatística
Statistics Portugal

Sociedade Portuguesa de Estatística

ESAC
European Statistical Advisory Committee

- What are official statistics?
- What are official statistics for?
- What is the “public good”?
- How can the public good be measured?
- Traditional sources of data
- Data from non-official bodies?
- Examples – and their problems
- The risks – and the promise

What are official statistics?

"[statistics] relating to an authority or public body and its activities and responsibilities"

"[statistics] having the approval or authorization of an authority or public body"

UK's Statistics and Registration Service Act, 2007:

(a) statistics produced by—
- (i) the Board,
- (ii) a government department,
- (iii) the Scottish Administration,
- (iv) a Welsh ministerial authority,
- (v) a Northern Ireland department, or
- (vi) any other person acting on behalf of the Crown, and

(b) such other statistics as may be specified by order by—
- (i) a Minister of the Crown,
- (ii) the Scottish Ministers,
- (iii) the Welsh Ministers, or
- (iv) a Northern Ireland department.

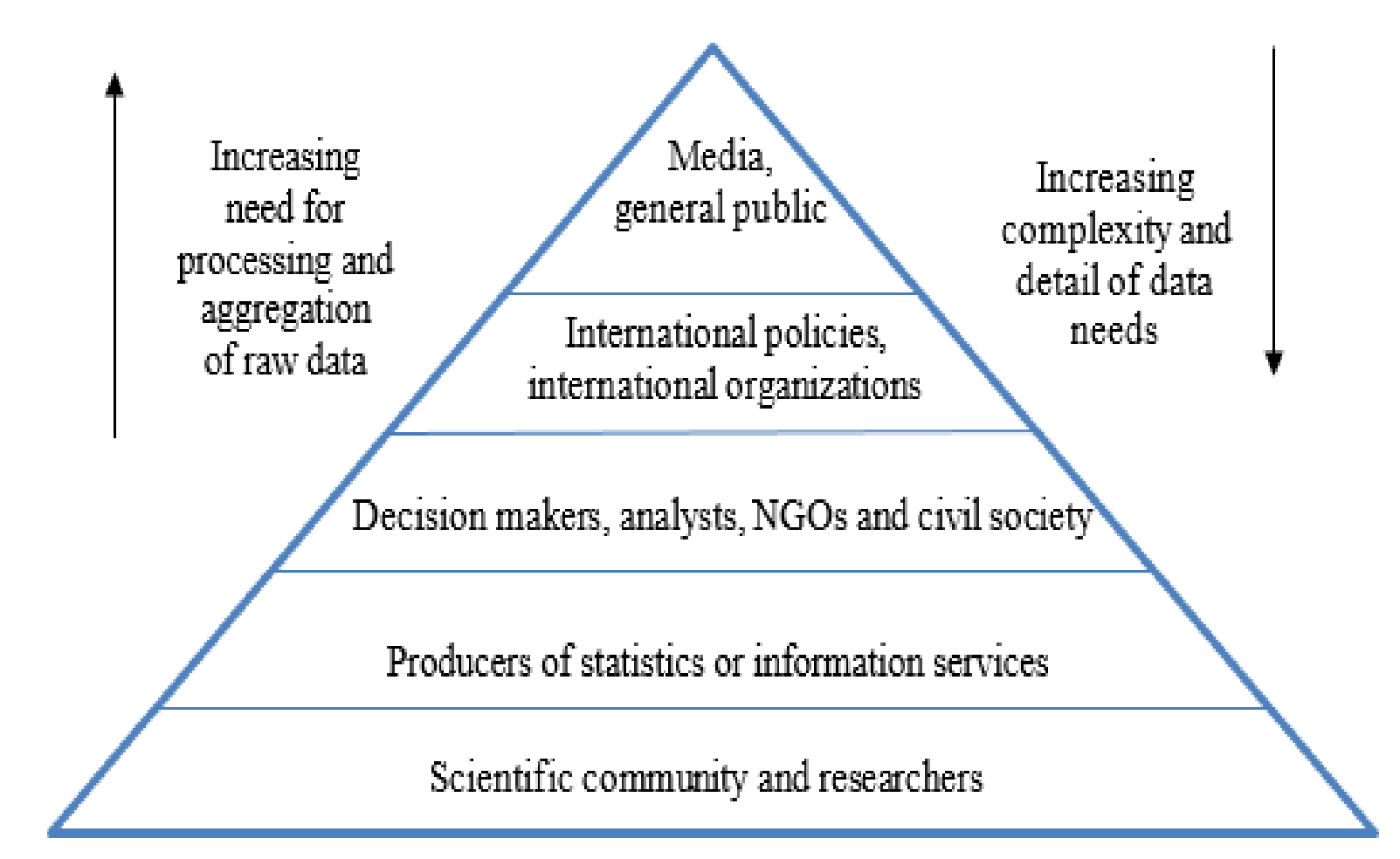
What are official statistics for?

UN Principle 1. Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation

This goes beyond government: ESAC has identified institutional and non-institutional users

The *Task Force on the Value of Official Statistics* identified:

1) Users with a general interest (e.g. citizens)
2) Users with a pre-defined interest (e.g. international organisations)
3) Users with a specific subject domain interest
4) Users with a reuse interest (e.g. organisations providing a service)
5) Users with a research interest



From: *Value of Official Statistics: Recommendations on Promoting, Measuring, and Communicating the Value of Official Statistics*, from the *Task Force on the Value of Official Statistics*, Conference of European Statisticians, June 2017

EUROPEAN STATISTICS DAY
20.10.2017
BETTER DATA. BETTER LIVES.
Lisbon, PORTUGAL

Official statistics are valuable if they promote *the public good*

But what is 'the public good'?

- helping the government
- helping individuals:
- helping business: where to locate new offices, investments,
- holding governments to account: are services effective
- shining a light
- accountability

EUROPEAN STATISTICS DAY
20.10.2017
BETTER DATA. BETTER LIVES.
Lisbon, PORTUGAL

The UKSA Code of Practice has 8 Principles:

- Meeting user needs
- Impartiality and objectivity
- Integrity
- Sound methods and assured quality
- Confidentiality
- Proportionate burden
- Resources (should be used efficiently and effectively)
- Frankness and accessibility

How can public good and value be measured?

The Task Force proposed a framework with three elements:

- observable ‘objective’ indicators such as number of downloads, citations (in all media), etc
- ‘subjective’ indicators from user satisfaction survey
- attempt to monetize value of official statistics

Should we also consider a counterfactual ?
What would be the situation if we did not have the statistic?

Quality is vital

relevance,
accuracy,
timeliness,
accessibility,
coherence

But note:

- what's good for one purpose may not be good for another
- may be good enough for today, but not tomorrow

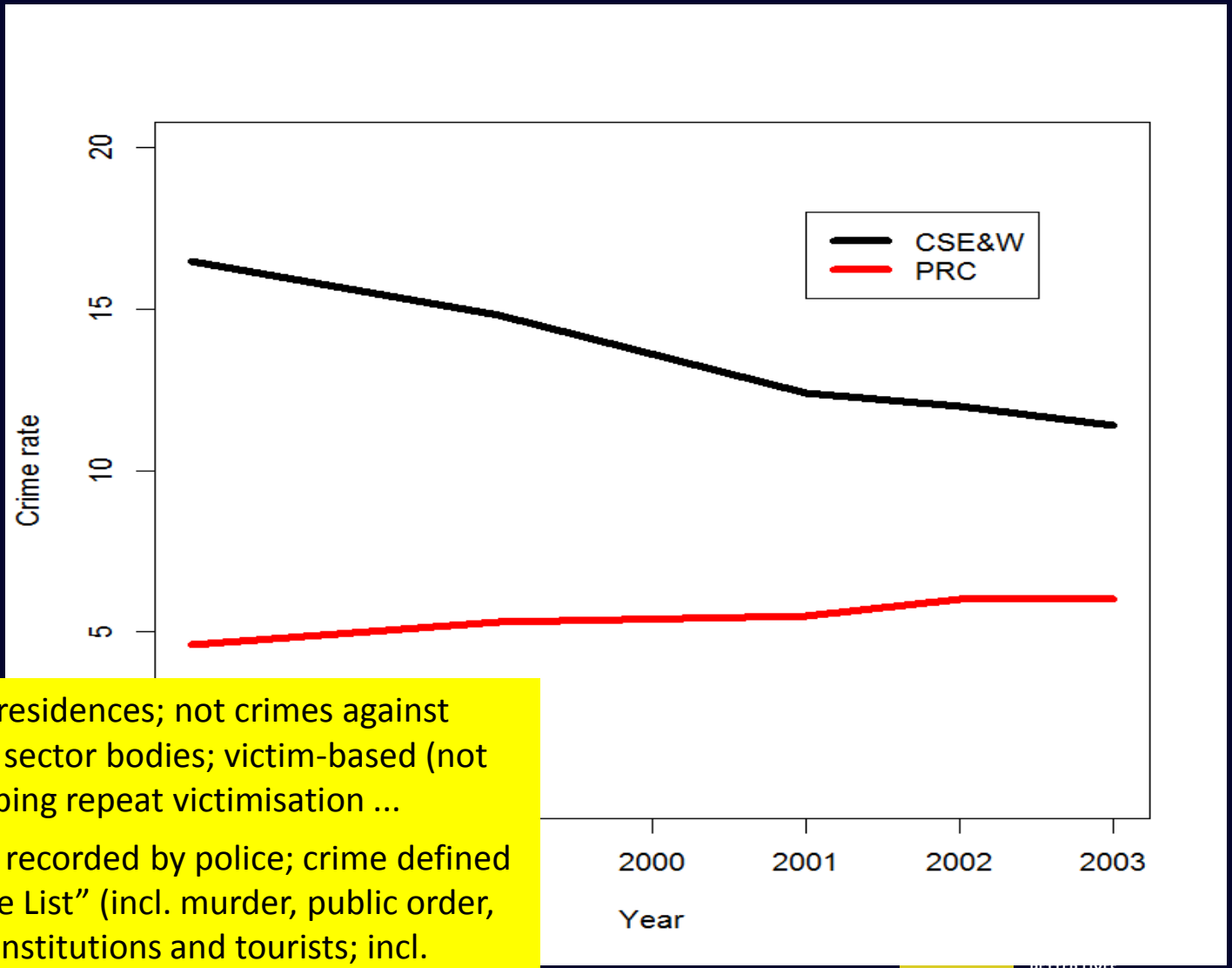
But official statistics are often criticised

- Timeliness
  - Time to collect data, e.g. inflation
    - UK: 700 goods and services from 20,000 locations
- Revision (e.g. GDP)
  - Because of new data arriving, e.g. GDP
    - Preliminary: 25 days after e/o Quarter; 44% of data
    - Second: 53 days; 83% of data
    - National Accounts: 85 days
    - Later revisions if statistical methods changed
- Inconsistency (e.g, crime statistics)

EUROPEAN STATISTICS DAY
20.10.2017
BETTER DATA. BETTER LIVES.
Lisbon, PORTUGAL

Crime rates, 1997-2003

Crime Survey England and Wales vs *Police Recorded Crime*



CSE&W: …not group residences; not crimes against commercial or public sector bodies; victim-based (not include murder); capping repeat victimisation ...

PRC: reported to and recorded by police; crime defined by "Notifiable Offence List" (incl. murder, public order, ...); incl. residents of institutions and tourists; incl. commercial bodies …

Independence is vital

Obviously if the data are distorted to present a positive image, they become useless as a measure

e.g. Commission for Health Improvement, 2003:

West Yorkshire Metropolitan Ambulance service had a substantial time lag between when a call was received and the time the clock to time ambulance response started

“In some cases, the clock appeared to start after the ambulance had actually arrived at the scene”

e.g. Campbell’s Law

"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

Argentina

National Statistics and Censuses Institute (INDEC) subject to political pressure from the government

"Private-sector economists and statistical offices of provincial governments show inflation two to three times higher than INDEC's number"

*The Economist, 2012**

Greece

Andreas Georgiou, former head of Greek national statistical authority, accused of distorting the figures despite Eurostat and the ASA saying they were accurate and "that they were produced and disseminated using appropriate processes and procedures based on European standards"

* *"The price of cooking the books". The Economist. 25 February 2012.*

EUROPEAN STATISTICS DAY
20.10.2017
BETTER DATA. BETTER LIVES.
Lisbon, PORTUGAL

Traditional sources of data

- Survey data

- Administrative data

But there is rapidly growing interest in the wider use of administrative data

ADVERTISEMENT

"Statistical challenges of administrative and transaction data"

Paper to be read at the RSS on Wednesday 15 November

Discussion contributions (in person or sent in) welcomed

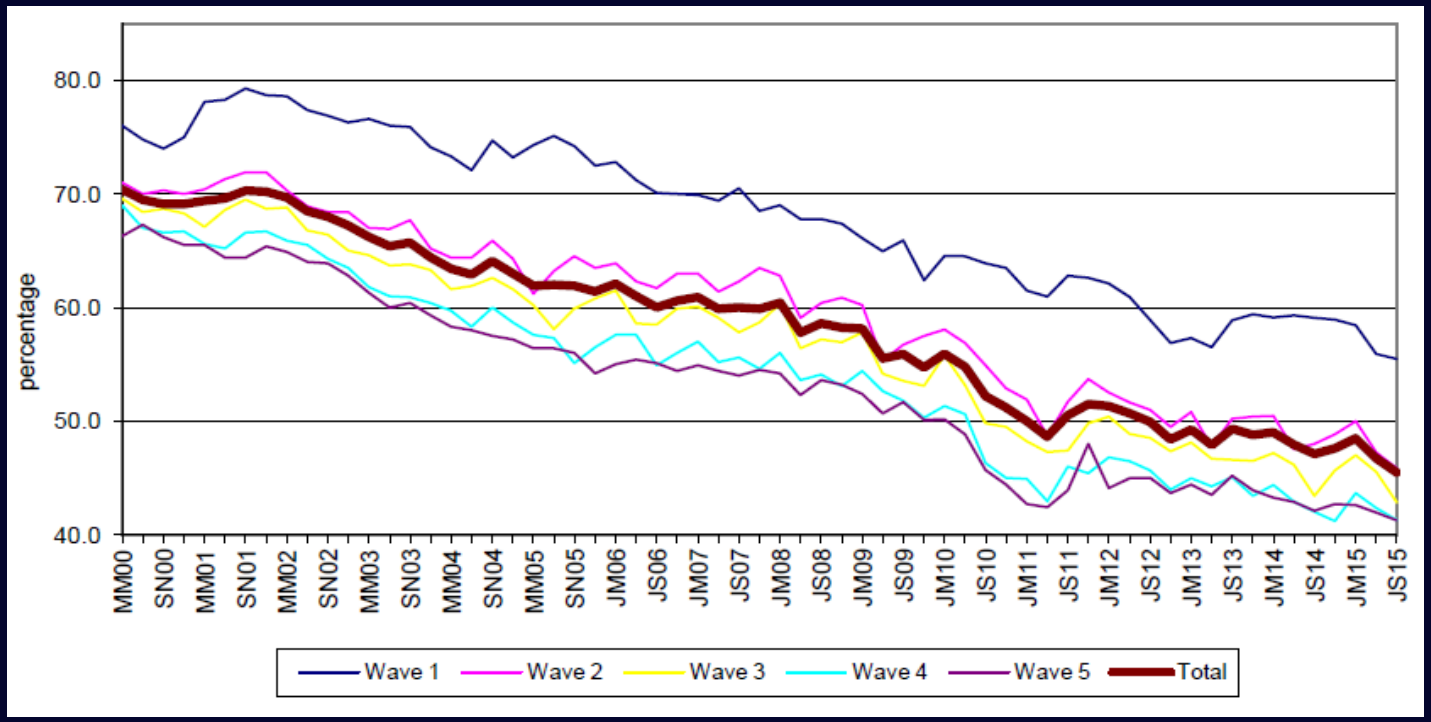
The paper can be downloaded from the RSS

http://www.rss.org.uk/RSS/Publications/Journals/Discussion_Papers/RSS/Publications/Journals_sub/discussion_papers.aspx?hkey=edfa4f9b-d001-49aa-bab9-e903b204e5d7

EUROPEAN STATISTICS DAY
20.10.2017
BETTER DATA. BETTER LIVES.
Lisbon, PORTUGAL

Survey data

- Well developed theory
- Uncertainty bounds (from sampling variation)
- *Concern about decreasing response rates*



UK's LFS quarterly survey wave-specific response rates:
March-May 2000 to July-Sept 2015

Source: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/labour-force-survey/index.html

Administrative data

Advantages of administrative data **???**

- *Free*, since already collected (for some other purpose)
- *All* the data are available
- *High quality*, since the effectiveness of the operation requires it
- *As up-to-date as it's possible to be*
- What people *do*, not what they *say they do*
- *Tighter definitions* than survey data

But are these claims justified?

- Free, since already collected (for some other purpose)
    - cost of cleaning, extracting, formatting, linking, …
    - free for the collecting organisation, but not others?
- All the data are available
    - not necessarily representative of the population to which one wishes to generalise
    - operational database will contain data relevant to the operation, not necessarily the question you want to answer
- High quality, since the effectiveness of the operation requires it
    - non-sampling distortions
    - not ideal for your question
- As up to date as it's possible to be
    - may not be instantly available to other organisations
- What people do, not what they say they do
    - age and gender of customer not relevant to transaction
- Tighter definitions than survey data
    - not necessarily tighter for your research questions

Recall the UK SARSA definition of official statistics as:
those produced by a collection of public (official) bodies

Which leads us to the question:
what about statistics which purport to do the same thing, but which are produced by other bodies?

"Eurostat supports the **modernisation of price statistics**, the aim being to ensure that **price collection methods** remain appropriate in a world of increasingly dynamic markets for consumer goods, dynamic pricing and ingenious ways of providing discounts."

Eurostat, 2017

Example 1: Supermarket scanner data for price indices

"A fifth of EU countries use [supermarket scanner] data already"

HICP, Practical Guide for Processing Supermarket Scanner Data, Eurostat, Sept 2017

- *The Netherlands CPI: 2002*
- *Norway CPI: 2005*

- Gives both prices and quantities
- Not merely a sample (by collector) of goods sold
- Extended periods
- Only items actually sold (not just on shelves)
- Includes real prices (discounts etc)
- Cheaper to collect
- Bigger data sets
- Churn is visible

But:

- Are the scanner data suppliers representative of the population you want to cover?
- Different types of outlet (even within one chain) may have different pricing strategies;
- Do the scanner data suppliers change internal strategies (special offers; 80/20 → product line rationalisation)?
- Will the supplier be in business in a year's time?
- Is there also a possible feedback/gaming danger?

NOTE:

Price indices are pragmatic measurements

The definition and the concept are two sides of the same coin

There is no “right” way of measuring inflation

Different methods have different properties and are suitable for different purposes

Example 2: Web scraping data for price indices

Many issues similar to scanner data

- Are the scanner data suppliers representative of the population you want to cover;
- Different types of outlet (even within one chain) may have different pricing strategies;
- Do the scanner data suppliers change internal strategies (special offers; 80/20 → product line rationalisation);
- Will the supplier be in business in a year's time?

+ others

- Length of time series
- Messy data
- Nothing on quantities sold
- Only web companies
- Search algorithms change

“The Billion prices project: using online prices for measurement and research”

Cavallo and Rigobon, 2016

- Half a million prices collected per day in US alone
- Compare US Bureau of Labor Statistics: 80,000 prices p/m
- 15 million products
- 900 retailers

Early days:

"These are early analyses using experimental techniques to help us develop our statistical methodology and are not comparable with headline estimate of inflation. We would strongly caution against their use in economic modelling and analysis use in economic modelling and analysis"

Research Indices Using Web Scraped Price Data, ONS 2016

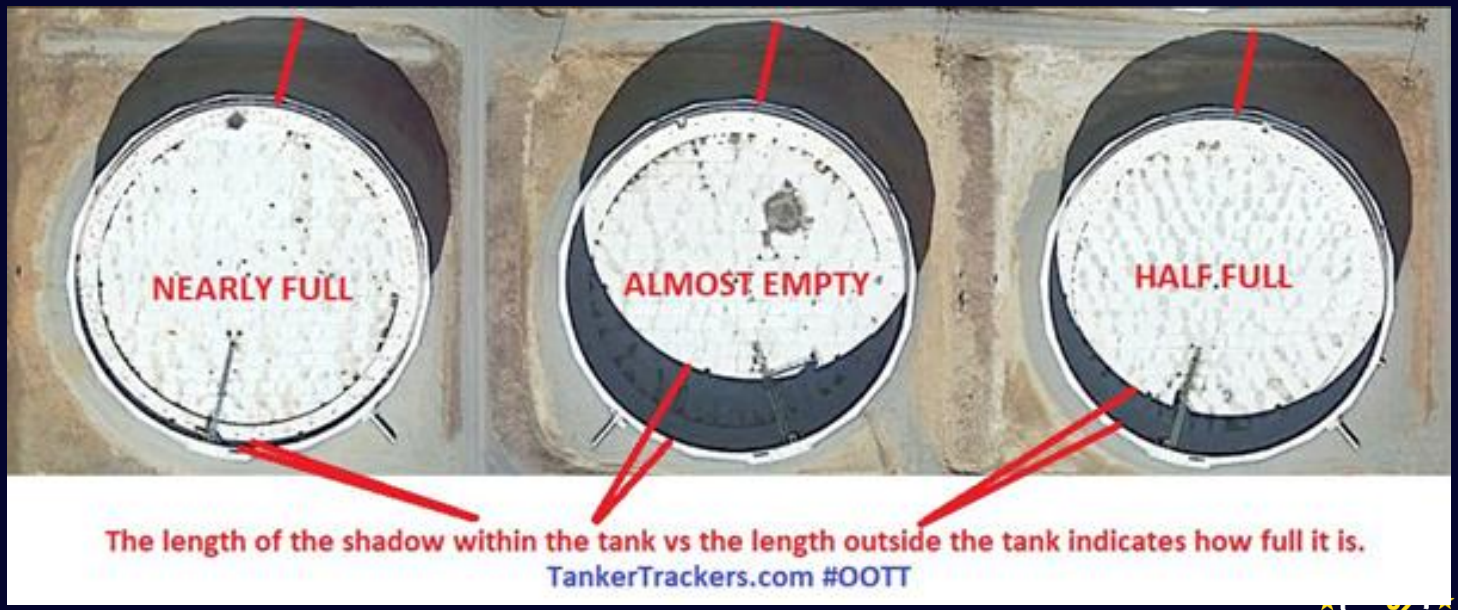
*"*A key challenge, addressed by our work, is that web-scraped price data are extremely messy and it is not obvious, *a priori,* how to reconcile them with standard CPI statistics*"*

Powell et al, 2017, Tracking and modelling prices using web-scraped price microdata

Example 3: Satellites

e,.g. Oil supplies

TankerTrackers.com



e.g. Crop yields

"Satellite data have repeatedly been shown to provide information that, by themselves or in combination with other data and models, can accurately measure crop yields in farmers' fields.

The resulting yield maps provide a unique opportunity to overcome both spatial and temporal scaling challenges and thus improve understanding of crop yield gaps."

Lobell, 2013

- Dirty data ?
- Missing data ?
- Elaborate measurement process

Example 4: Crowdsourcing

Merits:

- Costs
- Speed
- Quality
- Flexibility
- Scalability
- Diversity

….perhaps…..

- Winton *Hivemind*
- *Zooniverse*
- Amazon *Mechanical Turk*

  "a crowdsourcing Internet marketplace enabling individuals and businesses (known as Requesters) to coordinate the use of human intelligence to perform tasks that computers are currently unable to do."

For example:

- Data cleaning: deduplication, verification, …
- Creating training sets for machine learning systems
- Extracting information: from images, documents, …

Example 5: “Programmatic sampling”

e.g. Dalia Research

e.g. Qriously

“We intercept smartphone users while they use their usual apps by inviting them to take a short survey without leaving the app they are in. We do this by programmatically buying in-app ad space across 50,000 apps, showing an engaging question instead of a banner ad. Curious users who answer this initial question will be invited to answer more questions in a full-size survey interface.”

- Access to 60% of entire population in most major countries
- Dynamically selects a balanced demographic sample
- Weighting to match census data
- Filters for accidental clicks and suspicious answers

"At Qriously, we used our mobile polling methodology to run a number of polls on the EU referendum, and were the only ones to show a consistent lead for Leave before the day of the vote and to make an accurate outcome prediction when polls closed at 10pm"

Pollster	% voting leave
Qriously	55.8
Populus	45.0
ComRes	45.4
TNS	49.4
Opinium	49.4
YouGov	49.0
Ipsos MORI	48.0
Survation	48.4

- Uncertainty about the population being sampled
- Issues of privacy and confidentiality
- Issues of deliberately distorted data

e.g. one respondent who handed over the phone to his five-year-old daughter to answer the questions

Example 6: Telecoms data

94% of UK adults have mobile phones

- Location data
- Population flows
- Population densities at various places during the day
- Tourism data

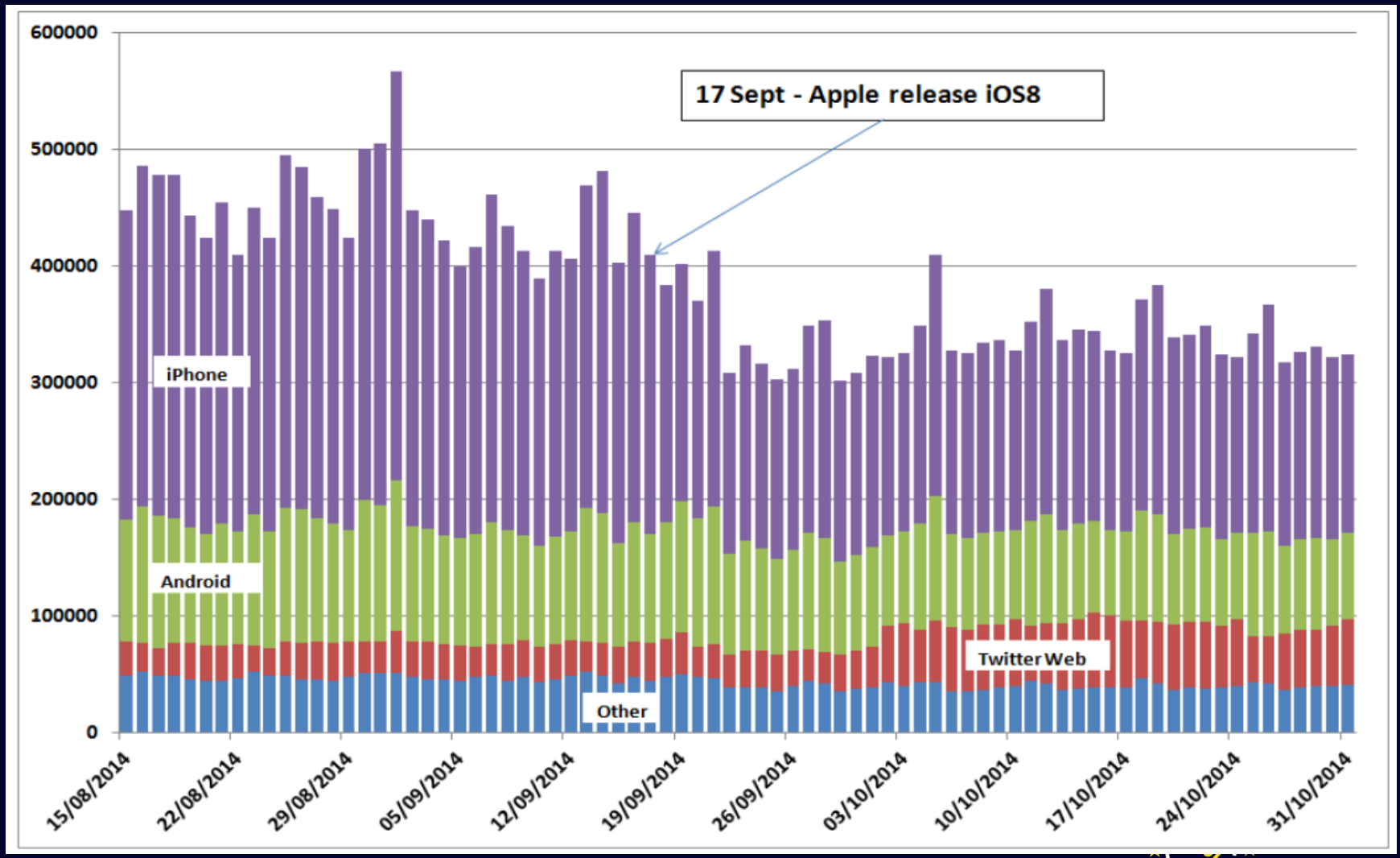
Do different service providers have different types of customer base with different behaviours?

Example 7: Twitter data

ONS Big Data project on using Twitter data to model internal national migration patterns

Bigdataproject2015qtr1progressreport-tcm77-407248.pdf

"In the second half of September 2014 there was a general drop in the daily volume of geo-located tweets from about 400,000 per day to about 300,000"



“Analysis by device type shows that this can almost entirely be explained by a sharp fall in the proportion of tweets from iPhone users

This can be explained by Apple’s release of the iOS8 operating system on 17 September 2014, which incorporated more flexible options for managing privacy with respect to location.”

Example 8: Smart meters

"A smart meter is a new kind of gas and electricity meter that can digitally send meter readings to your energy supplier for more accurate energy bills. Smart meters come with in home displays, so you can better understand your energy usage."

https://www.uswitch.com/gas-electricity/guides/smart-meters-explained/

- EU 2020 headline target of a 20% reduction in energy consumption
- Member states to ensure that at least 80% of consumers have smart meters by 2020 (wherever it is cost effective to do so)

Privacy issues

Selection bias issues for where installed

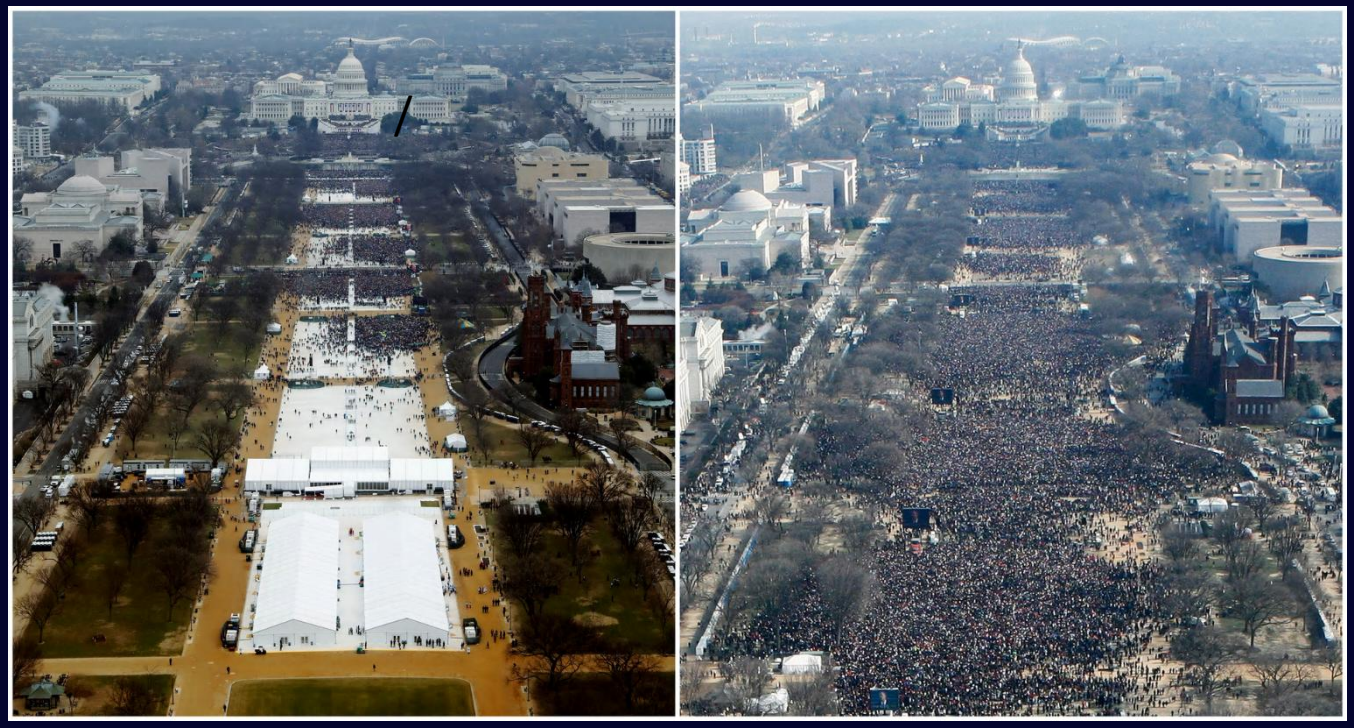
Summary of risks of new data sources:

- unknown provenance, opacity issues
- inconsistency over time
- concern about future access
- doubts about quality
- often not really free
- concerns about privacy and confidentiality
- conflicts of interest (recent examples where websites don't want to restrict postings)
- possible lack of impartiality: are websites publishers? (we know that newspapers will take attitudes which depend on the politics of the proprietors and editors)
- can we answer the question we want to?
- ownership of data
- discrimination and bias

And more?

Invented data

- Boris Johnson and his £350m
- Number at Trump inauguration



http://www.factcheck.org/2017/01/the-facts-on-crowd-size

Legalities

Should big data companies have to share data?

Regulation of big data companies – like utilities?

To conclude:

Exciting times
New opportunities

Value of new data sources for
- *Complementing* existing sources
- *Supplementing* existing sources
- *Triangulating* with existing sources

Much enthusiasm for how wonderful these new data sources are

But statisticians have a duty to tread carefully

EUROPEAN STATISTICS DAY
20.10.2017
BETTER DATA. BETTER LIVES.
Lisbon, PORTUGAL

thank you

EUROPEAN STATISTICS DAY
20.10.2017
BETTER DATA.
BETTER LIVES.
Lisbon, PORTUGAL