From price collection to price data analytics
Official Statistics production: Where we come from

The universe (entire statistical population)

The statistical data (sample)

The statistical model (to approximate the universe)

Amongst others: Quality control of data input

Official Statistics

=30%

=70%
Integration of large new data sources
no need for statistical models?
no need for theory?

The universe
(whole statistical population)

The statistical data („big data“)

The statistical model
(if necessary….?!)

Official Statistics

Amongst others:
Quality control of data input

=30%
=70%
Integration of large new data sources
no need for statistical models?
no need for theory?

The universe
(entire statistical population)

The statistical
data („big data“)

Amongst others:
Quality control of
data input

Official
Statistics

=30%

=70%
Integration of large new data sources

Quality control of scanner data and the web-scraped data → new measurement methods necessary

Is it relevant?
Is it accurate?
Is it complete?
## Relevance of scanner data

<table>
<thead>
<tr>
<th>Quality problem – Data Relevance</th>
<th>Measurement Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transaction data may contain transactions that are out of scope. -e.g. expenditures for business purposes (out of scope for consumer price indices)</td>
<td>Information by data providers; otherwise unresolved</td>
</tr>
</tbody>
</table>
Integration of large new data sources: Relevance

The statistical data (e.g. supermarket data food and non-food article)

Is it **relevant**?

- Large data-sources do not replace basic methodological work and checks concerning:
  - Coverage bias
  - Measurement error
  - Self selection bias

Large data sources do not make obsolete sound statistical models
## Relevance of web-scraped data

<table>
<thead>
<tr>
<th>Quality problem – Data Relevance</th>
<th>Measurement Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>are products offered online really sold and by whom?</td>
<td><strong>Information by data providers; otherwise unresolved</strong></td>
</tr>
</tbody>
</table>
## Accuracy of scanner data

<table>
<thead>
<tr>
<th>Quality problem – Data Accuracy</th>
<th>Measurement Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume and variety of data sets are too large to identify and clean erroneous/untrustworthy/inconsistent data sets with conventional methods.</td>
<td>Extent in % of erroneous/inconsistent data is monitored and excluded</td>
</tr>
</tbody>
</table>
# Accuracy of web-scraped data

<table>
<thead>
<tr>
<th>Quality problem – Data Accuracy</th>
<th>Measurement Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Website content may be IP-specific (a user who frequently checks a website or a web-scraper might lead to different price displays than first-time users)</td>
<td>Comparison of automatically and manually collected data</td>
</tr>
</tbody>
</table>
## Completeness of scanner data

<table>
<thead>
<tr>
<th>Quality problem – Data Completeness</th>
<th>Measurement Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume and variety of data sets are too large to identify missing values with conventional methods. (Scanner data: natural attrition of Unique identifiers is extremely high)</td>
<td>Number and level of target values are measured against historical values from previous deliveries</td>
</tr>
</tbody>
</table>
## Completeness of web-scraped data

<table>
<thead>
<tr>
<th>Quality problem – Data Completeness</th>
<th>Measurement Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Websites change frequently</td>
<td>Number and level of target values are measured against historical values from previous deliveries</td>
</tr>
<tr>
<td>Relevant variables and URLs might not be identified and scraped</td>
<td></td>
</tr>
</tbody>
</table>
Implementation of large new data sources: accuracy/completeness

The statistical data (estimate for Austrian retail market)
(e.g. supermarket scanner data for food and non-food)

Is it accurate?

<table>
<thead>
<tr>
<th>#</th>
<th>Shop ID</th>
<th>Art.-Code</th>
<th>Art. retailer classification</th>
<th>Product Description</th>
<th>Quantity sold</th>
<th>Sales in EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>212</td>
<td>1234</td>
<td>Soft drinks - cola</td>
<td>Cola, BrandX, 333ML</td>
<td>123</td>
<td>€129</td>
</tr>
<tr>
<td>2</td>
<td>212</td>
<td>1214</td>
<td>Soft drinks – cola</td>
<td>Cola, light, BrandY, L</td>
<td>255</td>
<td>€126</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>60.000.00</td>
<td>1234</td>
<td>9965</td>
<td>Bakery products</td>
<td>Brezel, brandZ, 500g</td>
<td>50</td>
<td>€126</td>
</tr>
</tbody>
</table>

60.000.000 data sets every month = 5.000 Articles X 4 Weeks X 1000 Shops X 3 Retailers

Before (with manual price collection):
10.000 data sets = 100 Articles X 1 (monthly collection) X 20 Cities X 5 supermarkets
Implementation of large new data sources: accuracy/completeness

The statistical data (e.g. supermarket data food and non-food article)

Is it **accurate?**

<table>
<thead>
<tr>
<th>#</th>
<th>Shop ID</th>
<th>Art.-Code</th>
<th>Art. retailer classification</th>
<th>Product Description</th>
<th>Quantity sold</th>
<th>Sales in EUR</th>
<th>Accurate &amp; complete?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>212</td>
<td>1234</td>
<td>Soft drinks cola</td>
<td>Cola, BrandX, 333ML</td>
<td>123</td>
<td>€129</td>
<td>YES</td>
</tr>
<tr>
<td>2</td>
<td>212</td>
<td>1214</td>
<td>Soft drinks cola</td>
<td>Cola, light, BrandY, L</td>
<td>255</td>
<td>€126</td>
<td>NO</td>
</tr>
</tbody>
</table>

Missing value for „Volume in Liter“

Large new data sources require automation of data cleaning and quality assessment processes
Implementation of large new data sources: accuracy/completeness

Analytical approach to quality control

1. Define measurable quality dimensions and elements of the data
2. Automate as many consistency and quality checks as possible

Examples:
- Extent in % of **erroneous** / inconsistent data is monitored and excluded
- Average # of missing values per data set
- Unreasonable changes of summary statistics
- Number and level of target values measured against historical values
- % of month to month attrition rates in product groups

3. Ability to adapt automated processes to ever-changing data structures and sources
Implementation of large new data sources: accuracy/completeness

3. Adapt automated processes to changing data structures and sources

IT
- integrates
- develops/writes programs
- executes
- updates

CPI experts
- imputes
- analyzes
- deletes
- interprets
- cleans

From price collection to price data analytics - Josef Auer and Ingolf Boettcher – Statistics Austria
Implementation of large new data sources: accuracy/completeness

3. Adapt automated processes to changing data structures and sources = Data science

„Data science“ (in price statistics)→integrate, clean, analyze and process continuously changing (non-standardized) large price data sources and turn them into compliant price statistics
Implementation of large new data sources:

3. Adapt automated price index compilation processes to changing data structures and sources = Data science

| Examples |
|-------------------|-------------------|
| **Scanner data**  | **Web-scraping**   |
| -retailer continuously update data-base structures to own data-warehouse needs | -frequently changing web-site architecture and product presentation |
| -high attrition rate of single articles, shops, product classes | -high attrition rate of single articles and categories |
# Price index compilation with scanner data

## new working steps

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Article identification and matching</td>
<td>Automated matching</td>
<td>Manual matching</td>
</tr>
<tr>
<td>2. Plauibility check / filter / imputation</td>
<td>Deletetion of implausible data sets</td>
<td>Sampling / Imputation</td>
</tr>
<tr>
<td>3. Index compilation</td>
<td>Geomean of sampled price relatives</td>
<td>Retailer Weighted aggregation indices</td>
</tr>
</tbody>
</table>
Price index compilation with scanner data
new strata

From price collection to price data analytics - Josef Auer and Ingolf Boettcher – Statistics Austria
Price index compilation with scanner data

1. Article Identification, matching and mapping

2. Plausi etc.

3. (1) HVPI Flash-Estimate + Plausi

3. (2) H/VPI Compilation + Plausi

H/VPI Publication
From price collection to price data analytics

Contact:

Josef Auer
josef.auer@statistik.gv.at

Ingolf Boettcher
ingolf.boettcher@statistik.gv.at