

WP5: Validating official statistics

A Guide to Data Integration for Official Statistics: High Level Group for the Modernisation of Official Statistics Data Integration Project

Detailed Report

Statistics New Zealand: Using data integration to validate official statistics

2016 ModernStats Data Integration Project Introduction and Purpose HL G-MOS Data Integration Project WPA: a framework for Data Integration WP0 : Data sets for common approaches WP1: Integrating survey and administrative sources WP2: New data sources and traditional sources WP3: Integrating geospatial and statistical information WP4: Micro-macro integration WP5: Validating official statistics Quality Framework for Data Integration Next Steps 2016 Experiment Proposals and Reports 2016 Project Members References and further information 2016 Experiment Reports

10. Validating official statistics (WP5)

Coordinator: Felipa Zabala (New Zealand)

Another specific area where data integration comes into play is using data from other sources as part of validating official statistics. There have been cases where other sources are seen as comparable to official statistics, and when they differ, the official statistics have been challenged. One example from the United Kingdom shows how the distribution of businesses listed in the "Yellow Pages" telephone directories was compared to the coverage of the statistical business register (<http://www.uncece.org/fileadmin/DAM/stats/documents/ces/sem.53/wp.7.e.pdf>). A further example concerns comparisons of inflation figures from MIT's Billion Prices Project against official price indices. These examples show that "other sources" are reaching a level of credibility that challenges the role of official statistical organisations. Market researchers are struggling with similar issues. Their business is already under increasing pressure from cheap or free internet panels. One thing is clear. These "other sources" are here to stay and they will increase in number and influence. Several SDG will use indicators produced from official sources which may not use standardised methodologies and data sources. The different approaches may lead to discrepancies in the results which will have to be analysed and explained.

Expanding the guidelines developed in 2016, if required. Include quality indicators that are essential to report on the quality of the validation process. Communicate findings with the HLG project developing quality indicators for the GSBPM.

This work package will consider the issues involved in integrating alternate data sources into the validation processes used for producing official statistics. Issues include assessing origin and quality of the source, including trustworthiness and commercial or other interests of the parties exploiting them; designing processes and modelling techniques which are sustainable and formalised (as ad hoc adjustments to the statistics would be difficult to defend); and, educating users on proper use and interpretation of information (both the general public and more specific user groups).

Activities:

- Identify issues related to systematically using data from other sources in the validation of official statistics.
- Recommend potential approaches and modelling techniques.
 - Identify issues related to systematically using data from other sources in the validation of official statistics.
 - Recommend potential approaches and modelling techniques.

Experiments:

- A comparative analysis of income data from New Zealand Income Survey with administrative data
- Linking the Statistical Register of Employment and the Labour Force Survey

Broad Activities

Work in 2016 focused on identifying different applications and methods for validating official statistics. Issues identified with use of administrative data to validate official statistics and lessons learnt from experiments as well as other validation projects carried out within organisation of contributing members are documented to provide initial guidelines in the use of administrative data to validate official statistics as well as recommend approaches and modelling techniques to resolve issues identified.

We propose, for 2017, to work towards:

- Investigating the relevance of the ESSNET Rules Repository and ESSVIP validation definitions handbook as tools for validation
- Testing recommended approaches and modelling techniques
- Describe different applications of using data integration for validating statistics (e.g. validating existing statistics, replacing sources, design of new statistics, improving the design of existing statistics, challenging results from alternative sources/suppliers of statistics)
- Develop initial guidelines in the use of administrative data to validate official statistics. Include a description of an initial set of minimum requirements (these could include minimum metadata, process steps, methods) to initiate validation process.
- Determine what validation methods exist or could be created. Include software available to carry out validation methods.
- Document issues identified with use of administrative data to validate official statistics
- Recommend approaches and modelling techniques needed to resolve issues identified in the use of administrative data to validate official statistics
- Test recommended approaches and modelling techniques
- Document experience and lessons learnt

- Expand the guidelines in the use of administrative data to validate official statistics, if required. Include quality measures or indicators that are essential to report on the quality of the validation process. Communicate findings with the HLG project developing quality indicators for the GSBPM
- Develop training materials to carry out validation process using data integration

Issues and Learning

The following section summarises the issues and learning identified so far for this part of the project.

Opportunities

There is an increasing need in the use of data integration in the production of an effective official statistics. Linking multiple data sources allows discovery and examination of underlying relationships between various aspects of society, thus, enabling an NSO to expand its use of external data sources in the production of official statistics. Data integration can occur either at the

- micro level, i.e., linking information from multiple data sources on an individual person of business firm (unit) or
- macro level, i.e., linking information from multiple data sources on a group of people or business firms (units).

Since data integration enables identification of records from multiple data sources that belong to a single individual or unit, data integration has been useful to validate official statistics, either through

- 1) use of external data sources to determine accuracy of survey results or
- 2) use of survey results to challenge results from alternative data sources of providers of statistics.

Either way, data integration provides opportunities for an NSO to determine if an external data source can be used

- to design new statistics
- to fully replace an NSO's existing sources of data collection
- to produce existing official statistics using some components of the external data source
- for benchmarking, imputation, validation, or other methods to improve statistics obtained from existing surveys.

The following are examples of opportunities at Statistics NZ that data integration has provided in its application to validating statistics.

First and foremost, data integration has enabled advancement of data integration skills which had led to the development of a [Data Integration Manual](#) and a [quality assessment framework](#). The manual is a guide to best practice and is a product of Statistics NZ's involvement in several large inter-agency data integration projects. The framework enables understanding of the error sources from individual data sources including those arising from integrated datasets providing assistance in determining the associated methodological and operational issues that may impact on quality resulting from producing statistical information from linked data sources.

The advancement of data integration skills has also led to the creation of [Statistics NZ's Integrated Data Infrastructure \(IDI\)](#). The IDI brings together linked datasets from a range of government agencies (including Statistics NZ's own data collections). The IDI is a large research database containing microdata about people and households and is continually growing. The IDI has paved the way to answer complex research questions to improve outcomes for New Zealanders.

Administrative data have been linked to examine and decide on their specific use in the production of official statistics. Inland Revenue data, specifically longitudinal payroll data from the Employer Monthly Schedule (EMS) returns was linked to produce new statistics - filled jobs, worker flows, and total earnings - that measure labour market dynamics at various levels – including industry, region, territorial authority, business size, sector, sex, and age. These statistics provide an insight into the operation of New Zealand's labour market.

Data integration has also been used for the improvement of a survey process as illustrated in the linking of the March 2013 Household Labour Force Survey (HLFS) to the 2013 Census data to analyse non-respondents to the HLFS. The project led to the deletion of a non-response adjustment step in the weighting procedure for the HLFS simplifying the HLFS estimation process.

Some validation projects involving the use of various administrative data sources have led to recommendations of using these data sources for either benchmarking income survey results, imputation or validation of income statistics rather than using the administrative data sources to replace various sources of income. The administrative data sources need not be integrated to the income surveys when using them for benchmarking or validating income statistics. In cases where data integration will be required for the above immediate uses, a new process – data integration – will need to be designed in the production process.

Linking the Census to administrative data sources in the IDI has been instrumental in the realisation of some of the goals of Statistics NZ's [Census Transformation Programme](#). The programme is investigating alternative ways of running New Zealand's future census including the feasibility of using linked administrative data to replace census questions.

Details on opportunities created by data integration are available in [Using data integration to validate official statistics](#).

Data integration has also paved the way in the development of new methods, e.g., new models. One good example is the production of population estimates using administrative data. Bryant and Graham (2015) use Bayesian modeling to estimate, specifically, regional populations in New Zealand based on administrative data on birth and death registrations, tax and NZ international passenger movements.

Challenges

Integrating multiple data sources face a number of challenges. A number of these follow.

Data integration projects carried out in NSOs may be subject to legislation, codes of practice, protocols and policies, some of which are stipulated in their Statistics Act. Staff working on a data integration project should be aware of the various policies and legislative provisions that affect their project.

Most of the data sources to be integrated are external to the NSO involved in data integration. The NSO had no control in the definition of the concepts and populations used in the collection of the data. Differences in concepts, classifications, populations and collection units are expected. Thus, there is an important need for detailed descriptive metadata to assist in the assessment of the quality of the data sources. The following dimensions of quality need to be assessed: accuracy, relevance, consistency, accessibility, comparability and timeliness.

Using someone else's data means an NSO cannot control any of the decisions on measurements and populations undertaken by an external data source provider. An NSO need to understand the design decisions so they can determine what to do to turn external data into the statistical information they want. These types of difference are expected of external data sources - differences in the definition of populations, concepts and classifications. These differences affect the usability of the external data source in the production of a statistical product specifically with regard to: the coverage of population, the validity of the target concepts, the availability and accuracy of descriptive metadata, sampling error, bias, legal basis for data, data collection methodology/questionnaire design, response burden, by product data versus survey question, confidentiality of the resulting output, and different consequences for different types of data provided. These differences need to be clearly explained and documented, and stored to ensure reuse and improvement of assessments. Good quality variables closely related to each other in different datasets would be ideal to use for linking.

Timeliness of external data sources, unless receipt of data is common and regular, will always be a challenge for linked datasets and for statistics produced from these datasets. These include:

- The promptness in picking up birthed units to an administrative data source. A high number of birthed units not picked up quickly enough by administrative data leads to potential bias in official statistics.
- Timing issues around getting the linking accomplished in time for production.

The cooperation of the dataset owner is also another challenge to address. The NSO needs to ensure the continuity and consistency of the quality of the data to be provided. However, contingency plans need to be in place in case the data source becomes unavailable. The NSO may also need to elicit assistance in determining the definition of concepts, classifications or populations in case these need to be redefined to better suit their needs.

After quality assessment of an external data source has been undertaken, the next challenge to address is the extent an external data source will be used to meet the statistical need. Are new methods required to convert the external data source into a form useful in the production of a statistical output?

Although administrative data may be freely available to an NSO, other external data sources may not necessarily be available for free. Costs may also be a challenge in accessing external data sources. Costs are also incurred in the quality assessments of external data sources and all these costs need to be determined and assessed before proceeding with any data integration project.

Another challenge is the resistance to changing any part of a production process that will involve the integration of an external data source especially when current approaches are widely accepted and well-grounded expertise has been established.

The need for standardised processes which are responsive to administrative changes in the data supplied and to new administrative data available to Statistics NZ should also be addressed when using external data to validate official statistics.

Risk mitigation

Staff working on a data integration project should be aware of the various legislative provisions and policies that affect their project.

Assess the quality of external data sources to determine the extent of its use in the validation of official statistics. A quality framework aimed at determining the optimal design of combining data source(s) that can minimise the cumulative effect of potential errors on a statistical output is essential.

A common issue with linked datasets is inconsistencies in the records linked. Where inconsistencies occur in records linked from two different data sources, it is important to know which of the two data sources is more reliable. Sometimes, even the order in which the datasets are linked is important in determining where an inconsistency arose. It is expected that as the number of datasets being linked together increases, the potential for efficiencies in detecting and treating inconsistencies in records increase as the number of variables increase. However, this may also increase the amount of editing required for the linked datasets. Issues to be addressed by an editing strategy for linked datasets can be summarised by its ability to: edit inconsistencies from the same unit from different sources, treat erroneous and missing variables in a record and ensure consistency in variables across a record for a time period and over time.

Sources of potential bias have been identified with regard to integrating datasets. These include:

- Coverage and conceptual issues may only apply for some groups of a population so care should be taken in generalising results.
- Some variables have the potential to affect the quality of linking and may be a source for potential bias in carrying out analysis on resulting datasets. Investigations on linkage rates across different subpopulations may be required.

The use of linked datasets even for validation purposes may result in a break in the data series that needs to be managed.

Extreme care should be taken in backwards and forward casting of linked data especially for longitudinal data. A person may link in one quarter but not in another due to data quality reasons (or may link to a different record).

A weight may be needed to adjust for missed links in linked datasets.

Methods to better estimate linkage errors are required to determine models appropriate to account for these linkage errors. Linkage errors contribute to potential coverage errors in the resulting target population. Care should also be undertaken when creating statistical units from integrated datasets wherein one dataset is sourced from an external dataset since the unit may be defined differently in the external dataset.

Data sourced externally may suffer from measurement errors, e.g., validity error, and these errors propagate when the data is integrated with other data sources to produce a statistical output. Hence, target concepts used in a dataset sourced externally should be well understood before being used in the production of official statistics.

Users of statistical information should be well informed about the definition of the concepts and populations used in all data sources used in a statistical output. A common understanding of statistical concepts between an NSO and users of the statistical information should also be ensured.

Standard Processes

Standard processes are recommended to perform the following:

- To integrate different types of data sources
- To edit inconsistencies in records linked
- To impute for values of a variable in linked records of an integrated dataset
- To determine weights to adjust for missed links in an integrated dataset
- To carry out validation methods

Recommended methods

There are a number of methods used to integrate data sources at a micro level. A good discussion of these methods is available from Chapter 7 of Statistics New Zealand's [Data integration manual: 2nd edition](#).

A number of methods to integrate data sources at aggregate or macro level are available from the following Eurostat website, https://ec.europa.eu/eurostat/cros/content/macro-integration_en.

ICT considerations

A validation system that involves standardised processes which are responsive to administrative changes in the data supplied and to new external data sources is preferred over an ad hoc system (e.g. done by humans).

Standards

- Concordance between statistical standards and others

- Standards used by source industries (e.g. education, travel, banking)
- Should we link GSIM to VTL Validation Transmission Language? ("Validation and Transformation Language (VTL) | SDMX – Statistical Data and Metadata eXchange," 2017)

Related work in other projects/organisations

ESS.VIP ADMIN Project. This project aims to find ways to optimise the use and accessibility of administrative data sources in the production of official statistics while guaranteeing the quality and comparability of these statistics. Details of work on this project are available from https://ec.europa.eu/eurostat/cros/content/essvip-admin-administrative-data-sources_en.

ESSnet project on Data Integration. This completed project focused on the methodologies and methodological issues of micro data integration. Details of work on this project are available from http://ec.europa.eu/eurostat/cros/content/data-integration-finished_en.

ESSnet project Integration of Survey and Administrative Data. The project aimed at developing the knowledge and expertise of participating NSOs in the use of integrated survey and administrative Data in the production of official statistics. Details of work on this project are available from http://ec.europa.eu/eurostat/cros/content/isad-finished_en.

ESSnet project on macro-integration. This project discusses various methods of integrating data sources at aggregated or macro level. Results of this project are available from https://ec.europa.eu/eurostat/cros/content/macro-integration_en.

References

Bakker, Bart (May 2010). *Micro-integration: State of the Art*. Paper presented at the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, The Hague, The Netherlands.

Biemer, Paul (2010), *Total survey error: design, implementation, and evaluation*, Public Opinion Quarterly, Vol. 74, No. 5, pp. 817-848

Burger, Joep, Davies, Jennifer, Lewis, Daniel, van Delden, Arnout, Daas, Piet, Frost, John-Mark (2013), *Guidance on the accuracy of mixed-source statistics*, Deliverable 6.3 of ESSnet Admin Data work package 6, <http://essnet.admindata.eu/WorkPackage?objectId=4257>

Bycroft, C. 2016. *Using quality frameworks to assess the potential use of administrative sources in the census*. Paper presented at the 2016 International Total Survey Error Workshop, Sydney, Australia.

Daas, Piet, Ossen, Saskia, Tennekes, Martijn, Burger, Joep (2012), *Evaluation and visualisation of the quality of administrative sources used for statistics*, Statistics Netherlands paper at Q2012 conference, <http://www.q2012.gr/default.asp?p=14>.

ESSnet (2013), *Guidance on calculating composite quality indicators for relevant dimensions of the ESS quality framework*, Deliverable 6.3 for ESSnet work package 6, <http://essnet.admindata.eu/WorkPackage?objectId=4257>

Laitila, T., and A. Holmberg. 2010. "Comparison of Sample and Register Survey Estimators via MSE Decomposition." In *Proceedings of the Q2010 European Conference on Quality in Official Statistics*, May 4-6, 2010. Available at: <http://q2010.stat.fi/sessions/special-session-34/>.

Statistics New Zealand. 2012a. *Linked Employer-Employee Data (LEED) project*. Available from www.stats.govt.nz.

Statistics New Zealand. 2012b. *Student Loans project*. Available from www.stats.govt.nz.

Statistics New Zealand. 2013. *Data integration manual: 2nd edition*. Available from www.stats.govt.nz.

Statistics New Zealand. 2014. *Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project*. Available from www.stats.govt.nz.

Statistics New Zealand. 2016. *Guide to reporting the quality of administrative data*. Available from www.stats.govt.nz.

Zabala, F. 2016. *Using administrative data to validate income in Statistics New Zealand's household surveys*. Paper presented at the 2016 International Total Survey Error Workshop, Sydney, Australia.

Zhang, Li-Chun (2012), *Topics of statistical theory for register-based statistics and data integration*, Statistica Neerlandica Vol. 66, nr. 1.

Skills

- Knowledge of regional or global approach and the ability to negotiate with data providers as a collective if necessary for them to make changes to ensure similarity in definition of population, statistical units, concepts and classifications
- Understanding of and evaluation of quality measures or indicators required to assess the quality of external data sources as well as data integrated from different types of data sources
- Application of Cost Benefit analysis for the optimal use of external data sources in the production of official statistics while guaranteeing the quality of these statistics

- Ability to forge agreements (legal)
- Knowledge of UN & /ESSVIP validation definitions - handbook
- Understanding of running and developing validation methods (micro validation and macro validation)
- Communicating reasons for difference in statistics derived solely from statistical surveys and statistics derived from mixed sources of data
- Relationship building with data providers and other information/statistics sources (needs to be active relationship)

Resources

- Methods or tools developed outside official statistics ICT industry - e.g. SAS or R community involvement of academia budget - money (either individual organisation or as a collective) business process specialist Appropriate methodologist expertise ICT - CSPA and collaboration in developing any generic methods and tools plus local ICT for country specific

Partnerships

- Forum for international collaboration - solving common problems

Promotion and advocacy

- Promoting results in conferences, papers, journals

Recommendations: Quality Assessment Framework

- The quality assessment framework, including the quality indicators, described in [Guide to reporting the quality of administrative data](#) is beneficial in carrying out validation studies. The quality framework is based on Li-Chun Zhang's [two-phase life-cycle model for integrated statistical microdata](#) (Figure 1) which expands the total survey error paradigm to include administrative data. The framework enables understanding of the error sources from the individual data sources including those arising from the integrated datasets. Zhang's two-phase life-cycle model assists in determining the associated methodological and operational issues that may impact on quality resulting from producing statistical information from linked administrative data sources. Phase 1 assesses the quality of an input data source that is intended to be used in the



Figure 4 Zhang's two-phase life cycle model

production of a statistical product. An NSO needs to understand the design decisions undertaken by the producers of the source to determine methods to turn the data into the statistical information required by the NSO. Quality of the input data source is assessed against the purpose for which it was collected. For a survey dataset, this purpose is defined for a statistical target concept and target population. For an external data source, the entries or 'objects' in the dataset might be people or businesses, but they could also be transaction records, or other events of relevance to the collecting agency. At this stage, evaluation is entirely with reference to the dataset itself, and does not depend on what an NSO intends to do with the data. Quality issues in the input data source will flow through into any use of the data in the production of a statistical product.

Phase 2 categorises the difficulties arising from taking variables and objects from source datasets and using them to measure the statistical target concept and population an NSO are interested in. In this phase, the NSO consider what they want to do with the data, and determine how well the source datasets match what they would ideally be measuring.

The quality assessment involves 3 steps.

Step 1: Initial metadata collation: Basic information is collected about each of the source datasets used in the validation project. The information relates to the source agency, purpose of the data collection, populations, variables and timeliness of the data.

Step 2: Phase 1 evaluation: Errors occurring in phase 1 of the quality framework are determined and categorised for each source dataset. This involves detailed consideration of how the methods, purpose, known issues, and other aspects of the original data collection contribute to each of the specific error categories in the phase 1 flow chart in figure 1.

Step 3: Phase 2 evaluation: As for the previous step, errors arising in phase 2 of the quality framework are listed and examined in a similar way, taking into account the dataset(s) being integrated to produce the final output. These errors are considered with respect to the intended statistical target concepts and population. The effects of phase 1 errors on the creation of statistical units, or the particular details of the misalignment between concepts on different datasets, must be understood.

- Understanding the error sources from the input data sources, including those arising from the linked datasets, is essential in determining the associated methodological and operational issues that may limit the use of administrative data in the validation of official statistics.
- The key quality dimension to use is accuracy. The normalised root mean square error is a useful tool to evaluate the accuracy of statistics sourced from integrated datasets. Laitila and Holmberg (2010) suggest an approach to estimating the total error of an estimator from integrated dataset by deriving lower and upper boundaries for a total mean square error measure.
- The [Guide to reporting the quality of administrative data](#) provides a metadata information template that encourages thinking about the key aspects of quality in an organised way. It is also a convenient way to record a standard set of information to compare different datasets. You do not necessarily need all the meta information. The basic information required are: name of data source agency, purpose of data collection, time period covered by the data, the population (target and actual) population of the dataset, the reporting units, a short description of key variables and the timing/delay information and method of collection.