Report for
the European Commission
Eurostat

Task N° 3

D3: Report presenting good practices on the use of administrative sources by statistical domain and by type of sources

Quality, methodology and research
Lot 1: Methodological support

Good practices when combining some selected administrative sources

Framework Contract N°: 11111.2013.001-2013.251 Lot 1

Contract ESTAT N°11111.2013.001-2017.400
Specific contract Ref. N°: 000094

25 April 2018

GOPA Consultants
Hindenburgring 18, 61348 Bad Homburg, Germany
Phone +49 6172 930-521 Fax: +49 6172 930-130 Email: statistics@gopa.de

Table of contents

Abbreviations

AD	Administrative data
AM	Account manager
AS	Administrative source
CBS	Centraal Bureau voor de Statistiek (The Netherlands)
CSA	Classification of statistical activities
EDI	Electronic Data Interchange
ESS	European Statistical System
ESSnet	Collaborative ESS networks
ESS.VIP ADMIN	ESS Vision 2020 Implementation Project for Administrative Data
ETL	Extract-Transform-Load
EU	European Union
FAO	Food and Agriculture Organisation of the United Nations
FSS	Frame structure survey
FTP	File Transfer Protocol
GSBPM	Generic Statistical Business Process Model
IACS	Integrated Administration and Control System (agriculture)
ID	Identifier
IT	Information Technology
Istat	Italian Statistical office
KOMUSO	Project on the quality of multisource statistics
LCS	Labour cost survey
NACE	Statistical Classification of Economic Activities in the European Community
NSI	National Statistical Institute
OECD	Organisation for Economic Cooperation and Development
SBR	Statistical business register
SBS	Structural business statistics
SEM	Structural equation modelling

SF	Statistics Finland
SFR	Statistical farm register (Italy)
URS	Statistical business register (Austria)
URV	Administrative business register (Austria)

1 Introduction

The ESS.VIP ADMIN project documents numerous activities that are making administrative data more usable for statistical production. This report aims to summarise a number of the efforts performed in the project, together with presenting other interesting approaches that are being taken by the ESS and by the NSIs. A careful analysis of the available material showed that the activities can be summarised under seven topics. Each topic is associated with a number of phases or sub-processes of the GSBPM, as applied to producing statistics using administrative data. A number of good-practice indicators were defined, in each topic. The following lists the topics, their connection with the GSBPM and the indicators.

1. Promotion of a culture for the use of administrative data in statistical production
   GSBPM phases: 1 & 2
   This topic refers to conceptual and organisational activities associated with the use of administrative sources, and to basic design decisions for the use of administrative sources.
   Indicators:

   Indicator 1a): Legal act for using administrative sources;

   Indicator 1b): Cooperation agreement with data owners (including, in the ideal case, technical and conceptual support with using administrative data);

   Indicator 1c): Development of a privacy and confidentiality policy;

   Indicator 1d): Active search for exploiting new administrative sources;

   Indicator 1e): Development of back-up strategies for administrative sources;

   Indicator 1f): Inventory of sources with respect to population coverage;

   Indicator 1g): Inventory of sources with respect to variable description;

   Indicator 1h): Quality policy for administrative sources.

2. Development of an appropriate infrastructure within the NSIs
   GSBPM phases: 3.1, 3.2 & 3.4
   This topic refers to the development of concepts for the use of administrative data in statistical production. It specifies and builds the necessary collection and process components for using administrative sources. It furthermore defines a top-level configuration of workflows.
   Indicators:

   Indicator 2a): Organisation of data provision by the administrative sources (including responsibilities in the participating organisations);

   Indicator 2b): Development of a data repository of the administrative sources (data warehouse of administrative data within the NSI);

   Indicator 2c): Development of a metadata repository of administrative sources;

Indicator 2d): Development of methods and standards for the attribution of pseudonyms;

Indicator 2e): Definition of data-exchange formats.

3. Implementation of quality standards for statistical products based on administrative data
   GSBPM phases: 3.1 to 3.4
   This topic refers to the adaptations of the statistical quality standards (Code of Practice) that are necessary, in connection with the use of administrative data. The benchmark is defined by the results of the KOMUSO project.
   Indicators:

   Indicator 3a): Quality standards (indicators) for the input quality of administrative data;

   Indicator 3b): Quality standards (indicators) for process quality, when using administrative data;

   Indicator 3c): Quality standards (indicators) for the output quality of products, when using administrative data.

4. Feasibility studies about the use of administrative sources
   GSBPM phases: 3.5 & 3.6
   This topic refers to the necessary tests to the envisaged production system using administrative data.
   Indicators:

   Indicator 4a): Assessment of the variables with respect to their use as substitutes for or supplements to traditional variables

   Indicator 4b): Assessment of the data integration processes, in the case of the production of multisource statistics;

   Indicator 4c): Assessment of the coverage of the administrative sources used;

   Indicator 4d): Assessment of the process design for production using administrative sources.

5. Implementation of standardised workflows for statistical products (statistical tables and registers, in particular, maintenance based on administrative data)
   GSBPM phases: 5 & 6

   This topic refers to all the methodological issues connected with the use of administrative data, in sub-processes of production.
   Indicators: Sub-processes for:

   Indicator 5a): Data matching and statistical linkage;

   Indicator 5b): Statistical matching;

   Indicator 5c): Editing and Imputation;

Indicator 5d): Alignment (harmonisation) of statistical units;

Indicator 5e): Life-sign approach for statistical units;

Indicator 5f): Alignment (harmonisation) of measurements (variables);

Indicator 5g): Substitution of variables;

Indicator 5h): Supplementation of variables;

Indicator 5i): Update of statistical registers;

Indicator 5j): Output preparation (e.g. macro editing).

6. Improvement of methodological knowledge for the use of administrative data
   GSBPM phases: (3.2 & 3.3), 5 & 6
   This topic considers good practice in the development of new tools and methods, when using administrative data.
   Indicators:

   Indicator 6a): Innovation: Development of new methods for using administrative sources;

   Indicator 6b): Knowledge transfer: Integration of the methods of an NSI that has shown good results. in other NSIs.

7. Carrying out analytical investigations into the use of administrative data
   GSBPM phase: 6
   This topic refers to the enhanced use of statistical products in the analysis of domain-specific problems.
   Indicators:

   Indicator 7a): Reliability of products based on administrative data: Consistent performance of production according to the required function;

   Indicator 7b): Validity of products based on administrative data: Conformity with the intended meaning;

   Indicator 7c): Analysis of new applications (e.g. micro-simulation).

This document discusses examples of good practice with respect to those criteria. Practically all NSIs in the ESS have begun transforming their more or less survey-oriented statistical production towards one, which makes more intensive use of administrative data. The validity of the examples presented depends on the level of this transformation. Another point must be taken into account is the domain of application. While using administrative data already has a certain tradition in economic statistics (e.g. National Accounts), in social and demographic statistics, it is often a new application.

The examples presented in this deliverable are taken from the results of the ESS.VIP ADMIN project but also from other sources in the ESS and other NSIs examples usually only cover a number of the above-mentioned topics. They should be understood as building blocks, which

hopefully contribute to a better understanding in the use of administrative sources in statistical production.

The report is organised as follows: Chapter 2 discusses good practices from a general point of view as they are rather independent of the application domain. It considers the organisation of Topics 1, 2 and 3 within the NSIs. Chapter 3 describes examples of good practice for demographic and social statistics, the main emphasis being on census applications. Chapter 4 describes examples of good practice with economic statistics. Finally, Chapter 5 summarises the findings of the examples. The approaches different countries apply are presented for each domain. The examples provided in Chapters 3 and 4 appear within a standardised structure, beginning with a short problem description and followed by a brief explanation of the important facts about the methods used. Each example of good practice is characterised by the following descriptors:

Countries:

Application domain: The Classification of Statistical Activities (CSA) is applied for the domain's designation.

Typology of sources involved: this typology employs three descriptors

Type of source: administrative / statistical

Type of aggregation level: micro-data / macro-data

Type of product: census / register / data / survey / intermediate / other

Topics in and indicators of good practice

Hyperlink to documentation

It should be noted that good practice must bear in mind individual countries' administrative and legal backgrounds, together with the way their NSIs are organised internally. Hence, these are not templates that one can apply automatically, but they can broaden thinking about solutions and about interesting questions.

Besides the examples of good practice described in detail in this report, Task 1 and Task 2 of the project reviewed the use of administrative sources in the statistical production system for different statistical domains and for different administrative sources. A template was defined for these reviews, which provides more detailed information about the sources used. Both the template and the applications considered can be found in the Appendix.

2 General good practices of NSIs in using administrative data

In this chapter the practices of three NSIs in Topics 1, 2 and 3 are considered. In each topic, the presentation follows the list of indicators for the topics. The examples are taken from Finland (SF), Italy (Istat) and The Netherlands (CBS).

2.1 Promotion of a culture for using administrative data

Indicators 1a, 1d): Inventory of administrative data, search for new sources

In the **Netherlands,** the Dutch Data Protection Authority is the place to start with a search for potentially new administrative sources. By law, this authority knows which organizations hold personal data. A formal agreement or a specific law is useful to define a procedure involving the NSI in analysing new administrative sources that are potentially useful for statistical purposes.

Indicator 1b): Cooperation agreements with data owners

In the **Netherlands** Chapter 5 of the Statistics Netherlands Act (2003b) grants legal permission for the use of data from public sources. Agreements also need to be reached with the data source holder, on the delivery and on any other arrangements made such as the additional need for feedback or assistance.

Indicator 1e): Backup Strategies for administrative sources

The CBS in the **Netherlands** developed a “fall-back scenario”, that is a combination of measures enabling Statistics Netherlands to deal with the unfavourable consequences of data provided by others becoming temporary unavailable. The main steps defined by CBS for the development of a “fall-back scenario” are the following:

- **Applicability:** A risk analysis needs to be performed and a fall-back scenario has to be drawn up for the statistics belonging to the list of ‘image-determining’ statistics of Statistics Netherlands;
- **Detailed description:** It is impracticable to prepare fall-back scenarios for all imaginable situations. Fall-back scenarios are often tailored to a specific practical situation. The best solution depends on what is missing and on the remaining information. The chosen solution must also address costs and time available.
- **Standard template:** A standard template has been created to determine the need for developing a fall-back scenario for a given statistic.

Indicator 1c): Development of privacy and confidentiality policy

Concerning personal data protection in **Finland**, the data collected for administrative purposes may be released to third parties, as well as for the purpose of compiling official statistics. The Finnish Statistics Act (2004) requires, wherever possible, that official statistics be compiled using AS. Concerning confidentiality, the Statistics Act stipulates that data received from administrative records are confidential. The flow of information only runs in one direction: from the administrative

authorities to the NSI, never the other way around. The Statistics Act also states that identifier data may only be collected by the NSI if they are needed to link different data sets.

Concerning the protection of personal data, in **Italy**, data collected for administrative purposes may only be released for the compilation of official statistics if they are declared in the National Statistical Program (PSN), which is defined by Istat, and which is approved by the Italian Authority for Confidentiality. The Italian statistical law (1989) concerning confidentiality, states that data received from administrative records are confidential. The flow of information runs in one direction only: from the administrative authorities to the NSI, never the other way around. The statistical law also requires that identifier data be collected by the NSI only if it is necessary in order to link different data sets.

In both cases described, it is possible to provide the owner of aggregate data with feedback, to stress quality problems with the data sources.

2.2 Development of an appropriate infrastructure inside the NSI

Indicator 2a): Organisation of data provision for administrative sources

Statistics Finland (SF) has defined a specific architecture for data provision and cooperation:

- SF appointed a contact person for each administrative-register authority: to monitor developments, and to maintain and improve the statistical applicability of the specific data;
- Each administrative register authority nominated a responsible person for statistical issues;
- Task force “Register Pool”. The group has the objective of promoting information exchange among register authorities, with a view to improving the usability and consistency of registers, developing the contents’ quality and accessibility, and increasing the cooperation among the participants. The Register Pool is appointed by the Ministry of the Interior for two years at a time.
- Every year SF arranges a meeting at Director General level, to monitor progress in cooperation.

In a similar fashion, in the **Netherlands,** the CBS has appointed an **Account Manager (AM)** for the most important data sources holders (Tax authorities, Municipal personal records database, etc.), who is expected to:

- To provide information to users and to gather it from the sources;
- To make and to monitor agreements.

The agreement (recorded in a formal way) with the holders is defined by the AM to establish the usability of a potential source, the delivery of the source (with all the information useful for the CBS, i.e. metadata about the administrative source), the use of data, and all mutual obligations involved.

The AM is the internal contact person for any problems regarding the source. Any contact by NSI sector experts with the data source holder is through the AM.

By law “Governmental organisations have an obligation to report to the data holder any suspicion of error in the data on an individual level”, but this is in conflict with the provisions for confidentiality as laid down in the Statistics Netherlands Act (CBS). For this, Statistics Netherlands avoids feeding back information about individual records, reporting instead on an overall level or – only in exceptional cases – with anonymised data.

For the usability of administrative sources for statistical purposes, **Finland** defined two levels of relationships between Statistics Finland and the Administrative sources holders:

- At an operational level, by the presence of contact person (inside the NSI) and a statistical contact person (inside the AS holder);
- At a strategic level, by creating a specific task force: The Register Pool.

In **Italy**, the compilation of administrative registers began in the 1990s. Nowadays, the Italian office Istat collects about 100 different dataset from public bodies. To guarantee the coherence and consistence of this information, Istat developed a specific centralised structure (Directorate).

Istat developed relationships with the main AD owners, taking into account the national law concerning statistics, by defining bilateral partnerships. Starting from the “Annual Plan of Acquisition of Administrative Data”, a specific agreement is defined each year. The plan includes:

- The sources and the variables to transfer to Istat;
- The metadata (essentially: definitions, classifications and target population) on the single sources, and the changes with respect to the previous year;
- The timetable (data and events);
- The exchange of the data.

Indicator 2b): Development of a data repository for administrative data

In **Finland** the most important factor facilitating the statistical use of administrative sources is the presence of a unified identification system across different sources. A personal identification code system has been in place since 1963. Similar systems exist for businesses, buildings, and dwellings.

In **Italy** the Integrated System of Administrative Micro-data (SIM) was developed. Its main activities are:

- Coordination of a group of Istat users, to define the “Annual Plan of Acquisition of Administrative Data”. The different parts of the productive structure are represented in the group;
- Collection of AD requirements from Istat statistics producers;
- Formulation of AD requests for each AD holder and for each AD source (a unique channel between Istat and the owner of AD);
- AD acquisition and storage;
- Procedures to ensure data confidentiality;
- Data loading;
- AD Integration (identifying the same units in different data sources and defining a unique identifier code. This statistical code is also useful in guaranteeing the data’s confidentiality);

- Recoding (developing the first check and correction of the data sources, in accordance with the statistical definitions);
- Dissemination to internal users.

2.3 Implementation of quality standards

The increasing use of data from administrative sources in **Finland** has led to increasing dependency on the quality of data collected by others. The NSI must bear this dependency in mind, when designing and re-designing statistical processes. Statistics Finland chooses two main lanes:

- A close cooperation with the AS holders, so as to have any information on changes affecting legislation or the procedures used by the owner in managing the AS as soon as possible;
- Special surveys developed by the NSI, or use of current surveys (with changes adopted for this goal) to monitor the quality characteristics of the administrative sources being used for statistical purposes.

In **Italy,** the first step to adding a new dataset to the SIM is to check the source's quality. A large part of the effort goes into checking consistency with the previous version. The eventually found discrepancies are analysed with the owner from two points of view: i) Identifying changes in the source's characteristics (change in the metadata or in the procedure implemented by the owner, to analyse and to archive the source); ii) Identifying actual mistakes in the source, which are exchanged with Istat.

In a general way, Istat choose two main lanes to monitor the administrative sources' reliability: close cooperation with the main AS holders, so as to be informed on changes in legislation or the procedures applied by the owner of the AS, as soon as possible; if necessary, to develop a special survey (or to use a current survey) to monitor the quality and characteristics of the administrative sources being used for statistical purposes.

3 Good practice in demographic and social statistics

This chapter considers examples of good practice in the domains of the population census. Two different scenarios can be distinguished. The first concerns good practice in the transformation from a traditional census towards a register-based census. The second concerns the production of the register-based census itself. Section 3.1 considers the first case, while section 3.2 focuses on the second scenario. Section 3.3 presents good practice in performing feasibility studies in the domain of labour statistics. Sections 3.4 and 3.5 present two examples of good practice in the development of new methodological knowledge necessary in the use of administrative sources. Finally, Section 3.6 shows the potential administrative sources have in advanced analysis.

3.1 Good practice in the transition towards a register-based census

3.1.1 Problem description

The main emphasis in demographic and social statistics is in the area of the population census. Worldwide, NSIs are making efforts to use administrative data for their censuses. The ESS.VIP ADMIN project launched national projects for transformation towards a register-based census in the following NSIs: Czech Republic, Croatia, Hungary, Latvia, Lithuania, Poland, and Slovak Republic. Naturally, the main emphasis is on the realisation of good practice in Topics 1 – 3 stated in the introduction. The results of these efforts are documented in Work Package 6 of the ESS.VIP ADMIN project and summarised in Deliverable 4 of Work Package 7. This summary uses the same set of indicators to describe the projects.

An overview of this project's findings is provided in the following. It is organised according to the results obtained for different indicators.

3.1.2 Method description

Indicators 1a & 1b): Legal acts concerning the use of administrative sources and cooperation

Besides conceptual analysis, the administrative and legal aspects that govern the use of administrative sources must be settled. Good practice also means thinking about the legal framework in the national statistical law obliging cooperation in data exchange between the administration and the statistical office.

Good practice also means discussing the evaluation made with respect to the above-mentioned characteristics with the data holders, and thinking together about improvements to the administrative registers and data. In particular, identifying the register units is of utmost importance and medium term solutions about possible improvements are reported.

Taking the different backgrounds into account, all NSIs involved displayed good progress in this direction.

Indicators 1f) & 1g): Inventory of administrative data, search for sources

Depending on the existing infrastructure, the first step is that of screening the available administrative sources in the country. This inventory needs to describe the existing sources (register or data) from different points of view:

- Coverage of the population by the registers;
- Availability of identifiers for the administrative units;
- Relationship between the administrative units and the persons, who define the register's population;
- Concepts of the variables available in the administrative sources;
- Measurement methods and variable domains;
- Stability of the concepts used in the sources;
- Maintenance and update policy of the registers.

Such an inventory has to be performed in cooperation with the holders of the administrative data.

Practically, all register projects in the ESS.VIP ADMIN WP 6: "Pilot studies and applications" carried out such an inventory.

Indicators 2a) & 2b): Organisation of data provision and development of a data repository

Development of an IT infrastructure within the NSIs is of utmost importance for a register-based census. That means, on one hand, secure and robust data exchanges, and on the other, an appropriate solution for storing the administrative data. Probably the best solution is the development of a data warehouse for administrative data. A number of countries participating in the ESS.VIP ADMIN WP 6: "Pilot studies and applications" have already started activities in this direction. Some countries, for example Hungary, can rely on already existing pilot implementations.

In connection with the development of the data warehouse, all aspects of confidentiality and security of the private data must be taken into consideration. The standard solution, with these issues, is that of using the technique of pseudonymisation of the data. Because pseudonymization is not only important in statistical applications of administrative data, good practice is the cooperation with a national data security agency and the specification in the national legislation.

Indicator 3a) Quality standards for the input quality of administrative data

All questions of quality must be solved, in connection with the development of the data repository of administrative data. Contrary to the traditional consideration of quality in statistics, the quality of register applications depends essentially on the quality of the inputs provided by the owners of the administrative sources.

A number of issues were raised in the different countries. A number of generally identified quality aspects refer to the issues above, in connection with indicators 1f) and 1g).

In the long run, good practice should be guided by the application of the standards defined by the KOMUSO project. Ideally, a model is developed, allowing the determination of the quality of statistical products on the basis of the quality of the administrative sources. Such an approach requires a metadata repository for the administrative data.

Indicator 3b) Process quality standards

The analyses of input quality lead to different solutions for the measurement of process quality. In particular, a number of NSIs in the project have started with the following activities for the evaluation of process quality:

- Definition of editing criteria and evaluation of editing results;
- Specification of matching criteria and evaluation of the matching results;
- Procedure for the alignment of variables.

In some countries, findings about process quality lead to the decision of having a combined census, as a first step in the transformation.

3.1.3 Summary of the use case

Countries: Bulgaria, Czech Republic, Croatia, Hungary, Latvia, Lithuania, Poland, Slovak Republic

Application domain: Population (CAS 1.1)

Typology of sources involved (employing three descriptors):

Type of source: administrative / statistical,

Type of aggregation level: micro-data,

Type of product: register / data / survey / intermediate.

Topics and indicators for good practice:

Indicator 1b): Cooperation agreements: Czech Republic, Hungary, Latvia, Lithuania, Slovak Republic

Indicator 1c): Development of a privacy and confidentiality policy: Czech Republic, Lithuania

Indicator 1f): Inventory of sources: Czech Republic, Croatia, Hungary, Latvia, Lithuania, Poland, Slovak Republic

Indicator 2a): Organisation of data provision: Bulgaria, Czech Republic, Croatia, Hungary, Latvia, Lithuania, Poland, Slovak Republic

Indicator 2b): Development of a data repository: Czech Republic, Hungary, Latvia

Indicator 2c): Development of a metadata repository: Czech Republic, Hungary, Latvia, Lithuania, Poland, Slovak Republic

Indicators 3a), 3b) & 3c): Quality standards: Hungary, Poland, Slovak Republic

Indicator 5a): Data matching and data linkage: Czech Republic, Hungary, Poland, Slovak Republic

Indicator 5c): Editing: Latvia, Lithuania, Poland

Indicator 5e): Life-sign approach: Czech Republic, Lithuania

Indicator 5f): Alignment of measurements: Lithuania

Sources: Bulgaria, Czech Republic, Croatia, Hungary, Latvia, Lithuania, Poland, Slovak Republic

3.2 Good practice in a register-based census

3.2.1 Problem description

In the ESS, the following countries have already organised the Census 2010 with administrative data: Austria, Belgium, Denmark, Finland, Iceland, the Netherlands, Norway, Slovenia, and Sweden. As an example, Austria's approach of is described.

3.2.2 Method description

The legal background to the Austrian census is defined by Austrian law. Main points of this law are:

- The mandate to organise the register-based census of persons together with the census of places of work and the census of buildings and dwellings.
- An e-government law, which realises the one-way principle of data flow (see: Section 2.1, Indicator 1f). A system of "Branch-specific personal identification numbers" has been developed for persons. The system is administrated by the Austrian data security agency.
- The organisation of a test register-based census, together with a traditional census, for a sample of 0,3% of the population. This enabled evaluating the feasibility of a register-based census and led to a number of improvements in the detailed specification of the register based census's workflow.
- The obligation for administrative data holders to transfer the data necessary for the census to the Austrian NSI.

The most important data sources for the register-based census are eight administrative registers and data sources, which provide information for all required census variables (central population register, register of buildings and dwellings, register of educational attainment, school statistics and statistics of higher education, data of the Main Association of Austrian Social Security Institutions, tax registers, data from the Austrian Labour Market Service, and the statistical business register. Besides these primary registers, a number of other data sources are used to realise the "Principle of Redundancy" (i.e. if possible for most variables, information is available from more than one data source).

Based on the information of the various administrative sources, a coherent data model was developed, consisting of seven main tables: Persons, Families, Households, Objects, Flats, Places of work, Businesses.

For the integration of the different sources with information about persons, the branch-specific personal identification number was efficient in supporting record linkage of the different data sources of personal information. A critical point was that of determining, whether a person belongs to the census's population. The life-sign method was applied to this identification of persons. Altogether, 21 different life-sign indicators were used.

In the case of the census of businesses and working places, the main sources were the statistical business register and the register of agricultural enterprises. Additional sources were necessary in order to find all places of work because the statistical business register only lists enterprises with a turnover of over 10,000 € per year and with at least one employee.

For the census of buildings and dwellings, the corresponding register was the main source of information. This register includes a number of objects that do not correspond to the definition of categories used by the census. Hence a number of alignments were necessary, in the units.

For the editing of the variables in the integrated data, the principle of redundancy was very helpful because it allowed the formulation of edit rules, in the case of contradictory information in different registers. For example, the variable “family status” usually occurs in different registers. Only for 19% of all persons does this variable occur in a single register.

Because the different variables are often logical dependent from each other (for example, age has some implications on family status), in a first step a hierarchy for the imputation was defined. The methods for imputation were chosen in dependence of the variables. Hot deck imputation and regression imputation were the prevailing techniques for attributes of persons. Rather tricky was the imputation of relationships between persons living in one household. Based on explanatory variables for the relationship (e.g. age or gender), a random procedure was used selection of the possible relationships between the persons living in a household. Missing values of variables for the buildings and dwellings were estimated by statistical models for the missing variables in dependence of explanatory variables. In many cases the methods were defined by regression trees based on decision rules.

All final tables were produced using the integrated and coherent data model. In order to guarantee confidentiality of the information in the tables, target record swapping was used.

3.2.3 Summary use case

Country: Austria

Application domain: Population (CSA 1.1)

Typology of sources involved:

Type of source: administrative data

Type of aggregation level: micro-data

Type of product: register / data

Topics and indicators for good practice:

Indicator 1a): Legal act

Indicator 1b): Cooperation with data owners

Indicator 1c): Development of a privacy and confidentiality policy

Indicator 1f): Inventory of sources with respect to population

Indicator 1g): Inventory of sources with respect to variables

Indicator 2a): Organisation of data provision

Indicator 2c): Development of a metadata repository

Indicator 2d) Methods for pseudonymization

Indicator 2e) Definition of exchange formats

Indicators 3a), 3b) & 3c): Quality standards

Indicator 4d): Assessment of process design

Indicator 5a): Data matching and data linkage

Indicator 5c): Editing and imputation

Indicator 5e): Life-sign approach

Indicator 5f): Alignment of measurements

Indicator 5j): Output preparation

Sources: Registerzählung 2011 (Quality report on the register-based census, in German)

3.3 Good practice in feasibility studies

3.3.1 Problem description

At the beginning, changing statistical production towards an intensified use of administrative data means testing the feasibility of the new production methodology. The example of good practice documents a feasibility study of substituting some of the variables in Belgium's Labour Cost Survey (LCS) with administrative data. The main emphasis is on the presentation of desired output tables. Existing data from administrative sources are compared with adapted results from surveys.

3.3.2 Method description

Belgium's NSI carried out a feasibility study about the use of administrative data for the LCS. This investigation good practice in how to proceed with a specific application. First, in accordance with the general plan, a selection of the possible administrative data sources was carried out. Of the available sources, four were selected, which showed promising results. In a next step, the organisational and legal issues concerning the use of the data were settled, and an IT environment for accessing and uploading the data was defined.

The feasibility study itself was not based on the matching of statistical units but on a tabulation based on the administrative sources. The results of tabulating the variables of interest according to NACE showed that there are three types of variables: those variables that are already almost ready for the substitution, those that are not yet ready for substitution, and those, for which no administrative source exists. In the latter two groups, the availability of information in the administrative data was improved in the meantime and, from comparing the results, it can be concluded that, in this case, the replacement of the LCS data by administrative data is possible.

3.3.3 Summary of use case

Countries: Belgium

Application domain: Labour (CSA 1.2)

Typology of sources involved:

Type of source: administrative data, survey data

Type of aggregation level: macro-data

Type of product: register / data

Topics and indicators of good practice:

Indicator 1b): Cooperation agreement with data owners

Indicator 1c): Privacy and confidentiality policy

Indicator 1f): Inventory of sources with respect to coverage

Indicator 1g): Inventory of sources with respect to variables

Indicator 2b): Development of a data repository

Indicator 2c): Development of a metadata repository

Indicator 3a): Quality standards for input quality

Indicator 4a): Assessment of variables with respect to substitution

Indicator 4b): Assessment of the data integration process

Indicator 4c): Assessment of coverage

Indicator 4d): Assessment of process design

Indicator 5f): Alignment of measurements

Source: Feasibility of substitution in LCS Belgium

3.4 Good practice in the development of methodological knowledge: aligning variables

3.4.1 Problem description

Classifications play an important role in official statistics, structuring the underlying population into groups. In order to use the classifications available in administrative data as a substitution for the classifications of surveys, one has to be aware that the classifications contain errors and that an alignment (harmonisation) is necessary. This makes it necessary to develop methods for the alignment of the administrative classifications to the statistical needs. The method describes the possible use of the administrative data by defining a model that estimates the variable from existing survey variables. These estimates can be used further on in other applications.

3.4.2 Method description

The development of methods for the alignment of categorical variables depends on the availability and the quality of the information. The problem can be structured by answering the following questions:

- Are the statistical variables being used for the definition of the alignment procedure free of random measurement errors?
- Are the administrative variables being used for the definition of the alignment free of random measurement errors?
- Is more than one administrative variable available for the alignment?

In the case of more than one administrative variable providing information about the classification, good practices involve evaluating the quality of the variables providing from the different administrative sources and developing a decision rule according to that evaluation. In some cases such a rule may simply be a majority rule.

Within the ESS.VIP ADMIN WP 4, Statistics Netherlands has demonstrated good practice from a more methodological point of view by applying latent-class modelling in the context of the variable Home ownership. The model estimates the categorical variable by means of a latent-class model allowing for errors in the statistical variables as well as the administrative variables. None of the variables needs to be a gold standard.

3.4.3 Summary of the use case

Country: Netherlands

Domain: Population statistics (CSA 1.2)

Types of sources: administrative data / statistical data

Types of aggregation: micro-data

Types of products: register / survey

Topics and indicators for good practice:

Indicator 4d) Assessment of process design

Indicator 5f) Alignment of measurements

Indicator 5h) Supplementation of variables

Source: Estimating classification errors under edit-restrictions in combined register-survey data

3.5 Good practice in macro-integration

3.5.1 3.5.1 Problem description

Using administrative data frequently also requires the development of new methods for the production of statistics. One important methodological challenge is the principle of univalency. This means that the information in different tables for the same domain shows coherent information. If production begins with a survey, and all tabulation is done from this source, this condition is usually not a serious problem. In the case of producing data from administrative data, the situation is quite different because different tables are obtained from different administrative sources. One way to overcome such problems is to use estimation methods for macro-integration. The traditional method involves the use of calibration methods such as RAS or Stone's method. These have the disadvantage that the connection between the figures in the table and the original micro-data is lost. An alternative class of methods are those methods for univalent estimation, which alter the original micro-data either by weighting or by mass imputation, in such a way that all the final tables are coherent.

3.5.2 Method description

The method of repeated weighting can be applied through ad-hoc methods in such a way that a sequence of dependent tables is defined and that appropriate weights are sequentially assigned to the data being used to produce the tables. A more theoretically oriented approach involves the definition of a divide-and-conquer algorithm. The algorithm breaks down the problem of estimating a large consistent table set into a number of smaller sub-problems, which are preferably estimated independently. In each step, parts of a table set are estimated but, contrary to repeated weighting, these parts are not the same as individual tables, but rather a combination of cells from different tables. In this new approach, estimation problems seen in repeated weighting do not occur. A consistent table set can be obtained, if it is actually possible to define independent estimation problems. There is therefore no need for the determination of problem-specific solutions, as is often necessary with repeated weighting.

3.5.3 Summary of the use case

Country: Netherlands

Domain: Population (CSA 1.1), other domains appear possible

Types of sources: administrative or statistical

Types of aggregation: macro-data

Types of products: tabular data

Topics and indicators for good practice:

Indicator 6a): Innovation: Development of new methods

Source: Divide-and-conquer solutions for estimating large consistent table sets

3.6 Good practice in analytical investigations

3.6.1 Problem description

Besides using statistical products in tables that describe economic and social phenomena, there is an increasing interest in the use of the data for analytical purposes. Rather than taking data from a single survey, the interesting questions usually integrate data from several different sources. The data can be either statistical survey data or administrative data. A challenge in the integration is that data sources often only cover a subset of the population. This calls for of statistical matching methods, in order to build up a synthetic data set.

3.6.2 Method description

Within the ESS.VIP ADMIN project, WP 4 Statistics, Italy has demonstrated good practice in how to create a data set for micro-simulation based on statistical matching. The example shows how a survey of labour and income dynamics, personal income tax data, employment insurance claims, and surveys of household spending can be integrated into one synthetic data set that preserves the confidentiality of individual information. The data set includes detailed socio-demographic information, weekly employment histories, expenditure patterns and tax deductions. The method for matching is based on hot-deck imputation.

A drawback of the method is that the conditional independence assumption is often difficult to verify.

3.6.3 Summary of the use case

Country: Statistics Italy (Stat Canada)

Domain: Population statistics (CSA 1.2), other domains are possible

Types of sources: administrative or statistical

Types of aggregation: micro-data

Types of products: register / data / survey

Topics and indicators for good practice:

Indicator 5b): Statistical matching

Indicator 6a): Development of new methods for the use of administrative data

Indicator 7c): Analysis of new applications for administrative data (synthetic data for micro simulation)

Source: The creation of a social policy simulation database (Stat Canada)

4 Good practice in economic statistics

This chapter provides an overview of good practice in economic statistics. Section 4.1 introduces the application of administrative data to maintaining a business register, Section 4.2 the development of a statistical register of agricultural holdings and Section 4.3 the substitution of survey data in agricultural statistics. Section 4.4 describes the use of new data sources in energy statistics.

4.1 Good practice in building and maintaining a business register

4.1.1 Problem description

Most statistical registers are built from administrative registers. As a rule, a register will contain information on a complete group of units: the target population (e.g. persons, buildings, businesses). The register's scope is usually but not always defined by legislation such as that pertaining to tax or social security, for example. Register units are also defined by a precise set of rules (for instance: the resident population in a given country). The units' attributes are updated in line with the changes they undergo. The creation of a statistical register from an administrative register requires a number of transformations. The most important are the alignment of the administrative units and the administrative variables according to the concepts of statistical units and statistical variables (harmonisation of variables and units). Another issue is the coverage of the register according to the statistical population. Besides the creation of the statistical register the register update policy is the most important task in the register's maintenance, which also includes the administration of versions of the register.

4.1.2 Method description

The main source for the production of the Austrian statistical business register (URS) is the Austrian administrative business register (URV). The URV compiles a number of different sources and is maintained by Statistics Austria, a good practice that facilitates the setup and maintenance of the URS. Besides the URV, a number of other sources are used for the URS, in particular data sources from the Austrian chamber of commerce, the Austrian tax information system, the register of the Main Association of Austrian Social Security Institutions, data from external registers such as those providing information about educational institutions. Statistics Austria has defined a comprehensive list of possible sources and workflows for the integration of the various sources into the URS. A time schedule for data provision is defined for all data providers. Most of the data are processed automatically and data matching uses advanced text processing methods for the identification of keys. Manual editing is minimised by editing rules, which are regularly updated. The classification of enterprises according to NACE Rev.2 is of utmost importance. An interface for enterprises facilitates contacts about the classification with the business owners.

Updating and maintaining of the register is performed with different periodicities depending on the data sources. The basic information about the enterprises is updated daily, while information from the tax register and the Main Association of Austrian Social Security Institutions monthly. The main tasks in maintenance are: capturing of new units, changes in the structure of units such as the

addition of new local units, new accounts, classification changes (NACE Rev.2), actualisation of turnover and number of employees, and changes in the activity status (e.g. close-down). The relevant data sources are identified for all types of changes. The fact that the administrative register (URV) and the statistical register (URS) are maintained by the NSI facilitates coherence between the two information systems. Workflows are defined for the update processes, which support tracking of the update process. The register stores the history of all units.

4.1.3 Summary of the use case

Domain: Business statistics (CSA 2.3)

Country: Austria

Types of sources: administrative

Types of aggregation: micro-data

Types of products: register / data / survey / intermediate

Topics and indicators for good practice:

Indicator 5a): Data matching and data linkage

Indicator 5c) Editing & imputation

Indicator 5i) Update of statistical registers

Source: Statistisches Unternehmensregister 2013/2014 (Quality report of the Austrian business register, in German)

4.2 Good practice in the development of a register of agricultural holdings

4.2.1 Problem description

In agricultural statistics, numerous countries have carried out feasibility studies on the use administrative data, and particularly of IACS data, so as to contribute to the improvement of a statistical register of agricultural holdings. At a European level, the studies were developed in the context of a clear tendency to use IACS data, not only as tools for the system of direct payments to farmers, but also as information systems widely used for implementation of the Common Agricultural Policy and as further sources of agricultural data. The studies reveal good practices about how best to proceed with managing new sources and/or developing ones that are already used. The goal is to build up a statistical infrastructure similar to a statistical business register, in agriculture, the so-called “Agricultural Holdings Register”. The feasibility study of the Italian NSI reports a number of interesting particularities in the validation step. The operations carried out in other countries are substantially the same but every country emphasises the most important elements of the endeavours of its national statistical and institutional system. Conclusions are quite similar. All feasibility studies underline the importance of the relationships with the data owners, the collection and the analysis of metadata, the need for unique identifiers, the opportunities of starting

from the register, in creating a coherent system of agricultural statistics, and finally the actions that are necessary to ensure the register's sustainability and maintenance.

The Italian statistical institute's experience in developing a Statistical Farm Register (SFR) is a good practice because of the integrated use of administrative and statistical sources it makes, together with the integrated use of different methodologies. A lot of items managed by the Italian NSI are common to many other EU Member States, first and foremost because the administrative sources useful in this area are very common (the two main administrative sources for almost all Member States are the IACS together with a register of livestock).

4.2.2 Method description

The SFR is based on the integration of 10 administrative or statistical sources. The common key for linking all the input databases is the holder's ID. In Italy, the ID is the so-called fiscal code.

The main operative phases necessary for the implementation of a SFR (which are common to many Member States) are:

- Pre-analysis;
- Identification, acquisition and analysis of the input databases;
- Treatment and integration of the input databases;
- Identification and implementation of the rules of eligibility;
- Result validation.

Pre-analysis

In this phase, the results of the last Census of the Agriculture (2010, in Italy) and of the related Post Enumeration Survey (made in 2011 to evaluate the Census's quality) were analysed.

The main aim of this activity was to:

- Identify the statistical and administrative sources to use for the SFR;
- Define the methodology for building up the SFR.

The Census of the Agriculture utilised a frame built by the integration of 17 statistical and administrative sources, the so-called "Integrated Base of the Administrative Sources". The census results, in particular the outcomes of the interview, provided important information on the reliability of the sources used in the census list.

With regard to the methodology of building up the SFR, the effectiveness of the design used to define the census list was evaluated, in particular the eligibility rules implemented in selecting units from the Integrated Base of the Administrative Sources. More specifically, the over-coverage and the under-coverage of the census list were analysed by means of the census results as well as those of the post-enumeration survey.

The main reasons of the census's over-coverage are: different definitions between census and administrative sources (e.g. units out of the field of observation), errors in the sources (e.g. non-agricultural units), sources are not updated (e.g. definitively ceased unit activity, whole

activity rented or sold to other holding), the failure to link units among different sources or clusters (resulting in duplications).

The main reason for under-coverage are “new” units deriving from units that were included in the census list (e.g. because of dismemberments, mergers or total land/livestock transfers). These could have been established during the time between the reference date of the administrative source used and the census date, or they could already be existent but not included in the census list due to a failure in updating the administrative sources.

Identification, acquisition and analysis of the input databases

In the light of the analysis shortly described in the previous chapter, 10 administrative and statistical sources have been identified for the 2013 Statistical Farm Register.

All of these administrative sources are acquired by Istat, at regular intervals, according to formal agreements with the responsible authorities concerning data and their related metadata. In particular, to obtain IACS data, Istat implemented a specific legal act.

The metadata, that is the informative content, completeness, data quality, definitions, and the classification used were studied in each source, so as to gain better understanding of how to treat them before the subsequent phase of physical integration.

The main problems observed in the analysis of the sources are:

- Differences in the definitions of the units;
- Differences in the definitions of variables;
- Different classifications;
- Source not available on time or relating to a different period;
- Data susceptible to political or/and fiscal regulation changes causing spurious structural breaks;
- Supplier can adapt legal procedures on the base of their owns interests (the administrative declaration does not correspond to the operator’s “economic reality”);
- Problems of matching data (unit duplications);
- Consistency of the sources.

Treatment and integration of the input databases

All the input sources were treated, in order to integrate them coherently and according to the statistical definitions and classifications. In general, this operative phase considered the following topics:

- Standardisation of the personal ID, of territorial variables, of units of measurement, etc.;
- Identification of the holder/farm according to the statistical definition;
- Identification and classification of the variables related to holder/farm;
- Identification of duplicated or out-of-target units, ceased or with duplicated land;
- Quality checks on outliers and missing values;
- Metadata analysis.

Once all input sources had been pre-treated, the physical integration was implemented, using the ID holder (Fiscal code) as the linkage key.

Identification and implementation of the eligibility rules

To select the eligible units for the FR from the integrated database, Istat has agreed the same methodology used in 2009 to create the census list, improved in the light of the census results and the preliminary analysis carried out.

Result Validation

This phase consists of three kinds of activities:

- Macro- and micro-comparisons with other administrative and statistical sources;
- Ex-post analysis using the Census fieldwork results;
- Analysis of the results of the "Special sample survey", developed for the purpose of evaluating the quality of the 2013 Farm Register.

A specific phase of improvement in Italy is marked by the special survey designed to evaluate the methodologies used to define the SFR.

The first activity concerned the ex-post analysis of the Farm Register data in comparison with Census fieldwork results. The aim of this study has been to evaluate the reliability of the rules of eligibility implemented to select the farms to include in the SFR from the "Integrated Base of the Administrative Sources", in accordance with European legislation. That analysis provides information on the sample design of the "Special sample survey" developed for the purpose of evaluating the quality of the 2013 Farm Register.

This study classified the rules of eligibility into two set:

- Rules coherent with the 2010 agricultural census outcomes (high value of the indicator for the units belonging to eligible rules and low value of the indicator for the units belonging to non-eligible rules);
- Rules not coherent with the 2010 agricultural census outcomes (low value of the indicator for the units belonging to the eligible rules and high value of the indicator for the units belonging to non-eligible rules).

These results were used, when planning the sample design of the special survey for evaluating the quality of the 2013 Farm Register. In fact, this survey's reference population is represented mainly by the units belonging to the population in 2013, which is not coherent with the outcomes of the 2010 agricultural census.

To select the units, a one-stage sample design has been chosen. Strata are a combination of units with uncertain probability to be eligible or non-eligible, and some structural characteristics (NUTS Level 2, areas with flowers in the cadaster, units belonging to IACS, UAA < 1 ha, etc.). Moreover, in order to confirm that the validity of the rules is coherent with the 2010 agricultural census, the units belonging to these rules are also considered. They have been selected and stratified in the sample according to NUTS Level 2, UAA, and Livestock Units.

The final results of the Special survey are used to validate or to implement the rules of eligibility to apply in the final version for the SFR.

4.2.3 Summary of the use case

Countries: Italy, Poland, Hungary, Greece

Domain: Agriculture (CSA 2.4.1)

Types of sources: administrative / statistical

Types of aggregation: micro-data

Types of products: register / data / survey

Topics and indicators for good practice:

Indicator 1f): Inventory of sources with respect to population coverage

Indicator 1g): Inventory of sources with respect to variable description

Indicator 2c): Development of a metadata repository for administrative data

Indicator 4a): Assessment of the variables with respect to their use as substitution

Indicator 4b): Assessment of the data integration process

Indicator 4c): Assessment of the coverage of the sources

Indicator 4d): Assessment of process design

Indicator 5d): Alignment of statistical units

Indicator 5d): Alignment of measurements

Indicator 5g): Substitution of variables

Source: (currently not available on the CROS portal)

4.3 Good practice in the substitution of variables in agricultural statistics

4.3.1 Problem description

In order to reduce response burden, the Austrian NSI carried out a feasibility study of the possible use of administrative data or other statistical surveys to substitute the variables related to the labour force in the Farm Structure Survey (FSS). A growing statistical burden affects the quality of the information provided by respondents and, in the near future, NSIs will stand to choose among a growing amount of data with a decreasing quality versus somewhat reduced statistics of an acceptable quality, obtained using administrative sources or re-using existing statistical surveys. The study shows good practice in the method of how to proceed with rationalising the statistical

system by using the information contained in already existing sources, for the total or partial substitution of statistical surveys or to complement them.

4.3.2 Method description

Firstly, statistical surveys and administrative sources are identified, including variables related to the Farm Labour Force. These occur in four primary statistical collections: EU-SILC, Micro Census, a register derived from administrative sources, and the Social Security data from the Main Association of Austrian Social Security Organisations.

The second step included analysing the metadata (definitions, classifications, questionnaires, coverages) and comparing them with the FSS and the other statistical surveys, as well as comparing the derived register and the Social Security data. The comparison of macro- and micro-data was carried out in order to obtain a clear idea of the differences in results that provide from the different data, and of the amount and the possibility of linking the micro-data from different sources.

Finally, the main data users' informational needs from the FSS were analyzed. They are those of international users (EU Commission, OECD, and FAO), of Austrian Institutions and of Statistics Austria internal users. The scope was to understand, whether it is possible, and to which extent the information on the Farm Labour Force can change without causing effective problems to the users.

The analysis showed that only the HV Social Security data was able to provide some variables to the FSS but also that there are some other key characteristics that cannot be determined from these data. On that basis, the study also provides different strategies for using the administrative data, highlighting pros and cons: (i) exclusive use of the administrative data with regard to recording the farm labour force, reducing the burden on the FSS questionnaire; (ii) the questionnaire is completed beforehand, using administrative data; (iii) use of the administrative data to support the interviewers during the course of telephone surveying, ensuring targeted questioning, (iv) use of the administrative data for the purpose of plausibility checks.

4.3.3 Summary of the use case

Country: Austria

Domain: Labour (CSA 1.2), Agriculture (CSA 2.4.1),

Types of sources: administrative / statistical

Types of aggregation: micro-data

Types of products: register / survey

Topics and indicators for good practice:

Indicator 4a): Assessment of the variables with respect to their use as substitution

Indicator 4b): Assessment of the data integration process

Indicator 4c): Assessment of the coverage of the sources

Indicator 4d): Assessment of process design

Indicator 5g): Substitution of variables

Source: (currently not available on the CROS portal)

4.4 Good practice in energy statistics

4.4.1 Problem description

Data concerning the energy consumption of households are usually captured by surveys, sometimes as a specific module of other surveys. Due to the fact that energy providers are usually institutions subordinated to or regulated by the government (or, at least, institutions of public interest), it appears possible to establish contacts with such data providers, in order to obtain information about the energy consumption of households from the providers. Statistics Sweden started a pilot project, which is showing positive signs in the area of an active search for new data sources (Topic 1a). A template was designed for a contract between the major network owner and the statistical office, which defines a win-win situation for both participants.

4.4.2 Method description

A common Swedish hub for energy data was designed so as to make the data available for statistical purposes. The hub stores micro-data with the following types of information:

- Installation data for every apartment;
- Customer data;
- Supplier Exchanges;
- Daily electrical data.

Differing definitions of statistical units, legal aspects and methodological issues pose major challenges to using the data (e.g. the energy delivery point may differ from the owner's address). These problems can restrict the usage of the data. The data display characteristics of big data but it turned out that, for data cleaning and data preparation, the traditional statistical methods for editing, imputation and integration are applicable.

The negations with the data provider were rather cumbersome but one can learn from these experiences that the use of the data is a promising option.

4.4.3 Summary of the use case

Country: Sweden

Domain: Energy statistics (CSA 2.4.2)

Types of sources: administrative

Types of aggregation: micro-data

Types of products: (big) data

Topics and indicators for good practice:

Indicator 1b): Cooperation agreement with data owners

Indicator 1d): Active search for exploiting new administrative data sources

Indicators 1f), 1g): Inventory of sources

Indicator 2b) Development of a data repository for administrative data

Indicator 4a): Assessment of the variables with respect to their use as substitution

Indicator 4b): Assessment of the data integration process

Indicator 4c): Assessment of the coverage of the sources

Indicator 4d): Assessment of process design

Source: Improvement of the use of administrative sources

5 Conclusions

5.1 General findings

A core element of good practice in using administrative sources for statistical production is the development of a close cooperation with the data owners. That cooperation should be supported by a paragraph in the national legal act for statistics. Provision of administrative data for statistical purposes should be part of legislation. For the handling of all issues of the privacy and confidentiality of statistical data, the national statistical act should be connected with the national e-government regulations. In the case of special surveys and census applications, it is useful to formulate specific regulations for the necessary provision of data.

Considering the internal organisation of the NSI, the production of multisource statistics should be promoted. This can be seen in connection with the general transition of statistical production from a stove-pipe principle towards a more integrated production. An important pillar of such a new view on production involves the building of a statistical data warehouse. Explicit strategies and methods for the ETL process must be developed. Most NSIs are at the beginning of such a development.

In connection with the statistical data warehouse, the development of a metadata repository is of utmost importance. The use of administrative data defines a number of new requirements from the metadata repository. The metadata repository is essential for the specification of the quality standards of multisource statistics, which encompass input quality, process quality and output quality.

With respect to the statistical production of multisource statistics, good practice is to rely on the Generic Statistical Business Process Model (GSBPM), which defines a useful framework for the organisation of statistical production. However, in order to accomplish statistical production in its details, it is necessary to test the new production system by means of extensive feasibility studies, covering all aspects going from the definition of populations and variables up to the detailed specifications of workflows of GSBPM sub-processes.

5.2 Methodological findings

For a number of the processes and in different domains, one can find examples of good practices in the ESS. In connection with those examples, the following new needs arose:

5.2.1 Data collection

Administrative data sources often use different units, which are not exactly the same as the intended statistical units. One needs to define a data model describing the relation between the administrative units and the statistical units. These methods are known under the term alignment (harmonisation) of units. Moreover, one must consider the coverage of the intended statistical population. Branch-specific methods need to be developed, to use when making decisions in coverage issues, such as the life-sign approach for individuals in the census.

5.2.2 Integration of data

When using administrative data, data-integration methods are much more important than they are in traditional survey production. Knowledge and skills in data linkage and data matching are of utmost importance. Besides the development of knowledge and skills inside the NSIs, the transfer of knowledge of good practices in other NSIs is important. Providing the data owners with feedback can often be conducive to a data provision that makes integration easier (e.g. use of appropriate identifiers).

5.2.3 Classify & code

Administrative data usually do not have sampling error, but measurement error may arise, consisting of different kinds of bias as well as a random error component. For the purpose of the alignment (harmonisation) of the measurements, complex statistical models can prove useful. An example is the use of latent-class models, in the case of categorical variables. The application of such methods requires feasibility studies, which test the model by comparing the results with existing survey data. This implies an additional effort in the transition phase to register-based statistics.

5.2.4 Editing & imputation

The building blocks for editing and imputation are the same as in traditional survey-based production. The main difference is the definition of special workflows for the administrative data. The editing of administrative data often involves very specific edit rules. These rules are frequently determined by how the administrative data is maintained. Good practice is the principle of redundancy, meaning that the values of the administrative variables should be checked using the values given by the different administrative sources. With respect to imputation one must be aware that, due to different concepts of population in administrative data, it may be necessary impute on specific subgroups of the population. In both cases, the design of a workflow (including the definition of a hierarchy for editing and imputation) is of utmost importance.

5.2.5 Derivation of new variables

The derivation of new variables plays a major role in the case of administrative sources and the process is rather similar to the case of variable alignment (harmonisation). In this case, advanced statistical methods such as structural equation modelling can also be applied. As with the alignment of measurements, feasibility studies are necessary to verify the model using comparable administrative and survey data.

5.2.6 Finalisation of data files

Building a data corpus with all the results obtained from the administrative sources requires carefully taking into account the relationships between different sources. Usually, it is not simply a data file but a database with interrelated tables.

5.2.7 Finalisation of outputs

The principle of univalency, which means that the different tables obtained from the administrative sources should provide coherent information, plays a much greater role in multi-source statistics than it does in traditional survey production. This is an area for the development of new methods.

Summing up, one can say that the ESS is on the way towards a more intensive use of administrative sources. This implies a certain amount of effort in learning the new production methods, in carrying out feasibility studies, in innovation and in adapting existing methods. The speed of transformation depends on the NSIs' background as well as on the nature specific

domains. Cooperation in the new development and in the transfer of knowledge between Eurostat projects (ESSnets, ESS.VIP) are very helpful in the process.

6 Appendix: Template for review of actual use of administrative sources

Attributes, category 1: Identifiers

Identifying variables		
Name	**Values**	**Label / explanation**
Source_ID		Number for identification
	integer	running number for sources inside the case study
Case_ID		Number of application
	integer	running number for application
Case_Country		Country of application
	string	Name of the country
Case_Own		Owner of the case study
	NSI	National statistical institute
	government	Governmental institution and Central Bank
	gov_reg	Institution subordinated or regulated by government
	pub_interest	Institution of Public interest
	private	Private Institution
	other	Other owners
	??	unknown
Case_OwnName		Name of the owner
	string	
	??	unknown
Case_Year		Validity of the documentation
	integer/string	Publication year of documentation / time period of validity
	??	Unknown
Out_Name		Name of the product
	string	

Remark:
Usually the case owner is a NSI, but sometimes it may happen that another institution is the owner of the case.

Attributes, category 2: Characterisation of output production

Characterisation of the output production		
Name	**Values**	**Label / explanation**
Out_Product		Type of the output production
	census	Statistical Census production
	register	Register production
	data	Data compilation
	survey	Statistical Survey production
	intermediate	Intermediate statistical production
	other	Other statistical form of production
	??	Unknown
Type_Product		Usage type of the statistical product
	feasibility	The product is a feasibility study
	standard	Output is a standard statistical product to be disseminated
	analytical	Output is used for analytical purposes
	intermediate	Intermediate statistical product
	??	Unknown
Use_Register		When the case study is a register application, specify
	build	Used for building a register
	maintenance	Used for updating and maintenance of register
	NA	Case study is not a register application
	??	Unknown
Use_Frame		Use of the product as frame for sampling
	yes	
	NA	In this case Use_Register = NA
	??	Unknown
Out_Dom_L1		Output domain taken from CSA level 1
	integer	Number of domain level 1
Out_Dom_L2		Output domain taken from CSA level 2
	integer	Number of domain level 2 or higher from CSA
Out_TargetUnit		Name of the statistical target unit
	string	Name of the unit
Out_Population		Reference population
	string	
Out_Period		Periodicity of the product
	more	More than annual /occasionally
	annual	Annual product
	quarterly	Quarterly product
	monthly	Monthly product
	continuous	Periodicity below monthly
	??	Unknown
Use_GSBPM		Characterisation of GSBPM usage
	direct	Direct tabulation
	subst_suppl	Substitution or supplementation of survey
Out_Flag		Flag for availability of documentation
	1	yes
	0	no
Out_Doc		Link to the documentation of the production
	link	

Attributes, category 3: Characterisation of input data sources

Characterisation of the input sources		
Name	**Values**	**Label / explanation**
In_Name		Name of the source
	string	
In_Sort		The type of the source
	statistic	Statistical source
	admin	Administrative source
In_Dom_L1		Source domain taken from CSA level 1
	integer	Number of domain level 1
In_Dom_L2		Source domain taken from CSA level 2
	integer	Number of domain level 2 or higher from CSA
In_Own		Owner of the source
	NSI	National statistical institute
	government	Governmental institution and Central Bank
	gov_reg	Institution subordinated or regulated by government
	pub_interest	Institution of Public interest
	private	Private institution
	other	Other owners
	??	Unknown
In_OwnName		Name of the owner
	string	
In_Prod		Production of the source
	census	Statistical Census production
	register	Register production
	data	Data compilation
	survey	Statistical Survey production
	intermediate	Intermediate statistical product
	other	Other statistical form of production
	occasionally	Source is produced occasionally in irregular periods
	??	Unknown
In_Gran		Granularity of the source as available in NSI
	micro	Micro-data
	macro	Macro-data
	time	Time series (macro)
	??	Unknown
In_Unit		Main source unit used for case study
	string	Name of the unit
	??	Unknown
In_Population		Description of the reference population
	string	
	??	Unknown
In_Period		Periodicity of the source when collected by NSI
	more	More than annual / occasionally
	annual	Annual source
	quarterly	Quarterly source
	monthly	Monthly source
	continuous	Periodicity below monthly
	??	Unknown
In_Frame		Frame for the case study
	yes	Source is the frame for the output of case study
	NA	All other cases
	??	Unknown

Remark:

*In the case of **In_Sort** = admin **In_Prod** has only register, data, or other as admissible values.*

Attributes category 4: Characterisation of the organisation of the production

	Characterisation of the organisation	
Name	**Values**	**Label / explanation**
Org_Legal		Legal arrangements for using the source
	yes	
	no	
	??	Unknown
Org_Provision		Organisation of data provision - relationship with the owner
	well_org	Existence of an administrative information system in NSI
	org	Specialised team for dealing with administrative data
	ad_hoc	Ad hoc contact with the owner
	??	Unknown
Org_Exchange		Organisation of exchange between NSI and owner/owners
	central	Centralised data exchange in the office
	decentral	Decentralised data exchange in the office
	mixed	Both centralised and decentralised
	??	Unknown
Org_Format		Format of data exchange
	EDI	Electronic data interchange
	FTP	Access by FTP
	table	Table Format (xlsx, csv,…)
	more	When mixed formats are used
	other	Other format
	??	Unknown
Org_Doc1		Documentation of the source
	yes	Existence and updating of metadata
	no	
	??	Unknown
Org_Doc2		Check for stability of definitions/classifications over time
	yes	
	no	
	??	Unknown
Org_Doc3		Statistical and technical checks done and documented
	yes	
	no	
	??	Unknown
Org_Privacy		Privacy and security arrangements
	yes	
	no	
	??	Unknown

Data file

The Excel file SC094_D3_DataTable.xlsx contains the data table and a pivot table which shows for demonstration the different domains of the examples. The values “??” are replaced by “unknown”. Note also that not all possible values occur for all attributes in the table.

The detailed descriptions in the Deliverable are prototypical examples of good practices in the different domains.