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?

<table>
<thead>
<tr>
<th>Orange</th>
<th>Colour</th>
<th>Softness</th>
<th>Ripeness</th>
<th>Orange</th>
<th>Colour</th>
<th>Softness</th>
<th>Ripeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Green</td>
<td>Firm</td>
<td>Unripe</td>
<td>9</td>
<td>Orange</td>
<td>Firm</td>
<td>Ripe</td>
</tr>
<tr>
<td>2</td>
<td>Green</td>
<td>Firm</td>
<td>Unripe</td>
<td>10</td>
<td>Orange</td>
<td>Firm</td>
<td>Ripe</td>
</tr>
<tr>
<td>3</td>
<td>Orange</td>
<td>Soft</td>
<td>Ripe</td>
<td>11</td>
<td>Orange</td>
<td>Soft</td>
<td>Unripe</td>
</tr>
<tr>
<td>4</td>
<td>Yellow</td>
<td>Firm</td>
<td>Unripe</td>
<td>12</td>
<td>Orange</td>
<td>Firm</td>
<td>Ripe</td>
</tr>
<tr>
<td>5</td>
<td>Yellow</td>
<td>Firm</td>
<td>Ripe</td>
<td>13</td>
<td>Green</td>
<td>Firm</td>
<td>Unripe</td>
</tr>
<tr>
<td>6</td>
<td>Orange</td>
<td>Soft</td>
<td>Ripe</td>
<td>14</td>
<td>Orange</td>
<td>Firm</td>
<td>Ripe</td>
</tr>
<tr>
<td>7</td>
<td>Green</td>
<td>Firm</td>
<td>Ripe</td>
<td>(end of training data)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Yellow</td>
<td>Soft</td>
<td>Ripe</td>
<td>15</td>
<td>Yellow</td>
<td>Firm</td>
<td>?</td>
</tr>
</tbody>
</table>
2. Use of big data in the production
2.1 Classification of individual products

• **Naïve Bayes classification:**

\[
P(ripe|yellow,firm) = \frac{P(yellow,firm|ripe) \cdot P(ripe)}{P(yellow,firm)} = \frac{P(yellow|ripe) \cdot P(firm|ripe) \cdot P(ripe)}{P(yellow) \cdot P(firm)}
\]

• 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

<table>
<thead>
<tr>
<th>Colour</th>
<th>Ripe</th>
<th>Unripe</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yellow</td>
<td>(P(\text{yellow}</td>
<td>\text{ripe}))</td>
<td>(P(\text{yellow}))</td>
</tr>
<tr>
<td>Orange</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NB: \(P(\text{ripe})\) = proportion of ripe oranges (independent of colour and softness).

Cross-tabulation of softness and ripeness

<table>
<thead>
<tr>
<th>Softness</th>
<th>Ripe</th>
<th>Unripe</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft</td>
<td>(P(\text{firm}</td>
<td>\text{ripe}))</td>
<td></td>
</tr>
<tr>
<td>Firm</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. Use of big data in the production

2.1 Classification of individual products

### Cross-tabulation of colour and ripeness

<table>
<thead>
<tr>
<th>Colour</th>
<th>Ripe</th>
<th>Unripe</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>1/9</td>
<td>3/5</td>
<td>4/14</td>
</tr>
<tr>
<td>Yellow</td>
<td>2/9</td>
<td>1/5</td>
<td>3/14</td>
</tr>
<tr>
<td>Orange</td>
<td>6/9</td>
<td>1/5</td>
<td>7/14</td>
</tr>
</tbody>
</table>

NB: $P(\text{ripe}) = 9/14$.

### Cross-tabulation of softness and ripeness

<table>
<thead>
<tr>
<th>Softness</th>
<th>Ripe</th>
<th>Unripe</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft</td>
<td>3/9</td>
<td>1/5</td>
<td>4/14</td>
</tr>
<tr>
<td>Firm</td>
<td>6/9</td>
<td>4/5</td>
<td>10/14</td>
</tr>
</tbody>
</table>
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 processes

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' → 'scar' (**insertion** of 's' at the beginning)
  - 'scan' → 'can' (**deletion** of 's' at the beginning)
  - 'scar' → 'scan' (**substitution** of 'r' for 'n')
2. Use of big data in the production

2.2 Treatment of re-launches

<table>
<thead>
<tr>
<th>Product description (or GTIN text)</th>
<th>Size of the string</th>
<th>Levenshtein distance</th>
<th>Levenshtein similarity(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Whole Milk 1L' (original)</td>
<td>13</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>'whole milk 1L'</td>
<td>13</td>
<td>2</td>
<td>85%</td>
</tr>
<tr>
<td>'whole milk 1 liter'</td>
<td>18</td>
<td>8</td>
<td>56%</td>
</tr>
<tr>
<td>'whole milk 1 litre'</td>
<td>18</td>
<td>8</td>
<td>56%</td>
</tr>
<tr>
<td>'Whole milk 1 ltr'</td>
<td>26</td>
<td>15</td>
<td>42%</td>
</tr>
<tr>
<td>'Whole Milk 2L'</td>
<td>13</td>
<td>1</td>
<td>92%</td>
</tr>
<tr>
<td>'1L Whole Milk'</td>
<td>13</td>
<td>6</td>
<td>54%</td>
</tr>
</tbody>
</table>

\(^1\) Calculated as \((1 - \text{Levenshtein distance} / \text{length of the longer string}) \cdot 100\%\).
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