A "big data" gaze at why electronic transactions and web-scraped data are no panacea

Jens Mehrhoff, Eurostat

15th Meeting of the Ottawa Group

Eltville am Rhein, 10 – 12 May 2017

Structure of the presentation

1. The supposed population of transactions
2. Not more data are better, better data are better!
3. Electronic transactions and web-scraped data
4. Panacea's potion?: changes rather than levels
5. Are we impaled upon the horns of a dilemma?

"Is an 80% non-random sample 'better' than a 5% random sample in measurable terms? 90%? 95%? 99%?" (Wu, 2012)

1. The supposed population of transactions

- A (non-random) **sample of quotes** from abstracts for this meeting:
  - *"Scanner data have big advantages over survey data because such data contain transaction prices of **all items sold**…"*
  - *"…bilateral methods … do not capture the **full population dynamics** expressed by scanner data…"*
  - *"A further solution would be the use of transaction data (scanner data) to capture **all … prices** on the market."*
  - *"It is the first time that the evolution of … prices has been traced down using a dataset that covers the **population of transactions**…"*

European Commission

1. The supposed population of transactions

<table>
<tr><td rowspan="4">Transactions not recorded electronically</td><td>Electronic transactions data not available to NSIs</td></tr>
<tr><td>Available transactions data deleted by cleansing</td></tr>
<tr><td>Unmatched data not used in index calculation</td></tr>
<tr><td>Actual information exploited from "big data" sample</td></tr>
</table>

2. Not more data are better, better data are better!

- Let us consider a case where we have an **administrative record** covering f_a percent of the population, and a **simple random sample (SRS)** from the **same population** which only covers f_s percent, where $f_s \ll f_a$.
- How large should f_a / f_s be before an estimator from the **administrative record dominates** the corresponding one from the **SRS, say in terms of MSE**?

Source: Meng, X.L. (2016), "Statistical paradises and paradoxes in big data," *RSS Annual Conference*.

2. Not more data are better, better data are better!

- Our key interest here is to **compare the MSEs of two estimators** of the finite-sample population mean $\bar{X}_N$, namely,

$$\bar{x}_a = \frac{1}{n_a}\sum_{i=1}^{N} x_i R_i \text{ and } \bar{x}_s = \frac{1}{n_s}\sum_{i=1}^{N} x_i I_i\,,$$

  where we let $R_i = 1$ $(I_i = 1)$ whenever x_i is recorded (sampled) and zero otherwise, $i = 1, \ldots, N$.
- The **administrative record has no probabilistic mechanism** imposed by the data collector.

2. Not more data are better, better data are better!

- Expressing the **exact error**, where $f_a = n_a/N$:

$$\bar{x}_a - \bar{X}_N = \frac{\mathrm{E}[xR]}{\mathrm{E}[R]} - \mathrm{E}[x] = \frac{\mathrm{Cov}[x,R]}{\mathrm{E}[R]}$$

$$= \underbrace{\rho_{x,R}}_{\substack{\text{Data}\\ \text{Quality}}} \cdot \underbrace{\sigma_x}_{\substack{\text{Problem}\\ \text{Difficulty}}} \cdot \underbrace{\sqrt{\frac{1-f_a}{f_a}}}_{\substack{\text{Data}\\ \text{Quantity}}}.$$

- Given that $\bar{x}_s$ is **unbiased**, its MSE is the same as its variance.

European Commission

2. Not more data are better, better data are better!

- The **MSE** of $\bar{x}_a$ is more complicated, mostly because R_i depends on x_i:

$$\text{MSE}[\bar{x}_a] = \text{E}\left[\rho_{x,R}^2\right] \cdot \sigma_x^2 \cdot \left(\frac{1 - f_a}{f_a}\right).$$

- For **biased estimators** resulting from a large self-selected sample, the **MSE is dominated (and bounded below) by the squared bias term**, which is **controlled by the *relative* sample size** f_a.

2. Not more data are better, better data are better!

- To guarantee $\mathrm{MSE}[\bar{x}_a] \leq \mathrm{Var}[\bar{x}_s]$, **we must require** (ignoring the finite population correction $1 - f_s$)

$$f_a \geq \frac{n_s \rho_{x,R}^2}{1 + n_s \rho_{x,R}^2}, \text{ or equivalently}$$

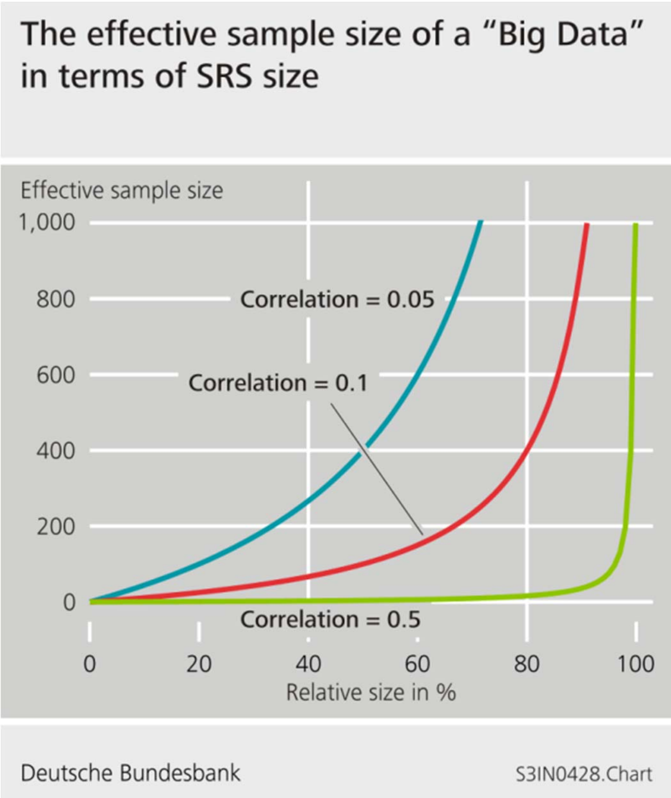
$$n_s \leq \left(\frac{f_a}{1 - f_a}\right) \frac{1}{\rho_{x,R}^2} = \left(\frac{n_a}{N - n_a}\right) \rho_{x,R}^{-2}.$$

- A **key message** here is that, as far as statistical inference goes, what makes a **"big data"** set big is typically **not its absolute size**, but its **relative size to its population**.

2. Not more data are better, better data are better!

- Therefore, the question **which data set one should trust more** is unanswerable without knowing N.
- But the general message is the same: when dealing with self-reported data sets, **do not be fooled by their apparent large sizes**.
- This reconfirms the **power of probabilistic sampling** and reminds us of the **danger in blindly trusting that "big data"** must give us better answers.
- **Lesson learned:** What matters **most is the quality**, not the quantity.

2. Not more data are better, better data are better!



- Imagine that we are given a **SRS** with $n_s = 400$:
  - If $\rho_{x,R} = 0.05$ and our intended **population is the USA**, then $N \approx 320{,}000{,}000$, and hence we will need $f_a = 50\%$ or $n_a \approx 160{,}000{,}000$ to place more trust in $\bar{x}_a$ than in $\bar{x}_s$.
  - If $\rho_{x,R} = 0.1$, we will need $f_a = 80\%$ or $n_a \approx 256{,}000{,}000$ to dominate $n_s = 400$.
  - If $\rho_{x,R} = 0.5$, we will need over 99% of the population to beat a SRS with $n_s = 400$.

3. Electronic transactions and web-scraped data

- What **price** would be most representative of the sales of the **same product** sold at a number of different prices for a month? The answer is the ***unit value*** (CPI Manual, 2004):

$$UV^t = \frac{\sum_{i=1}^{N} p_i^t q_i^t}{\sum_{i=1}^{N} q_i^t} = \frac{\mathrm{E}[p^t q^t]}{\mathrm{E}[q^t]}.$$

- **<u>Estimators</u>**
  - Electronic transactions data: $\widehat{UV}^t = \frac{\sum_{i=1}^{N} p_i^t q_i^t R_i}{\sum_{i=1}^{N} q_i^t R_i} = \frac{\mathrm{E}[p^t q^t R]}{\mathrm{E}[q^t R]}.$
  - Web-scraped data: $\widehat{UV}^t = \frac{\sum_{i=1}^{N} p_i^t R_i}{\sum_{i=1}^{N} R_i} = \frac{\mathrm{E}[p^t R]}{\mathrm{E}[R]}.$

3. Electronic transactions and web-scraped data

- <u>**Error of web-scraped data**</u>

$$\frac{\mathrm{E}[p^t R]}{\mathrm{E}[R]} - \frac{\mathrm{E}[p^t q^t]}{\mathrm{E}[q^t]} = \underbrace{\frac{\mathrm{Cov}[p^t, R]}{\mathrm{E}[R]}}_{\text{Systematic Undercoverage}} - \underbrace{\left(\frac{\mathrm{E}[p^t q^t]}{\mathrm{E}[q^t]} - \mathrm{E}[p^t]\right)}_{\text{Missing Quantities}}$$

- The **second term would not disappear** even when full population coverage could be achieved.

3. Electronic transactions and web-scraped data

- Since, caused by product substitution,

$$\frac{\mathrm{E}[p^t q^t]}{\mathrm{E}[q^t]} - \mathrm{E}[p^t] = \frac{\mathrm{Cov}[p^t, q^t]}{\mathrm{E}[q^t]} < 0,$$

  there are just two relevant cases to distinguish:

  1. Mainly the **upper end of the market** is covered, i.e. $\mathrm{Cov}[p^t, R] > 0$, and hence the **total error is necessarily positive** (albeit *a posteriori* to an unknown degree).
  2. Mainly **discounters and the like** are covered, i.e. $\mathrm{Cov}[p^t, R] < 0$, so that it is **no longer possible to guess at what the likely sign of the total error is**.

3. Electronic transactions and web-scraped data

- **Error of electronic transactions data**

$$\frac{\mathrm{E}[p^t q^t R]}{\mathrm{E}[q^t R]} - \frac{\mathrm{E}[p^t q^t]}{\mathrm{E}[q^t]} = \underbrace{\frac{\mathrm{Cov}[p^t q^t, R]}{\mathrm{E}[q^t R]}}_{\text{Turnover Undercoverage}} - \underbrace{\frac{\mathrm{Cov}[q^t, R]}{\mathrm{E}[q^t R]/UV^t}}_{\text{Quantity Undercoverage}}$$

- The error of electronic transactions data is **more complicated**.

3. Electronic transactions and web-scraped data

Sign of the total error	$\frac{\mathrm{Cov}[q^t, R]}{\mathrm{E}[q^t R]/UV^t} > 0$	$\frac{\mathrm{Cov}[q^t, R]}{\mathrm{E}[q^t R]/UV^t} < 0$
$\frac{\mathrm{Cov}[p^t q^t, R]}{\mathrm{E}[q^t R]} > 0$	Indefinite	Positive
$\frac{\mathrm{Cov}[p^t q^t, R]}{\mathrm{E}[q^t R]} < 0$	Negative	Indefinite

4. Panacea's potion?: changes rather than levels

- The **MSE** can be written as the **sum of the variance of the estimator and the squared bias** of the estimator:

$$\begin{aligned}&\mathrm{MSE}\left[\left(\widehat{UV}^{t}-\widehat{UV}^{t-1}\right)\right]\\&\quad=\mathrm{Var}\left[\left(\widehat{UV}^{t}-\widehat{UV}^{t-1}\right)\right]+\mathrm{Bias}^{2}\left[\left(\widehat{UV}^{t}-\widehat{UV}^{t-1}\right)\right]\\&\quad=\mathrm{MSE}\left[\widehat{UV}^{t}\right]+\mathrm{MSE}\left[\widehat{UV}^{t-1}\right]\\&\qquad-2\,\mathrm{Cov}\left[\widehat{UV}^{t},\widehat{UV}^{t-1}\right]-2\,\mathrm{Bias}\left[\widehat{UV}^{t}\right]\mathrm{Bias}\left[\widehat{UV}^{t-1}\right]\end{aligned}$$

- If $\widehat{UV}^{t}$ and $\widehat{UV}^{t-1}$ are **positively correlated** and their **bias is in the same direction**, the **total MSE *of the change* will be lower** than the sum of the MSEs.

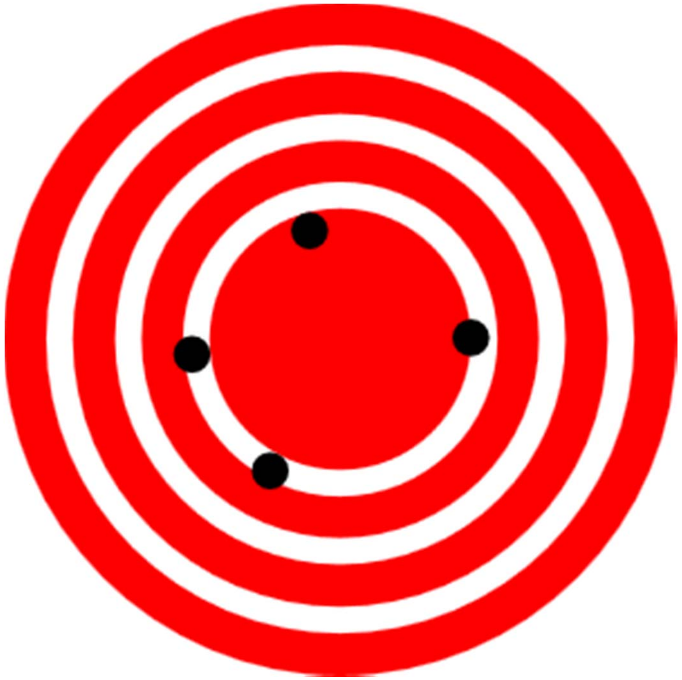
European Commission

5. Are we impaled upon the horns of a dilemma?

Source: Wikipedia.

- **Electronic transactions and web-scraped data** can be **very precise** – but at the same time may have **limited accuracy**.
- The paradox: the "bigger" the data, the surer we will **miss our target**!

5. Are we impaled upon the horns of a dilemma?



Source: Wikipedia.

- Price data from **traditional surveys** will not be collected perfectly in reality because of **non-probabilistic selection errors** as well.
- **The combination of survey data with "big data" is the ticket to the future.** (Groves, 2016, *IARIW General Conference*)

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