Demystifying big data in official statistics – it's not rocket science!

Jens Mehrhoff, Eurostat

9th Biennial IFC Conference

Basel, 30 – 31 August 2018

1. Definition of big data

- **Four possible interpretations of *big data*** – at least:
  - **'Data science'**: e.g. linking micro data
  - **New data sources**: e.g. Google or social media
  - **IT architecture**: e.g. distributed computing
  - **Large data sets**: e.g. granular/administrative data
- More often than not, ***big data* in official statistics are simply large data sets or the IT architecture handling them**.

European Commission

2. Use of big data in the production of official statistics

- **Case study: Electronic transactions data** ('scanner data') for measuring the average change in prices → large but structured data set
  1. **Classification of individual products into *homogeneous* groups**: supervised machine learning
  2. **Treatment of *re-launches***: probabilistic record linkage (fuzzy matching)
  3. **Index calculation**: multilateral methods (here: time-product dummy) – *time will not allow, please see*: https://www.youtube.com/watch?v=4zHpD5jzMMM

European Commission

2. Use of big data in the production

2.1 Classification of individual products

Example: Is a *yellow* and *firm* orange ripe?

Orange	Colour	Softness	Ripeness	Orange	Colour	Softness	Ripeness
1	Green	Firm	Unripe	**9**	Orange	Firm	Ripe
2	Green	Firm	Unripe	**10**	Orange	Firm	Ripe
3	Orange	Soft	Ripe	**11**	Orange	Soft	Unripe
4	Yellow	Firm	Unripe	**12**	Orange	Firm	Ripe
5	Yellow	Firm	Ripe	**13**	Green	Firm	Unripe
6	Orange	Soft	Ripe	**14**	Orange	Firm	Ripe
7	Green	Firm	Ripe	**(end of training data)**			
8	Yellow	Soft	Ripe	**15**	Yellow	Firm	**?**

European Commission

2. Use of big data in the production
2.1 Classification of individual products

- **Naïve Bayes classification:**

$$P(\text{ripe}|\text{yellow},\text{firm}) = \frac{P(\text{yellow},\text{firm}|\text{ripe}) \cdot P(\text{ripe})}{P(\text{yellow},\text{firm})}$$
$$= \frac{P(\text{yellow}|\text{ripe}) \cdot P(\text{firm}|\text{ripe}) \cdot P(\text{ripe})}{P(\text{yellow}) \cdot P(\text{firm})}$$

- Relies on the **assumption** that every feature being classified is **independent of all other features**.

European Commission

2. Use of big data in the production

2.1 Classification of individual products

Cross-tabulation of colour and ripeness

Colour	Ripe	Unripe	Total
Green			
Yellow	*P*(yellow|ripe)		*P*(yellow)
Orange			

NB: *P*(ripe) = proportion of ripe oranges (independent of colour and softness).

Cross-tabulation of softness and ripeness

Softness	Ripe	Unripe	Total
Soft			
Firm	*P*(firm|ripe)		*P*(firm)

European Commission

2. Use of big data in the production
2.1 Classification of individual products

Cross-tabulation of colour and ripeness

Colour	Ripe	Unripe	Total
Green	1/9	3/5	4/14
Yellow	***2/9***	1/5	***3/14***
Orange	6/9	1/5	7/14

NB: *P*(ripe) = ***9/14***.

Cross-tabulation of softness and ripeness

Softness	Ripe	Unripe	Total
Soft	3/9	1/5	4/14
Firm	***6/9***	4/5	***10/14***

European Commission

2. Use of big data in the production

2.1 Classification of individual products

- **Naïve Bayes classification:**

$$P(\text{ripe}|\text{yellow},\text{firm}) = \frac{P(\text{yellow}|\text{ripe}) \cdot P(\text{firm}|\text{ripe}) \cdot P(\text{ripe})}{P(\text{yellow}) \cdot P(\text{firm})}$$

$$= \frac{(2/9) \cdot (6/9) \cdot (9/14)}{(3/14) \cdot (10/14)}$$

$$= \frac{28}{45} = 0.62$$

European Commission

2. Use of big data in the production
2.1 Classification of individual products

- The **accuracy of supervised machine learning**, i.e. the proportion of automatically correctly classified products, is **around 80% for supermarket scanner data**. That means that **one out of five products is misclassified**.
- Hence, while machine learning can give **reasonable suggestions for the classification**, it eventually **needs to be assisted by human beings**; it is no panacea!

European Commission

2. Use of big data in the production
2.2 Treatment of re-launches

- **Re-launch**: A new attempt to sell a product or service, often by **advertising it in a different way or making it available in a different form**, e.g. different packaging → different GTIN.
- **Record linkage**: The task of **finding records** in a data set that **refer to the same entity** across entities that **may not share a common identifier**.
  - **Entity**: product or service; **Identifier**: GTIN ('barcode')

European Commission

2. Use of big data in the production
2.2 Treatment of re-launches

- **Levenshtein (1965) distance**: Minimum number of operations needed to **turn one string into another**.
  - **Operations**: insertion, deletion, or substitution of a character
- **Examples:**
  - 'car' → '**s**car' (**insertion** of 's' at the beginning)
  - '**s**can' → 'can' (**deletion** of 's' at the beginning)
  - 'sca**r**' → 'sca**n**' (**substitution** of 'r' for 'n')

European Commission

2. Use of big data in the production
2.2 Treatment of re-launches

Product description (or GTIN text)	Size of the string	Levenshtein distance	Levenshtein similarity[1]
'Whole Milk 1L' (*original*)	13	0	100%
'whole milk 1L'	13	2	85%
'whole milk 1 liter'	18	8	56%
'whole milk 1 litre'	18	8	56%
'Whole milk 1 ltr'	26	15	42%
'Whole Milk 2L'	13	1	92%
'1L Whole Milk'	13	6	54%

[1] Calculated as (1 – Levenshtein distance / length of the longer string) · 100%.

European Commission

2. Use of big data in the production
2.2 Treatment of re-launches

- The **last string** leads to horrible results because language allows us to **swap the order of words**.
  - There are still **plenty of other ways to improve**: capitalisation, trimming, character encoding, et cetera.
- However, **1 litre of milk is different from 2 litres**; while '1L', '1 liter', '1 litre', and '1 ltr' are all the same.
  - Hence, **do not trust the results blindly**! They would be the input into a user interface, for a **computer-assisted classification** – so use them as suggestions.

European Commission

3. Other potential uses of big data

- A recent survey by the Irving Fisher Committee on Central Bank Statistics (IFC) showed that there is **strong interest in big data in the central banking community**.
  (http://www.bis.org/ifc/publ/ifc-report-bigdata.pdf)
- The IFC Executive decided to select a **few case studies** for piloting the usefulness of big data:
  - **1. Administrative data; 2. Internet data; 3. Commercial data; 4. Financial market data**
- The **IFC / Bank Indonesia Satellite Seminar** to the ISI RSC 2017 explored the topic of big data from a central banking perspective (see *IFC Bulletin No 44*).
  (http://www.bis.org/ifc/publ/ifcb44.htm)

European Commission

4. Discussion and outlook

- The future direction, after the hype, is more like **big data will be supplementing rather than replacing official statistics**; a **genuine change in paradigm is rather doubtful** in the short to medium term.
- This has to been seen not least against the background of the **lower quality (keyword: coverage bias) of such *experimental* statistics**.
- Just one question: Will the lower production costs outweigh the potentially considerably higher **non-monetary costs of misguided policy decisions**? (Others include **governance and resource issues**.)

European Commission

Contact

JENS MEHRHOFF

European Commission
Directorate-General Eurostat
Price statistics. Purchasing power parities. Housing statistics

BECH A2/038
5, Rue Alphonse Weicker
L-2721 Luxembourg
+352 4301-31405
Jens.MEHRHOFF@ec.europa.eu

European Commission