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

Jens Mehrhoff, Eurostat

Second Statistics Conference
Banco Central de Chile

Santiago, 3 – 4 October 2017

Structure of the presentation

1. Definition of big data
2. Use of big data in the production of official statistics
3. Other potential uses of big data
4. Discussion and outlook

'But the "big data" that interests many companies is what we might call "found data", the digital exhaust of web searches, credit card payments and mobiles pinging the nearest phone mast.'
Tim Harford, Financial Times, 28 March 2014.

1. Definition of big data

- **Gartner's (2001) '3 Vs':**
  - **Volume**: amount of data ('found/organic' data)
  - **Velocity**: speed of data in and out (real time)
  - **Variety**: range of data types and sources ('data lake')
  - (**Veracity**: quality of data (inconsistency))
- **Big data** usually includes data sets with **sizes beyond the ability of commonly used software tools** to capture, curate, manage, and process data **within a tolerable elapsed time**.

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**.

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)

2. Use of big data in the production

2.1 Classification of individual products

- **Supervised learning**: The computer is presented with **example inputs and their desired outputs** and the goal is to **learn a general rule that maps inputs to outputs**.
  - **Classification**: identifying to which of a set of categories a new observation belongs, on the basis of a training set

- **Unsupervised learning**: **No labels are given** to the learning algorithm, leaving it on its own to **find structure in its input**.
  - **Clustering**: grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups

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	**?**

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**.

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)

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***

2. Use of big data in the production

2.1 Classification of individual products

- **Naïve Bayes classification:**

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

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

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

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!

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')

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')

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%.

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.

2. Use of big data in the production

2.3 Index calculation

- **Purpose**: measuring the **average rate of change in consumer prices** going from a base period 0 to a comparison period *t*
- **Bilateral**: $P[0,t] = \mathrm{P}[p(0),p(t),s(0),s(t)]$ – *biased*
- **Multilateral**: $P[0,t] = \mathrm{P}[p(0),\ldots,p(t),s(0),\ldots,s(t)]$
- **Time-product dummy**: $P[0,t] = \exp \delta^t$ from weighted regression with time and product dummies $\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t$

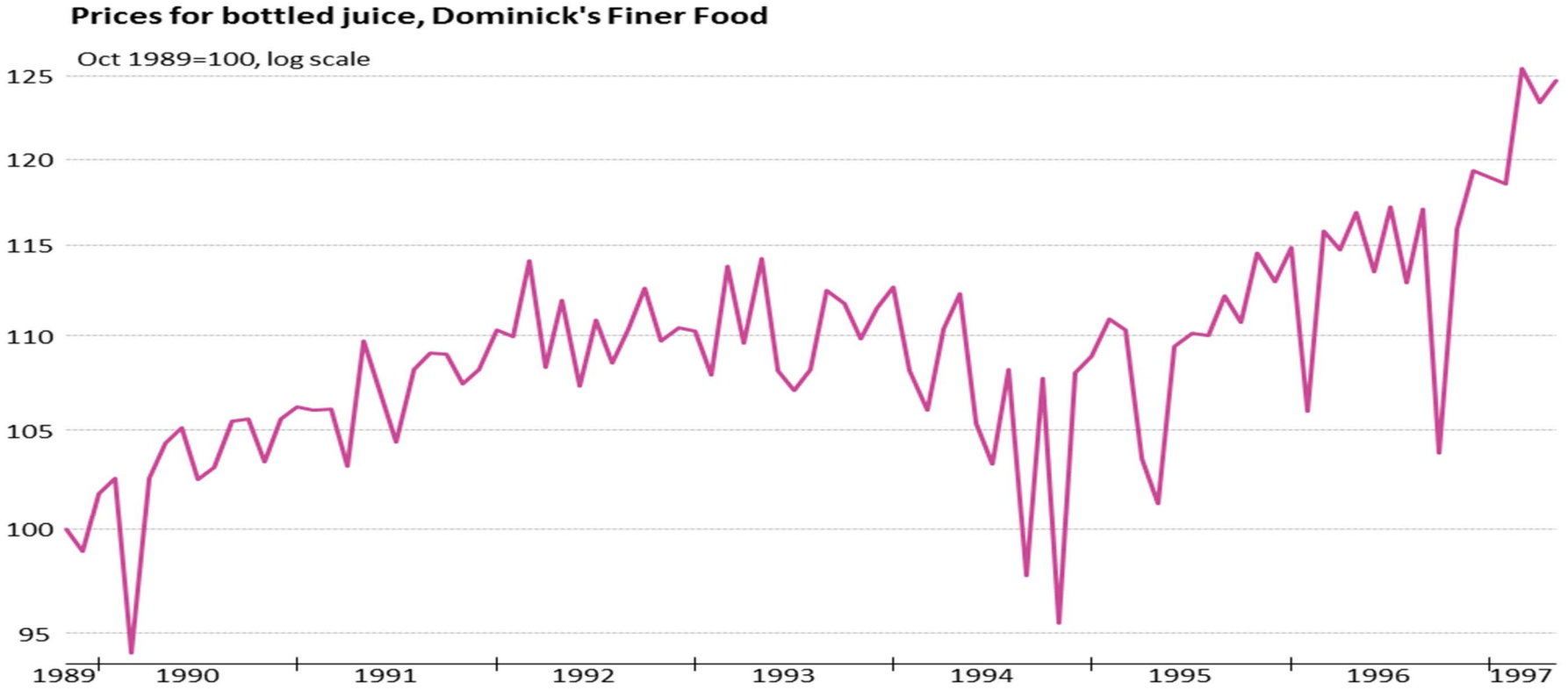
2. Use of big data in the production
2.3 Index calculation

- **Example: Dominick's Finer Food[1]** (now defunct Chicago-area grocery store chain) – electronic transactions data for 93 stores and 398 weeks from September 1989 to May 1997
  - 98 884 285 observations (after cleansing)
  - 13 845 products (excluding re-launches)
  - 29 categories (from analgesics to toothpastes)
- **Monthly aggregation at the category level.**

[1] James M. Kilts Center, University of Chicago Booth School of Business

2. Use of big data in the production

2.3 Index calculation



European Commission

2. Use of big data in the production

2.3 Index calculation

- **New and disappearing products**: 'dynamic' universe due to product churn
- **Just 5% of products are available in all 91 months**; 2½% are available in a single month, only; the average availability is 34 months.
- Within-category **average duration of products** ranges from 27 months (30%) for 'cigarettes' to **51 months (56%) for 'canned soup'**.

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)

3. Other potential uses of big data
3.1 Now/forecasting

- **Case study: now/forecasting consumer prices** using AMADEUS booking data for package holidays in Germany
- Package holidays have a **comparatively high weight** and prices show **pronounced volatility**.
  - **National sub-index contributes** up to a tenth of a percentage point **to euro area all-items inflation**.
  - On the other hand, analysis and forecasting are hampered by the **lack of a further breakdown**.

3. Other potential uses of big data

3.1 Now/forecasting

- While the **prices enter** the consumer price index only in the month **when the travel commences**, booking data contain **actual transactions for departures in the future**.
  - For example, **in March** already more than **half of the expected journeys in Summer** (July/August).
  - Enables **forecast to be based on hard data** rather than time series models alone.
  - Also **higher level of disaggregation** possible.

3. Other potential uses of big data

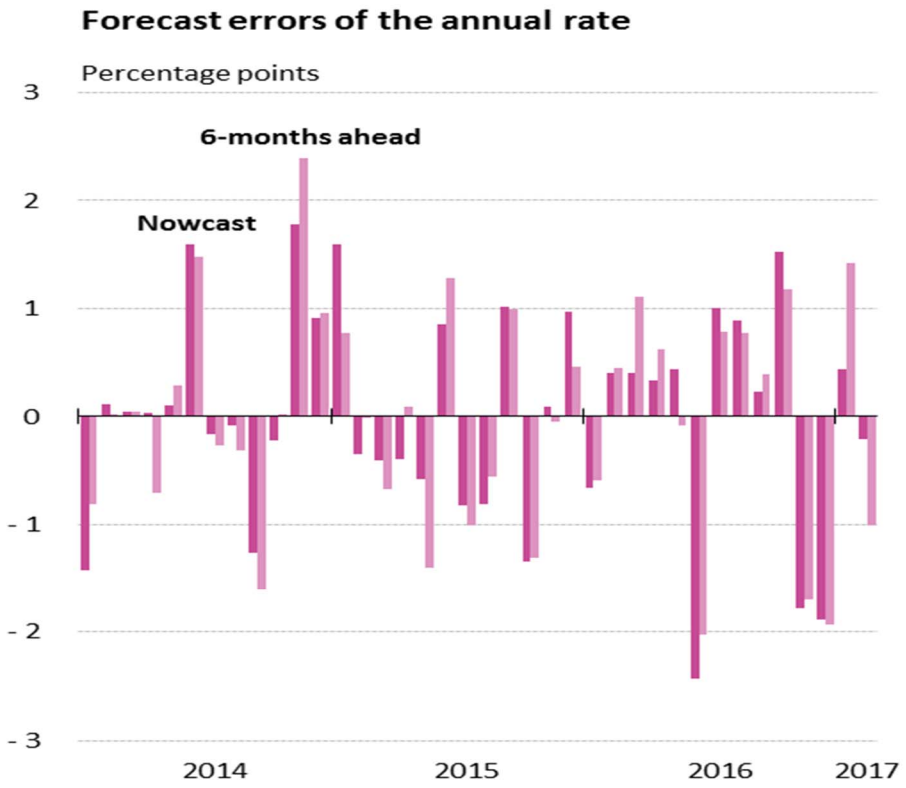
3.1 Now/forecasting

In-sample forecast evaluation (Jan 2014 – Feb 2017)

Book time, in months	0	1	2	3	4	5	6
Mean error, in percentage points	+0.01	+0.01	0.00	0.00	–0.01	–0.01	–0.01
Mean absolute error, in percentage points	0.82	0.81	0.78	0.76	0.85	0.78	0.83
Standard deviation of error, in percentage points	1.07	1.06	1.03	1.01	1.06	1.03	1.04
Reliability of sign, as a percentage	78.95	73.68	78.95	86.84	84.21	86.84	86.84
Signal-to-noise ratio (relative to mean absolute error)	2.97	2.99	3.12	3.18	2.85	3.13	2.92
Gain (relative to seasonal ARIMA model)	1.31	1.31	1.35	1.39	1.31	1.35	1.34

3. Other potential uses of big data

3.1 Now/forecasting



- According to the gain, the ***root mean square error*** **can be lowered** by 30 to 40%.
- Still, there are **some singular but large forecast errors** of 2 to 3 percentage points remaining.
- **Delineation of turning points versus outliers** proves to be particularly challenging in real time.

3. Other potential uses of big data

3.2 New/complementary indicators

- **Example: offer-based price index for housing**
- The following aspects are relevant for **real estate market analysis**:
  - **How long is an object offered?**
  - **How has the price changed over the course of time?**
- A **short marketing period and low price concessions** could be indicators for a **tense property market**.
- But it would also be important to know the **margin between the offer price and the transaction price**.
- ***This is another link of big data to digitalisation.***

4. Discussion and outlook

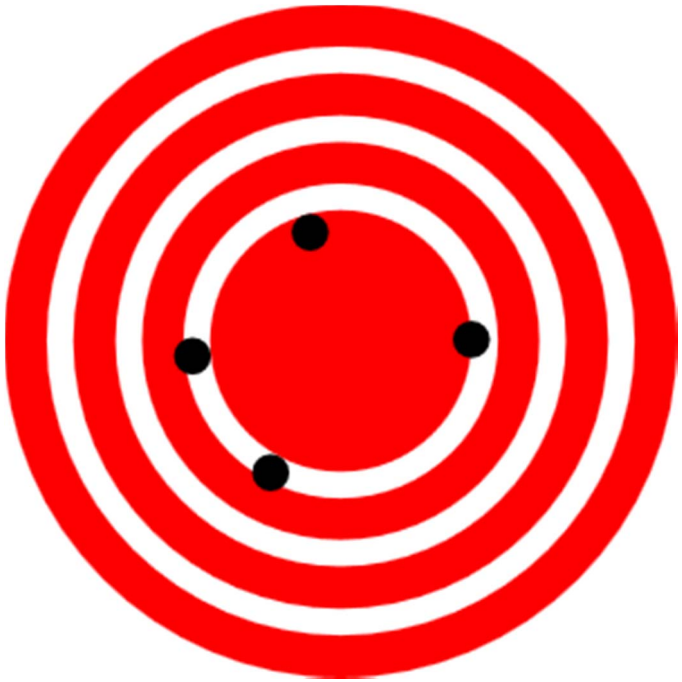
- The future direction, after the hype, is more like **big data will be supplementing rather than replacing official statistics**; a **genuine change in paradigm is rather doubtful** in the short to medium term.
- This has to been seen not least against the background of the **lower quality (keyword: 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**.)

4. Discussion and outlook

Source: Wikipedia

- **Big data** can be **very precise** – but at the same time may have **limited accuracy**.
- **Not more data are better, better data are better!**
- The paradox: the "bigger" the data, the surer we will **miss our target**.
  (Meng, 2016, *RSS Annual Conference*)

4. Discussion and outlook



Source: Wikipedia

- Big, or 'organic', data is **not capturing all behaviours in the society, just some**; and we might not know **which ones are missing**.
- The **combination of survey and census data with big data** is the ticket to the future. (Groves, 2016, *IARIW General Conference*)

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