

WP1: Integrating survey and administrative sources

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

[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](#)

6. Integrating survey and administrative sources (WP1)

Coordinator: Kaja Malesic (Slovenia)

Some countries already have extensive experience integrating survey and administrative data sources and there have been collaborative projects in this area, for example Eurostat projects.

The administrative data may have existed for some time but not been used. It may be integrated using record linking or statistical matching or may use modelling approaches. It may involve pooling or combining information from multiple surveys, including surveys not conducted by the NSOs themselves.

There are common challenges faced in this type of integration. The quality of administrative dataset may be good enough for administrative purposes but not sufficient for statistical purposes. Transforming administrative datasets into statistical datasets may require improving the quality and dealing with conceptual differences, especially when we want to use administrative data in a direct way. In the case of surveys carried out with the use of data from administrative sources it is crucial to gather all data (from survey and administrative sources) in one database.

Examples of sources that can potentially be integrated are: Labour Force Surveys and social insurance register and/or educational registers, data from ministries of culture and cultural associations to produce statistics on museum attendance and there are several examples of administrative data being combined with survey data for producing indicators traditionally collected through censuses.

Activities

- Definition of statistical framework and boundaries of the work.
 - Select Administrative sources and a survey dataset
 - Define a case using records from the administrative dataset instead of records from a survey (a smaller frame) or adding additional variables from the administrative survey (a shorter a survey questionnaire) or both.
- Design expected outputs of the integration
- Design and test of transformation of the administrative dataset into the statistical dataset. - A cleaning method has to be designed
- Design and test of the integration of administrative data and survey data.
- Evaluation of statistical outputs obtained from the integration of administrative datasets and survey datasets.

Experiments

- Integrating potential information sources for the statistical data production on job vacancies
- Linking the Statistical Register of Employment and the Labour Force Survey
- System of consultation and geographic location of schools

Broad Approach

[Job Vacancies and Overtime \(JVOT\) - Hungary, Slovenia and Serbia](#)

- document experience and lessons learnt on integrating survey and administrative sources for Job Vacancies and Overtime (JVOT) statistics, including linking, data quality and methodological issues and classifications issues (e.g. same classification, different sources and different data can result in subjectivity)
- develop plans for sharing (research) data; via sandbox if possible
- involvement of Big Data sources, if possible
- document experience and lessons learnt related to using national and international classifications and/or mapping approaches and how to handle discrepancies when updated.
- document experience and lessons learnt for data integration methodology, e.g. classifications and coding to a unified/same level.
- develop approach for managing quality and validation.

[System of consultation and Geographic location of schools](#)

Share and document experience and lessons learnt on the System of consultation and Geographic location of schools

Issues and Learnings

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

Opportunities

The degree and systems of integrating administrative and survey sources vary greatly across countries; some have fully developed register-based statistical systems, while others are just starting to integrate the data. In the official statistics production process, administrative and survey data can be integrated in different ways. Usually, administrative data are the source for the population frame for sample surveys. They can also be used to supplement surveys in questionnaires, for a part of the population, for a set of variables, for estimation or for the data validation and editing process. In some cases, administrative data can replace the sample survey; in these cases the statistics are based entirely on administrative sources. Administrative data can also be a source for establishing and maintaining statistical registers data, which are further used in implementing surveys.

Sample surveys are generally more flexible than administrative sources as they are designed to meet a precise purpose. Administrative sources are on the other hand the result of a legislative system. Administrative sources usually offer better coverage of target populations and in general have high response rates. As these data are already being collected for administrative purposes, it is cost-effective and cheaper to acquire the data than to conduct a sample survey. Also, there is no additional respondent burden. The ability of administrative data in covering whole populations enables the production of local area data to a level of detail not permitted by sample surveys, which is also of advantage in implementing local policies.

Challenges

There are a number of challenges in integrating administrative and survey data. Since administrative data are collected for non-statistical purposes, the difference in concepts might lead to coverage problems as well as bias problems. In some cases, such as business statistics, units do not necessarily correspond directly to the definition of the required statistical units. This requires some modelling for converting the administrative units into statistical units. It is likely that there will also be differences in the definitions of variables. It is important to have a thorough knowledge of the impact of these differences. Sometimes it is possible to influence the administrative definition by co-operating with the responsible authority.

Another issue is classifications. In cases of different classifications, the usual step is to use correspondence tables and conversion tools based on additional variables that may be available for converting into more correct classification code. However, even the same classifications may result in different data, especially when classifications are complex or the rules of a classification are difficult to apply. In administrative sources, there would often be respondent coding, while a sample survey may have open questions and coding is often done by experts. Co-operation between the NSI and the administrative authority is a good way to solve a part of the classification problem. The NSI can provide experience and may be the one responsible for maintaining the classification. Another issue that concerns classification is a decision to use directly translated international classifications or national classifications. It depends on what national data are needed; however, the first option is usually harder to implement in case of changes and revisions compared to having national classifications. To change a classification in an administrative source is a demanding task since there can be many data providers that need to become familiar with the changes.

Problems to overcome are also the missing data and errors. Missing data happen due to unit or variable non-response, but in administrative sources the causes can be different. It is important to identify if errors and missing data are systematic and apply appropriate validation and editing rules.

Timeliness is one more point in integrating administrative and survey data. Administrative data may not be available in time or may not coincide with the statistical reference period. It can be resolved by analysing the impact and if necessary adjusting it by models.

Risk mitigation

Legal basis

The first and most important for the use of administrative sources for statistical purposes is the legal basis. It is sound if national legislation is aware of already existing administrative sources rather than recollecting data. The usage of administrative sources is often stated in a Statistical Act. To assure public acceptance, a Personal Data Protection Act is also important. It determines the rules on processing personal data in a way that the legal rights of the individuals concerning privacy and integrity of individual's data are not violated.

Collaboration with administrative data providers

Administrative records are data collected for the purpose of implementing various non-statistical programs concerning legal requirements such as taxation, housing, pensions, social benefits, trade in goods, etc. Statisticians may have to make compromises concerning coverage, data quality, classifications, etc., in administrative sources. Collaboration of the NSIs with administrative authorities in the preparation of legal documents establishing and maintaining an administrative source is a good solution to overcome this problem. The approval of the NSIs in passing legislation on administrative records may be stated in a Statistical Act.

Institutional methodological groups

Control of the methods by which the administrative data are collected and processed rests with the administrative agency. They are specialized in formulating transparent rules and procedures. The NSIs have experience in data collection, classifications and data validation. In some cases, the same data are used by several institutions, so continuous collaboration in institutional methodological groups is recommended to develop a system that is satisfactory for administrative and statistical purposes.

Cooperation agreements

Cooperation agreements are signed to divide the tasks between the parties of the agreement, to define the rules and conditions of transferring data such as timeliness, technical implementation and metadata.

Unified identification system

Existence of a unified identification system across different sources is one of the most important aspects in the integration of administrative data. If there is no such system, it is much more difficult to link different sources. In such a case linking and matching methods must be applied.

Collaboration in legal acts and policies

Collaboration in institutional methodological groups

Recommended methods

All administrative sources are different, which may result in using different methods. Some of the methods relevant to the use of administrative data are:

- data linking and data matching
- data modelling to overcome conceptual differences
- data cleaning
- data validation, data editing and imputations
- small area or domain estimation to obtain coherent detailed output
- post calibration to improve data coherence

ICT considerations

A significant number of tools exist for record linkage and matching (e. g. Statmatch in R...). These will be documented.

Standards

These are some links to standards which are relevant to integrating administrative and survey sources:

- Quality measures and quality reporting in the ESS <http://ec.europa.eu/eurostat/web/quality/quality-reporting>
- Generic Statistical Process Model (GSBPM) to define the statistical process <http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0>
- Classifications http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM&StrGroupCode=CLASSIFIC&StrLanguageCode=EN

Related work in other projects/organisations

UNECE Assist. Knowledge base on the use of administrative and secondary sources in statistics <http://www1.unece.org/stat/platform/display/adso/ASSIST>

ESS.VIP ADMIN Project. The project purposes are to support the EU Member States to reap the benefits (decrease costs and burden, increase data availability) of using administrative data sources for the production of official statistics, and to guarantee the quality of the output produced using administrative sources, in particular the comparability of the statistics required for European purposes. https://ec.europa.eu/eurostat/cros/content/essvip-admin-administrative-data-sources_en

ESSnet project on Data Integration. The project focused on methodologies for data integration (Record Linkage, Statistical Matching, Micro integration Processing) and on statistical aspects to be considered to make those methods concretely applicable by NSIs. http://ec.europa.eu/eurostat/cros/content/data-integration-finished_en

ESSnet project Integration of Survey and Administrative Data. The project purpose was to promote knowledge and application in practice of sound methodologies for the joint use of existing data sources in the production of official statistics. http://ec.europa.eu/eurostat/cros/content/isad-finished_en

Eurostat (2013): The use of registers in the context of EU-SILC: challenges and opportunities. <http://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF>

UNECE (2007): Register-based statistics in the Nordic countries. Review of the best practices with focus on population and social statistics. <http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2764>

UNECE project on Quality Indicators for the [Generic Statistical Business Process Model \(GSBPM\)](#). This is an on-going project aimed at developing quality indicators to monitor the quality of the statistical production process for each of the phases of the GSBPM, including the sub-phase, 'integrate data'. The project is currently reviewing and updating the quality indicators to include the use of administrative data in the production of official statistics. On-going work is available from <http://www1.unece.org/stat/platform/display/QI/Quality+Indicators+Home>.

Skills

These are some of the essential skills needed for integrating administrative and survey sources:

- leadership and negotiation skills are useful for participating in policy development and in discussions with administrative data providers.
- legal skills relate to legal basis for obtaining data, data protection and a co-operation agreement between the administrative authority and the NSI.
- subject-matter statistician skills cover expertise in knowing data content, understanding and analysing data, knowing the statistical process and dissemination methods.
- methodological skills relate to all statistical processes such as sampling frame preparation and selection of observation units, data linkage and matching, weighting, time series analysis and seasonal adjustment, data protection, etc.
- programming, software and database skills are needed for construction of microdata databases and for establishing and maintaining generic and non-generic process programs (e.g. for aggregation and tabulation, data protection).

Resources

The resources needed for integrating data include budget, IT infrastructure and human resources. Administrative data are usually cheaper than sample surveys as they are already being collected for administrative purposes, but they would still require some budget. The fact that acts in favour of data integration is the rapid development in the IT area, i.e. hardware equipment as well as a wide range of software tools. The IT infrastructure needed for integrating data covers servers, tools for database development where microdata and metadata are stored (e.g. Oracle, SQL), software for data processing (e.g. SAS, R) and different tools for data processing and dissemination. Human resources include subject-matter statisticians, methodologists and IT experts.

Partnerships

- collaboration and sharing experience with statistical institutions
- collaboration with data providers
- collaboration among institutions in the country (e.g. tax offices, employment office) – together deciding the methodology with stats office providing classifications
- conferences, papers...
- sharing information and experience by organizing common events (e.g. once per year) with data providers and data user
- measure data quality
- take into account the common goal of institutions (collecting data only once, reducing unnecessary expenses)
- good and continuing collaboration with data providers
- international recommendations for statistics (concerning improvement of data access in some countries)
- personnel exchange to gain experience learn from good practices Work

Promotion and advocacy

- conferences, papers...
- sharing information and experience by organizing common events (e.g. once per year) with data providers and data user

Recommendations

- measure data quality