Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

Demystifying big data in official statistics – it's not rocket science![1]

Jens Mehrhoff,
Eurostat

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

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

Jens Mehrhoff, Eurostat

Abstract

The talk will initially define big data and discuss the interpretation in the area of official statistics. We will then focus on the use of big data in the production of official statistics, referring to the case study of 'scanner data'. Simple classification rules and similarity measures are introduced, which help in grouping items together. An empirical example shows how a price index can be calculated from this new data source. At all stages of the presentation two things are key: demystifying machine learning and the like, while, at the same time, highlighting the limits of what is technically possible.

Keywords: big data; machine learning; classification; record linkage; scanner data

JEL classification: C43; C55; C81

Contents

1. Definition of big data

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. The '3 Vs' definition of big data according to Gartner (2001) comprises:

- Volume: amount of data ('found/organic' data);
- Velocity: speed of data in and out (real time); and
- Variety: range of data types and sources ('data lake').

But large data are not necessarily big data. The Square Kilometre Array (SKA) radio telescope, on the other hand, will ultimately be the largest scientific instrument on Earth, both in physical scale and in terms of volume of data it will generate. Just in its first phase, the telescope will produce some 160 terabytes of raw data per second – more than 4 times the 2016 global internet traffic. These data are no longer analysable by humans.

This is a massive amount of data. If a byte of data is equivalent to a small grain of rice, 160 terabytes per second would take

- half a second to cover Basel 10 cm high in rice,
- one day to cover the entire European Union 10 cm high in rice,
- three-and-a-half weeks to cover the whole surface of the Atlantic Ocean 10 cm high in rice, and
- two millennia to fill up the Atlantic Ocean with rice from its seabed to the surface.

More often than not, big data in official statistics are simply large data sets or the IT architecture handling them. There are, at least, four possible interpretations of big data in the area of official statistics:

- 'Data science': e.g. linking micro data;
- New data sources: e.g. Google or social media;
- IT architecture: e.g. distributed computing; and
- Large data sets: e.g. granular/administrative data.

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

As a case study for the use of big data in the production of official statistics we refer to electronic transactions data ('scanner data') for measuring the average change in prices. This is a large but structured data set. Simple classification rules and similarity measures are introduced, which help in grouping items together. An empirical example shows how a price index can be calculated from this new data source.

The process is split into three parts:

1. Classification of individual products into homogeneous groups via supervised machine learning;
2. Treatment of re-launches via probabilistic record linkage (fuzzy matching); and
3. Index calculation via multilateral methods (here: time-product dummy).

2.1 Classification of individual products

According to Mitchell (1997) 'a computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.' The task in our setting is the classification of individual products into homogeneous groups, the performance measure is accuracy, i.e. the proportion of automatically correctly classified products, and the experience is training data.

Supervised learning is where 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 is an example with the aim of identifying to which of a set of categories a new observation belongs, on the basis of a training set. In contrast to this is unsupervised learning, where no labels are given to the learning algorithm, leaving it on its own to find structure in its input. An application is clustering with the objective of 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.

Example: Is a *yellow* and *firm* orange ripe? Table 1

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	**?**

The naïve Bayes classifier relies on the assumption that every feature being classified is independent of all other features:

$$\begin{aligned} P(\text{ripe}|\text{yellow,firm}) &= \frac{P(\text{yellow,firm}|\text{ripe}) \cdot P(\text{ripe})}{\mathrm{P}(\text{yellow,firm})} \\ &= \frac{P(\text{yellow}|\text{ripe}) \cdot P(\text{firm}|\text{ripe}) \cdot P(\text{ripe})}{P(\text{yellow}) \cdot P(\text{firm})}. \end{aligned}$$

Unlike in Bayesian econometrics, the prior information, i.e. *P*(ripe), actually exists. Plugging in the numbers derived from a cross-tabulation of colour and softness with ripeness, one gets

$$
\begin{aligned}
P(\text{ripe}|\text{yellow,firm}) &= \frac{P(\text{yellow}|\text{ripe}) \cdot P(\text{firm}|\text{ripe}) \cdot P(\text{ripe})}{P(\text{yellow}) \cdot P(\text{firm})} \\
&= \frac{(2/9) \cdot (6/9) \cdot (9/14)}{(3/14) \cdot (10/14)} \\
&= \frac{28}{45} = 0.62.
\end{aligned}
$$

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!

2.2 Treatment of re-launches

A relaunch is 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 and different GTIN (Global Trade Item Number, 'barcode').

Record linkage has the task of finding records in a data set that refer to the same entity across entities that may not share a common identifier. In our case the entity is the product or service and the identifier is the GTIN.

The Levenshtein (1965) distance is defined as the minimum number of operations needed to turn one string into another. Operations are insertion, deletion, or substitution of a character:

- 'car' → '**s**car' (insertion of 's' at the beginning);
- '~~**s**~~can' → 'can' (deletion of 's' at the beginning); and
- 'sca~~**r**~~' → 'sca**n**' (substitution of 'r' for 'n').

Example: Which products are the same? Table 2

Product description (or GTIN text)	Size of the string	Levenshtein distance	Levenshtein similarity
'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%

Levenshtein similarity is calculated as (1 – Levenshtein distance / length of the longer string) · 100%.

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 (and look also at second and third best results).

2.3 Index calculation

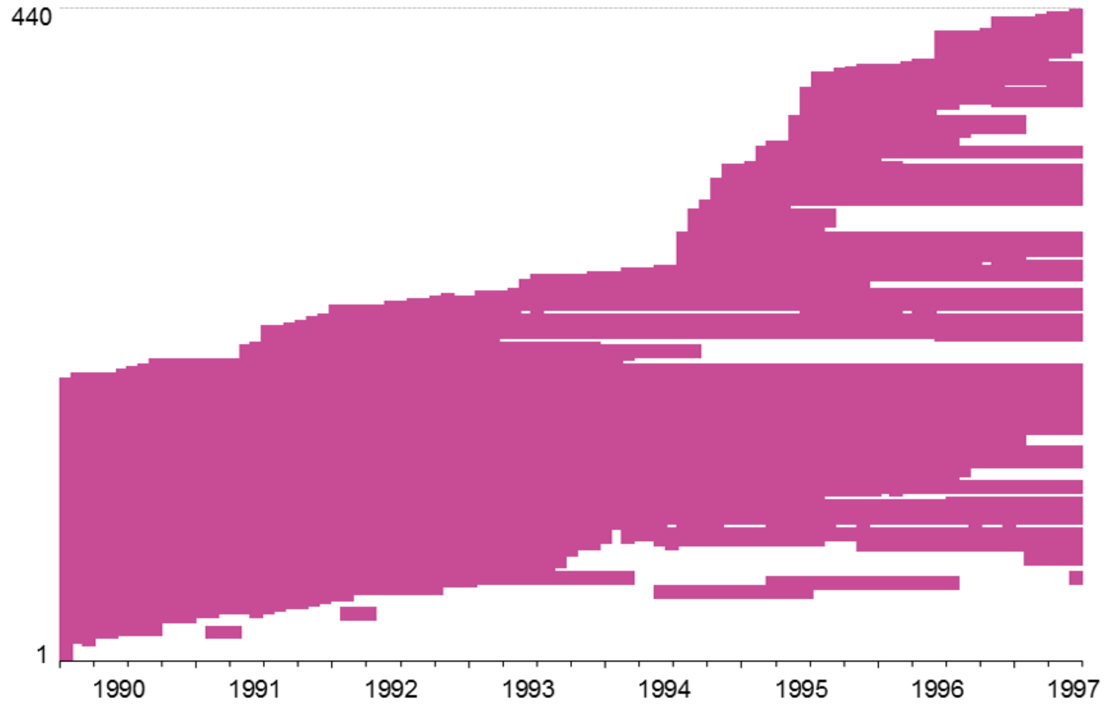
The purpose of a price index is measuring the average rate of change in consumer prices going from a base period 0 to a comparison period *t*. Traditional bilateral indices would be drifting when chain-linked due to effect of sale prices and stockpiling. Hence, multilateral indices are applied instead. We exemplify this by the time-product dummy (TPD) method, where the index P^t = exp δ^t ($P^{t=0}$ = 1) from an expenditure share-weighted regression of the logarithmic prices on time and product dummies:

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t.$$

The data are taken from the Dominick's database of the James M. Kilts Center at the University of Chicago Booth School of Business.[1] Dominick's Finer Foods was a Chicago-area grocery store chain and historic data are provided for academic research purposes. The data set covers 93 stores for 399 weeks from 14 September 1989 to 7 May 1997 and totals 98,884,285 observations (after cleansing) of 13,845 products (excluding re-launches) in 29 categories (from analgesics to toothpastes).

For the sake of exposition we aggregate the weekly store-level Universal Product Code (UPC, incorporated by GTIN) data to chain-wide item codes (tracking the same product across multiple UPCs) at monthly frequency using the month in which the week ends (weeks run from Thursday to Wednesday), covering the period from October 1989 to April 1997 (91 months; 467,605 observations of 13,818 products). Below we present the results for bottled juice as an example.

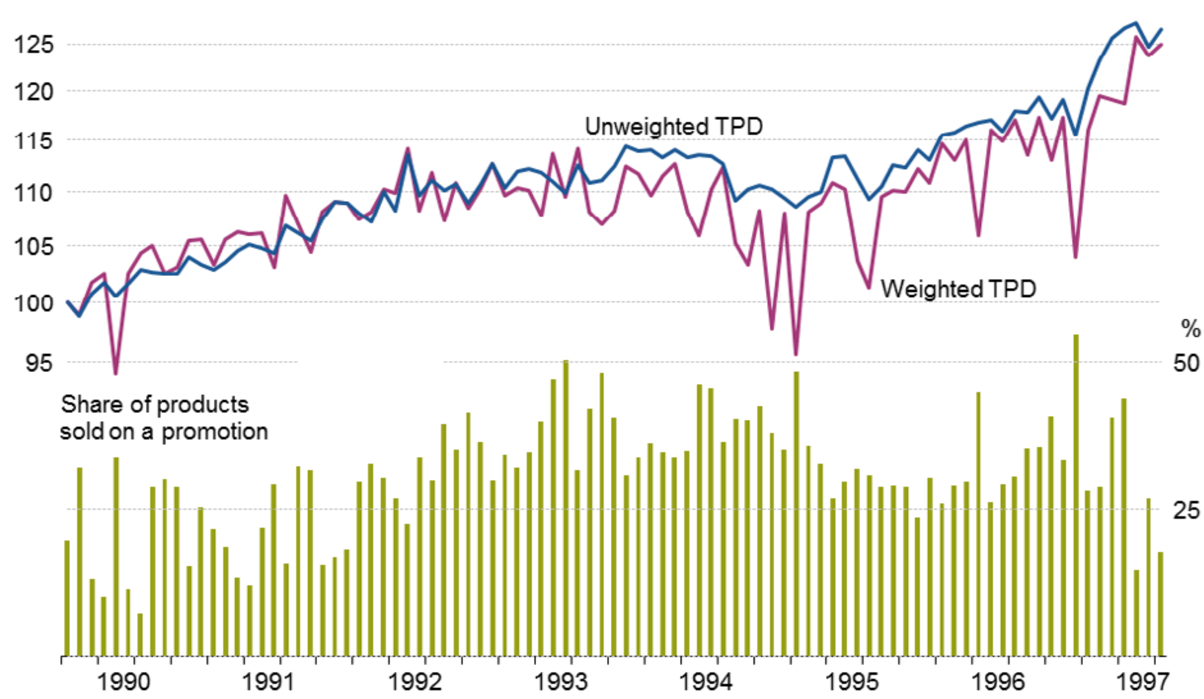
Product churn by item code, bottled juice



[1] Mehrhoff, J. (2018), *Promoting the use of a publically available scanner data set in price index research and for capacity building*, available from https://ec.europa.eu/eurostat/web/hicp/overview

The within-category average duration of bottled juice products in the sample is 37 months (average over all categories: 34 months). Notably, more than 10% (5%) of products are available throughout the 91 months studied. On the other hand, 3½% (2½%) of products are available in just one month. Compared to the first month, half of the initial products are still sold after 82 months.

Prices for bottled juice, Dominick's Finer Foods
Oct 1989 = 100, log scale



As it can be seen from a comparison of the weighted and the unweighted TPD index, where the latter is less affected by quantity increases due to price decreases – very much like web-scraped data –, the troughs are highly correlated to sale periods. But the stronger signal in the weighted TPD index comes at a cost: there is now also much more noise in the time series than in the unweighted version.

3. Discussion and outlook

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). 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).

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?

The Irving Fisher Committee on Central Bank Statistics revealed other issues including managing legal, financial and ethical risks (data governance), huge implications for information systems (IT resources), and that necessary skills may not be available in-house (staff resources). A fundamental change in institutions' business models, though, is not (yet) in sight.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

Demystifying big data in official statistics – it's not rocket science![1]

Jens Mehrhoff,
Eurostat

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

European
Commission

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**.

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

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:**

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

- 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:**

$$
\begin{aligned}
P(\text{ripe}|\text{yellow,firm}) &= \frac{P(\text{yellow}|\text{ripe}) \cdot P(\text{firm}|\text{ripe}) \cdot P(\text{ripe})}{P(\text{yellow}) \cdot P(\text{firm})} \\
&= \frac{(2/9) \cdot (6/9) \cdot (9/14)}{(3/14) \cdot (10/14)} \\
&= \frac{28}{45} = 0.62
\end{aligned}
$$

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.

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)

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**.)

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