A Quality Framework for Official Statistics Production using Combined Sources

ENP 2018 Georgia
Why do we need a framework?

- Drive to use more ad[Administrative data: data that is collected by sources external to statistical offices,”](https://www.un.org/esa/sustdev/documents/un_ece_nagoya.pdf) Statistical quality from Combined and/or Integrated data

How do we design a

- Design data collections that meet the needs of our users
- Integrate data sources using appropriate methodologies
- Balance cost and quality
  - requires a measure of quality!

---

Administrative data: data that is collected by sources external to statistical offices,” while “administrative sources are data holdings containing information which is not primarily collected for statistical purposes.” (UNECE, 2011)
**Administrative agency**
- **Statistical Agency**

**Metadata for admin sources**

**Determine limitations of admin data source**

**Determine limitations of survey data**

**Determine how each limitation interacts with the design and affects the statistical output**

**Measure(s) of final output quality; compare and evaluate alternatives by quality/cost ratio**

---

**General Structure**

**Initial Collation**

**Phase 1 evaluation**

**Phase 2 evaluation**

**Phase 3 evaluation**
Input Quality

- Evaluate each source with respect to its own original purpose.
- Collate basic metadata, and understand what is known about the data and its processing, and any limitations or issues.
- Don’t make blanket judgements about ‘good’ or ‘bad’ data.
- It is the particular statistical purpose and statistical designs that can fulfil such a purpose that will help deciding if data is useful or not.
Li-Chun Zhang (2012) phase one

Output Quality

- Looks at data with the **output** (statistical) target concept in mind!
- Requires clear definition of the statistical target population and concepts
- Output data is purpose specific and multivariate
- What is done to the input data to produce an output dataset?
- Estimation (in a generic sense) and estimates who give the ultimate results and the quality indicators
Phase two

Measurement (variables)

Target concept
Harmonized measures
Re-classified measures
Adjusted measures

Representation (units)

Input data (single-source and/or integrated micro data)
Transformation from object to unit
Target population
Linked sets
Aligned sets
Statistical units
Integrated (secondary) micro data

Errors:
- Relevance error
- Mapping error
- Comparability error

Error types:
- Coverage error
- Identification error
- Unit error
Phase three

- Processes and Errors introduced when the final outputs/estimates are derived
- Quality Evaluation, Guesstimates

Output Dataset

Data adjustments
Modelling
Analytic processing

Output Estimates

Quality reporting for users
Analysis from Quality Monitoring
Evaluation of Statistical Design
Environment setting

Central repository of available data

Dataset

Standard methods and tools

Standard quality report

Quality measures and indicators

Output dataset

Quality monitoring for analysts

Quality reporting for users

Need

Develop & Design

Production system
Quality information is metadata and the key to efficient reuse
Case Studies in the paper

1. Redesign of the Building Activity Survey

2. Evaluating administrative data for personal income

3. Population estimation in New Zealand
Building Activity Survey

- **Old design:** Quarterly, sample survey, retrospective postal collection of value of completed work.

- **Idea:** Utilize admin data from local authorities on building consents. Save money (given quality) decrease postal forms, change software environment and modernize processes.
# Building Activity Survey: Initial metadata collation

<table>
<thead>
<tr>
<th>Information object</th>
<th>Building consents</th>
<th>Building Activity Survey data (before redesign)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source agency</strong></td>
<td>Local government authorities</td>
<td>Stats NZ</td>
</tr>
<tr>
<td><strong>Purpose of data collection</strong></td>
<td>Track new construction work and provide an early indicator of building activity planned throughout New Zealand.</td>
<td>Provide an estimate of the value and volume of work put in place on construction jobs in New Zealand.</td>
</tr>
<tr>
<td><strong>Target set</strong></td>
<td>All building consents issued by local authorities in New Zealand with a value of $5,000 or greater.</td>
<td>All construction jobs in New Zealand active during the reference quarter.</td>
</tr>
<tr>
<td><strong>Main variables collected</strong></td>
<td>Consent date, consent value, building type, geographic location.</td>
<td>Dollar value of work put in place during the reference quarter.</td>
</tr>
<tr>
<td><strong>Mode of collection</strong></td>
<td>Administrative lists requested from each local authority on a monthly basis.</td>
<td>Quarterly (panel) sample survey using building consents as the sampling frame.</td>
</tr>
</tbody>
</table>
## Building Activity Survey: phase 1 Errors

<table>
<thead>
<tr>
<th>Error type</th>
<th>Building consents</th>
<th>Building Activity Survey data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity error</td>
<td>The target concept is the amount recorded on the form. There is no validity error.</td>
<td>Work done on a job is a well-defined concept for respondents to understand, so the data is very closely aligned with our target data. Minimising validity errors.</td>
</tr>
<tr>
<td>Measurement error</td>
<td>Values are often rounded down by applicants to reduce financial incentive (lower fees) to have a lower value.</td>
<td>Respondents can make mistakes or round numbers.</td>
</tr>
<tr>
<td>Processing error</td>
<td>The main errors that occur in processing are related to capturing the building type in some cases it is extremely difficult to determine the building type based on the description given on the form. Processing errors at this point are because the variable – work put in place – is not always scanned from the survey form. There are some errors in capturing the information accurately.</td>
<td></td>
</tr>
<tr>
<td>Frame error</td>
<td>Cases of consents being given the wrong consent are not being included in the data extraction for a given building type provided to us by the consenting authority.</td>
<td>Some construction work does happen unconsented jobs, especially small ones.</td>
</tr>
</tbody>
</table>
## Building Activity Survey: phase 1 Errors

<table>
<thead>
<tr>
<th>Error type</th>
<th>Building consents</th>
<th>Building Activity Survey data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection error</td>
<td>Every consent in the frame is included in the data by definition.</td>
<td>Actual sample drawn from the consents can be incorrect when building consents data contains errors. This results in a building job being placed into the wrong sample stratum. Sampling errors also arise from the random sample drawn in the lower value strata in the old design.</td>
</tr>
<tr>
<td>Missing/redundancy error</td>
<td>We do not get missing records on the consents because the consent itself is the unit of interest – any consent issued is available in the data.</td>
<td>Unit and item non-response are difficult to distinguish on the Building Activity Survey because (aside from simple confirmations of contact details and so forth) only one statistically important variable is collected on the questionnaire. There is about 10–15% non-response to the survey.</td>
</tr>
</tbody>
</table>
Building Activity Survey: phase 2-3 Errors

- Integrating unit records from Consents and Survey responses
  - Almost no errors from data integration

Corrects for non-response, Erroneous respondent values (Comparability errors)

Modelling Errors of building work done for small construction jobs below the cut-off and hence deliberately not sampled. (Mapping error)

Seasonal adjustment
Outcomes from the Case studies

- Efficiencies from using combined/integrated data versus designs with single survey collection
- Awareness and informed decision making about new processes, quantified risks, process knowledge
- Empirical validation of Quality indicators
- Improved Quality Documentation
- Staff Competence Increase
# The Appendix has Definitions and Q-Indicators

## Quality indicators for selection errors

<table>
<thead>
<tr>
<th>Quality indicator / measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adherence to reporting period</td>
<td>Proportion of units that provide data for a different period than the required reporting period for the administrative dataset. This may be due to lags, delay, or non-compliance with reporting period.</td>
</tr>
<tr>
<td>Dynamics of births and deaths</td>
<td>Changes in birth and death rates of units in the data over time.</td>
</tr>
<tr>
<td>Readability</td>
<td>Proportion of records that can be accessed using existing software for reading data.</td>
</tr>
<tr>
<td>Inconsistent objects/units</td>
<td>Proportion of units that are (and cannot be made) internally inconsistent. Examples are objects involved in non-logical relations with other (aggregates of) objects in the data source.</td>
</tr>
</tbody>
</table>
References


