Office for National Statistics

Data Sources for the SBR

Andrew Allen

Overview

- Main Administrative Sources of SBR data
- Statistical Sources
- Combining Sources
- Potential New Sources

Three types of source

- Administrative - collected for non-statistical purposes, such as taxation and regulation.
- Statistical – economic surveys, profiling, SBR improvement surveys
- Other Sources – private suppliers, directories, internet etc

Commonly used administrative sources

- Business Registration/License to trade
- Tax – VAT, employee income tax etc
- Trade Associations - chamber of commerce
- Social security – related to payment of staff.
- Government units
- Utility registers
- Industry Specific lists from regulators
- Published accounts

Advantages of administrative data

- Coverage - very important to get all economically active units on the SBR
- Cost - usually free or charge, or just marginal cost of system extract
- No additional response burden on business
- Timeliness - regular updates

Disadvantages of administrative data

- Gaining legal Access
- Units are not statistical - collected for a different purpose.
- Classification systems different
- Timeliness and lags in processing
- Changes to regulations and procedures

Administrative data quality

- Need to build knowledge of source and assess quality
- Regular monitoring of quality
- Need to have procedures for dealing with source conflict

Legal Issues

- A legal framework allowing access to sources will usually be required
- Formal agreement on delivery flows and security
- Building and maintaining a good working relationship with suppliers

Practical use of administrative data

- Keep administrative and statistical data separate
- Establishing unique identifiers
- Be aware of thresholds
- Handling changes to data sources

Practical issues

- Transforming to Statistical unit: most administrative data are legal units.
- These need to be transformed to enterprise i.e. smallest combination of legal units with autonomy.
- Transformation algorithms will differ by country. Profiling can be used for largest businesses.

Statistical Sources

- **Economic Census**
- Still used in many countries - but expensive, and intercensal updating required.

- Also issues with non-recognisable business places e.g. within residential unit.

Feedback from enterprise/establishment surveys

- Vital for feedback on address changes, deaths, classification changes etc.

- But has serious limitations:
- Does not find new births
- Feedback bias on only selected enterprises.
  (not an issue for those sampled with certainty)

SBR improvement surveys

- A method of updating difficult to reach segments e.g. Where there is a conflict in administrative data
- Can be tailored to all large, with sampling of smaller of businesses
- Can be used to assess quality from other sources e.g. classification

Local Unit Source

- Not many countries have an Administrative source for Local Units
- Survey of enterprises is one option for identifying local units.
- Focus on largest enterprises – more likely to be multisite. Small enterprises likely to be single site !

Profiling

- Using accounts and interviews with enterprise officials to determine structure of large complex enterprises.
- Expensive , but important for getting good structure and statistical data from large businesses.

Combining Sources

- A comprehensive SBR requires the combination of administrative and statistical sources.
- E.g. administrative data will provide source and new records.
- Statistical sources could be used to estimate missing characteristics.

Record linking and matching.

- Often administrative data does not have a common identifier.
- In this case matches have to be based on a similarity measure.
- Typically using characteristics such as name and address

Standardizing

- To achieve good matching results standardisation of text is required
- Making sure same format, tidying up spaces, taking out different abbreviations etc.
- Then text strings comparison techniques can be applied.

Probabilistic matching

- A likelihood ratio is calculated
- Below certain threshold – non match – i.e. they are separate businesses
- Above upper threshold – match – they are the same business
- Grey area in middle – potential matches – need further work

Computational issues.

- Blocking can be used to reduce pairs to be matched. e.g. Block by geographical region
- If used need to iterate with different blocking method.
- Remember : there are commercial products that can save a lot of this work.

Other data Sources

- Commercial data suppliers
- Accounting service /payroll providers
- Internet searching
- Directories etc.
- Watch out for coverage gaps!

That’s all

- Any more questions.