Office for National Statistics
Swyddfa Ystadegau Gwladol

Istat

Instituto Nacional de Estatística
Statistics Portugal

eurostat

cbs Statistics Netherlands

statistics estonia

economie

DESTATIS
wissen.nutzen.

STATISTIKOS DEPARTAMENTAS
PRIE LIETUVOS RESPUBLIKOS VYRIAUSYBĖS

Checking the Usefulness and Initial Quality of Administrative Data

Frank Verschaeren
Statistics Belgium

Overview

- Administrative data in business statistics
- ESSnet Admin Data
- Is There a Problem ?
- Getting Useful Data
- Initial Checking of the Data
- Ways to resolve Data Issues
- Next steps

Administrative data

- Increasingly used in production of Official Statistics
  - due to considerations of cost and burden
  - better data processing and storage capacity
- Historically used less within business statistics

- Guidance and best practice are lacking
  - ESSnet AdminData established

ESSnet Admin Data

- Part of the Modernisation of European Enterprise and Trade Statistics (MEETS) programme
- Runs Sept. 2009 – Aug. 2013
- 8 European Member State NSIs working in collaboration across 9 work packages
- Information Centre:

http://essnet.admindata.eu/

"… to share best practices and make recommendations in the uses of administrative and accounts data in the production of business statistics."

ESSnet Admin Data

WP2 started in 2010

a: Checklist to assist Member States when investigating the usefulness of administrative data

b: Checklist for the quality of administrative data inputs

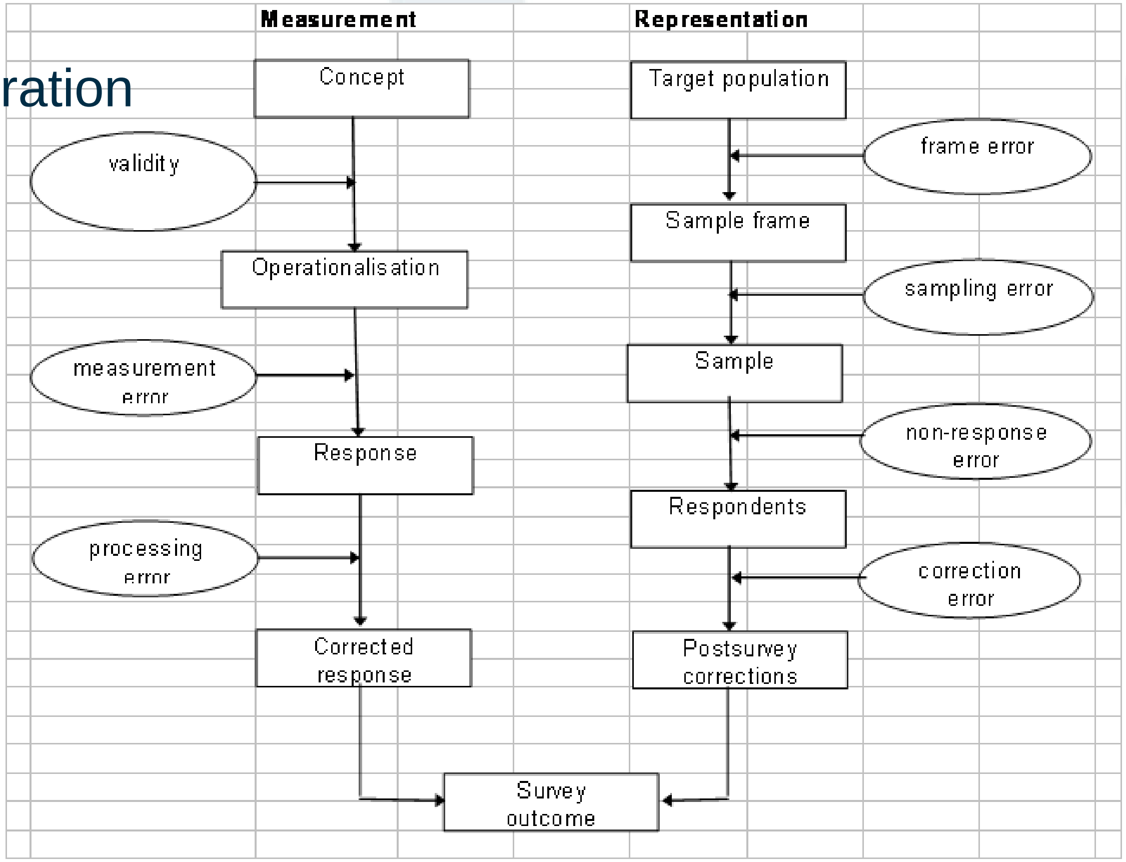
Is There a Problem ?

Similarities with primary data collection

= Same Sources of Error as with survey data

Life cycle of a survey

→ life cycle of a registration



Is There a Problem ?

Differences:

- Administrations have different needs
  - Concepts and target population not tailored to NSI’s needs
  - Continuity can be threatened by administrative changes
  - Administrative purpose, individual consequences for respondent
  - Some variables get more attention than others
- External collection and processing procedures
  - Often specific data correction procedures and limited editing
  - Inaccuracies: data or metadata wrong, or both?
- In many cases not possible to re-contact the respondent
  - Standard methods for checking / correcting do not always apply
- Administrative datasets are usually very large
  - Searching large datasets is computationally very demanding

Getting Useful Data

Most important question:

Whether or not to use the data at all

Pre-evaluation:

Checklist to consider relevant issues:

- Provides a structured way of looking
- Assures that attention is paid to important preconditions
- Limited effort compared to exploring actual data
- Can show where further clarification needed

Getting Useful Data

getting the data

Not only content is important, also delivery related aspects

Is a good working relationship with administrative data holders enough?

Something to say on:

- How data are transferred to the NSI (standardisation)
- Content and layout of administrative forms
- Possible changes in the future

Next steps:

- WP2 collects country experiences
- recommendations on the basis of these experiences

Initial Checking of the Data

Current situation: no common approach to checking

WP2’s objective: present good practices

→ Reference document:

- Techniques for checking the data
- Guidelines
- Generic enough for use in different contexts

→ Contribute to efficient, transparent and harmonised use of administrative data

Initial Checking of the Data

Two different ways of looking at the data

A) Starting from domain expert knowledge

Understand what’s behind the data, to detect problems

3 domains: VAT – Employment - Accounts

B) Starting from the data

Reverse-engineering the metadata and test the data against the ‘real’ data rules

Data profiling and monitoring

Initial Checking of the Data

A. Domain expert knowledge: detecting suspicious values

Method 1 – Quartile distances in industry Turnover

(Hoogland and Van Haren) extreme values in the distribution of VAT Turnover within a particular industry and size class.

If Turnover > Q3 + [C × (Q3 – Median)] Or Turnover < Q1 – [C × (Median – Q1)]

parameter C is derived from analysing past data

Method 2 – Period on period ratios

(De Jong) period on period ratios for each business based on the contribution that business’s Turnover makes to its (*industry and size*) class.

Score= turnover / median turnover

$$\text{TestRatio} = \begin{cases} \text{Score}_t / \text{Score}_{t-1} & \text{if } \text{Score}_t > \text{Score}_{t-1} \\ \text{Score}_{t-1} / \text{Score}_t & \text{otherwise.} \end{cases}$$

Predefined threshold for testratio

Initial Checking of the Data

A. Domain expert knowledge: detecting suspicious values

Method 3 – Comparison with reporting history for the business *(Hoogland and Van Haren, Lorenz…)* compares the current VAT Turnover figure with previous historic values for the same business.

If Turnover > C

And Turnover > 10 × mean Turnover for the business in the past 24 months

parameter C is e.g. € 100 million: focus on large reported turnover

Method 4 – Quartile differences combined with measure of influence (inspired by *Hoogland)* where influence is proportion of Turnover to total Turnover in its class.

Influence = Turnover / Total turnover

If Turnover > Q3 + [C × (Q3 – Median)] Or Turnover < Q1 – [C × (Median – Q1)]

First step: quartile difference, then eliminate low influence

Initial Checking of the Data

A. Domain expert knowledge: detecting suspicious values

Method 5 – Hidiroglou-Berthelot ratio between the Turnover value for a business in the current period and the Turnover value for the same business in the previous period.

Ratio r = Current Turnover / Previous Turnover

Transform:

If r < median then t = r – median / r else t = r – median / median

Define:

$E = t \times \max(\text{current Turnover}, \text{previous Turnover})^{V}$

The parameter V can take any value between 0 and 1. A value of 0 results in every business having the same importance

$d_{Q1} = \max\left[(Q2 - Q1), \left|A \times Q2\right|\right]$ $d_{Q3} = \max\left[(Q3 - Q2), \left|A \times Q2\right|\right]$

Where A is usually 0.05

Suspicious businesses are defined:

If $E < Q2 - C \times d_{Q1}$ Or $E > Q2 + C \times d_{Q3}$

Initial Checking of the Data

A. Domain expert knowledge : detecting suspicious values

Also methods for detecting:

suspicious patterns

unit errors

in reported VAT Turnover.

The five methods described above were tested on two years of UK VAT data

Estimated false hit rate: proportion of businesses identified as suspicious, but which had VAT Turnover similar to reported survey Turnover. The survey Turnover came from the Annual Business Survey

Other work = Methods for dealing with errors in VAT Turnover

Initial Checking of the Data

B. Starting from the data

Data profiling and monitoring

- Describe existing techniques for discovering:
  - invalid values
  - invalid combinations of values
  - Unreasonable results (outliers)
- Collect Country experiences: case description
- Make recommendations on:

how to - avoid duplication of checks and files

- streamline and automate checks

Ways to resolve Data Issues

Current situation: editing/imputation = starting point

- OK: describe and compare methods
- + More comprehensive approach possible?
  - detect quality issues
  - asses impact
  - investigate root cause
  - develop/implement remedies
  - monitor results

→ Better to eliminate the cause than to address symptoms

Next steps

- Work to further develop description and comparison of checks
- Get input from NSI’s across Europe
- Test and improve checklist

experience with checking administrative data?

willing to share information?

contact me!

Thank you for your attention

Contact:

Frank.Verschaeren@economie.fgov.be

More information:

http://essnet.admindata.eu/