IMPROVING THE METHODOLOGY FOR USING ADMINISTRATIVE DATA IN AN AGRICULTURAL STATISTICS SYSTEM

Administrative Data and the Statistical Programmes of Developed Countries

October 2015
Working Paper No. 3
Global Strategy Working Papers

Global Strategy Working Papers present intermediary research outputs (e.g. literature reviews, gap analyses etc.) that contribute to the development of Technical Reports.

Technical Reports may contain high-level technical content and consolidate intermediary research products. They are reviewed by the Scientific Advisory Committee (SAC) and by peers prior to publication.

As the review process of Technical Reports may take several months, Working Papers are intended to share research results that are in high demand and should be made available at an earlier date and stage. They are reviewed by the Global Office and may undergo additional peer review before or during dedicated expert meetings.

The opinions expressed and the arguments employed herein do not necessarily reflect the official views of Global Strategy, but represent the author’s view at this intermediate stage. The publication of this document has been authorized by the Global Office. Comments are welcome and may be sent to global.strategy@gsars.org.
Improving the Methodology for Using Administrative Data in an Agricultural Statistics System

Technical Report 2: Administrative Data and the Statistical Programmes of Developed Countries
Drafted By

Emily Berg and Jie Li

and

Center for Survey Statistics and Methodology
Iowa State University
United States of America
Executive Summary

Data collected from sources other than surveys and censuses are used extensively by the statistical offices of developed countries. Such data can maximize the consistency of published cross-sector figures and help to align survey data with information from non-statistical sources.

A variety of mechanisms are involved. Administrative data can be used directly for a statistical product if the coverage of target populations and concepts is complete and data on the administrative source are accurate. If micro-data are available, administrative data sources often contribute to sampling frames that are later used for surveys and censuses. Aggregate totals from an administrative source can be used as calibration controls. Differences between administrative and survey data can highlight possibilities for improvements in the quality of both.

Many of the quality dimensions traditionally used to evaluate survey and census data also apply to administrative data. The appropriate form of quality analysis depends on the intended use of the administrative resource. If a database is to be used in direct tabulation, high coverage and low measurement error are essential. If the administrative source is to be integrated with other databases at the micro-level, then sound identifying variables are essential. If the primary role of the administrative data is to reduce variances in estimators through calibration or small-area estimation, high correlations between variables on the administrative source and those on the survey are valuable.

Because the collection of administrative data is usually outside the control of statistical offices, administrative databases may not meet the standards required for statistical purposes, and data judged adequate for administrative purpose may not be reliable enough for statistical uses. An administrative database will not cover a whole population if the collection process is only relevant for a subset. Inconsistencies in definitions and coding systems arise from misalignment between administrative and statistical objectives and as lack of coordination among agencies. Public scepticism may arise if there are concerns about confidentiality.

To overcome these challenges, statistical offices should develop relationships with administrative agencies and understand their data source. Combining several data sources can improve coverage and enable analysis of measurement error. Manual review of embedded samples can detect and correct errors, and automated editing procedures can improve data quality. Model-based methods for record-linkage have been developed to handle inconsistencies or incompleteness in identifying variables. Software tools for probabilistic record linkage can help in addressing technical challenges associated with managing
large quantities of data and are useful for combining individual administrative registers into register systems for statistical use.
# Table of Contents

Executive Summary ................................................................. i
Background Information .......................................................... v
Acronyms and Abbreviations ................................................... vi

1. Introduction .............................................................................. 1
   1.1. Defining Administrative Data ............................................ 1
   1.2. Why Use Administrative Data ? ....................................... 5
   1.3. What Are the Challenges ? ................................................ 6
   1.4. Overview of Uses Around the World ................................. 6
   1.5. Outline ........................................................................... 8

2. Sources of Administrative Data for Agriculture ..................... 9
   2.1. Soils Information .............................................................. 10
   2.2. Crop Insurance and Subsidiary Programs .......................... 10
   2.3. Land Registration and Cadastral Records .......................... 11
   2.4. Government regulations and Monitoring Programs ............ 11
   2.5. Private/Non-Governmental Organizations and Sources of Operations ......................................................... 13
   2.6. Reporting Systems .......................................................... 13
   2.7. Taxation Data .................................................................. 14
   2.8. Connections to Core Data Items ....................................... 15

3. Benefits of Administrative Data ............................................... 16
   3.1. Cost Savings ................................................................. 17
   3.2. Reduced Respondent Burden .......................................... 18
   3.3. Improvements in Micro-Data Quality ............................... 19
   3.4. Improvements in Efficiency of Macro-Level Estimators ...... 20
   3.5. Small Area Statistics ....................................................... 21
   3.6. Timeliness ..................................................................... 22

4. Uses of Administrative Data ................................................... 24
   4.1. Direct Tabulation ........................................................... 24
   4.2. Frame Construction/Improvement .................................... 25
   4.3. Survey Design ............................................................... 26
   4.4. Model- Assisted Calibration Estimators ............................ 27
   4.5. Non-response Adjustments and Imputation ...................... 29
   4.6. Model Based Small Area Estimation and Forecasting ........ 31
   4.7. Cut-off Surveys ............................................................. 34
   4.8. Uses in Data Collection .................................................. 36
5. Defining the Quality of Administrative Data .................................. 37

5.1. Quality Dimensions ..................................................................... 39
5.2. Fitness for Use: Evaluating Quality with Respect to Intended Use ............................................................... 40

6. Challenges and Limitations of Administrative Data .................. 43

6.1. Lack of Standardization between Multiple Data Sources ...... 43
6.2. Lack of Control ........................................................................... 45
6.3. Coverage of Population and Concepts ..................................... 47
6.4. Computational and Methodological Challenges ..................... 48
6.5. Confidentiality and Public Perception ...................................... 50
6.6. Methodological Challenges in the Construction of the Austrian Business Register ................................... 50

7. Mechanisms to Improve the Quality of Estimates Derived from Administrative Data .......................................... 52

7.1. Proactive Involvement in Administrative Processes and Engagement with the Public ..................................... 52
7.2. Combining Multiple Data Sources ........................................... 56
7.3. Audits and Embedded Samples .................................................. 64
7.4. Systems and best Practices for Quality Assurance and Control. 64
7.5. Methodological and Technical Tools for Data Integration ...... 68

8. Probabilistic Record Linkage ......................................................... 72

8.1. Steps Involved in Data Cleaning and Record Linkage .......... 73
8.2. Software ..................................................................................... 74

9. Registers ....................................................................................... 80

9.1. What is a Register? Administrative and Statistical Registers ... 80
9.2. Combining Administrative Registers .......................................... 81
9.3. Statistical Farm Register in the EU Acceding Countries ...... 82

10. Summary, New Directions and Lessons for Developing Countries... 84

10.1. New Directions ........................................................................ 87
10.2. Lessons for Developing Countries ............................................ 88

References ....................................................................................... 94
Background Information

Task 2: Review
The objective of Technical Report 2 is to review and analyse country experiences and research in developed countries on the use of administrative sources for producing agricultural data and lessons for developing countries.


Weeks 5–10: Analysis of experiences in developed countries: The proceedings of workshops such as ESSnet ISAD (2008) and the United Nations 2011 handbook on the use of administrative data for statistical purposes will provide a basis for this investigation. Each institution will use its own experiences with integrating administrative and survey data through collaborative agreements with the USE will also inform this analysis.
### Acronyms and Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABS</td>
<td>Australian Bureau of Statistics</td>
</tr>
<tr>
<td>ADSAS</td>
<td>Administrative Data System for Agricultural Statistics</td>
</tr>
<tr>
<td>APHIS</td>
<td>Animal Public Health Information System</td>
</tr>
<tr>
<td>CEAP</td>
<td>Conservation Effects Assessment Project</td>
</tr>
<tr>
<td>CTS</td>
<td>cattle tracing system</td>
</tr>
<tr>
<td>FAO</td>
<td>Food and Agriculture Organization of the United Nations</td>
</tr>
<tr>
<td>FSA</td>
<td>Farm Services Agency</td>
</tr>
<tr>
<td>GIS</td>
<td>geographic information system</td>
</tr>
<tr>
<td>IACS</td>
<td>Integrated Agricultural Control System</td>
</tr>
<tr>
<td>MDG</td>
<td>Millennium Development Goal</td>
</tr>
<tr>
<td>NASS</td>
<td>National Agricultural Statistics Service</td>
</tr>
<tr>
<td>NMR</td>
<td>National Methodological Report</td>
</tr>
<tr>
<td>NRCS</td>
<td>Natural Resources Conservation Service</td>
</tr>
<tr>
<td>NRI</td>
<td>National Resources Inventory</td>
</tr>
<tr>
<td>OECD</td>
<td>Organization for Economic Cooperation and Development</td>
</tr>
<tr>
<td>PARIS21</td>
<td>Partnership in Statistics for Development in the 21st Century</td>
</tr>
<tr>
<td>STES</td>
<td>short-term economic statistics</td>
</tr>
<tr>
<td>USDA</td>
<td>United States Department of Agriculture</td>
</tr>
<tr>
<td>VAT</td>
<td>Value-Added Tax</td>
</tr>
</tbody>
</table>
Introduction

Data collected from sources other than surveys and censuses are used by statistical agencies in many countries. Increases in computer resources and government budget cuts have led statistical offices to explore new options for generating official statistics from administrative data (Karr, 2012; Smith et al., 2013). Tabulations of register systems have augmented data collections for censuses in Austria and Finland (Berka et al., 2012; United Nations, 2011), and Statistics Canada decided to use taxation data in place of “detailed expense questions” for most farms in the 2016 Census of Agriculture (Smith et al., 2013). Totals constructed from administrative data have been used as auxiliary information to improve the efficiency of survey estimators (Nusser and Goebel, 1997; Slud, 2004; Sarndal and Lundstrom, 2005). Academic and government institutions have devoted journal articles, as described for example in Karr 2012, and workshops such as the Organization for Economic Cooperation and Development (OECD) 2015 to facilitate discourse on the use of administrative data to produce official statistics.

1.1. DEFINING ADMINISTRATIVE DATA

Given the diversity of types and uses of data from sources other than surveys or censuses, the concept of administrative data is difficult to define. Our remarks in this section expand on the discussion of defining administrative data in section 2.1 of Technical Report 1.

Definitions of administrative data focus on information arising as a by-product of the administration of government monitoring programmes and regulations (United Nations, 2011) such as taxation, farm-subsidy programmes and social security benefits, as in Definition 1.1.1:

1.1.1. DEFINITION

Administrative sources are data holdings containing information collected by public sector offices to meet demands of government regulations.
This definition is consistent with the concept adopted by the short-term economic statistics (STES) taskforce (OECD, 2015), which assigns four characteristics to administrative data:

- the agent that supplies the data to the statistical agency and the unit to which the data relate are usually different, in contrast to most statistical surveys;
- the data were originally collected for a definite non-statistical purpose that might affect the treatment of the source unit;
- complete coverage of the target population is the aim;
- control of the methods by which the administrative data are collected and processed rests with the administrative agency. (OECD, 2015)

The STES taskforce adds: “...it is normal to accept (and expect) that the administrative agency will be a government unit that is responsible for implementing an administrative regulation” (OECD, 2015). Because Definition 1.1.1 is specific to government programmes, it excludes potentially valuable sources of data of an administrative in nature, and it omits data obtained from private institutions and non-governmental organizations.

To expand the class of administrative data sources, the Conference of European Statisticians adopts a definition that includes: “... data collected by sources external to statistical offices” (United Nations, 2011), and Nordbotten (2008) defines administrative data as information “… collected primarily for non-statistical purposes, and adopted for producing statistics.” These definitions encompass non-traditional sources such as satellite data, private information on commercial transactions and meteorological data. The United Nations 2011 handbook formalizes this in the following:

1.1.2. DEFINITION

Administrative sources are data holdings containing information that is not primarily collected for statistical purposes.

Definition 1.1.2. accomplishes the objective of generality, but it may be inadequate for two reasons. First, a reviewer has noted that it is ambiguous as to whether information from expert reporters is “administrative”. Such data are relevant for agricultural statistics, for example, because they may be collected through processes in which a sample of qualified informants gives assessments of crop or livestock characteristics (see section 2.6). But because the reporters are not necessarily a probability sample and because their assessments may be subjective, such reporting systems are criticized for lacking the rigour of sample
surveys. Statistical offices nonetheless use data gathered through reporting systems to create estimates and populate databases, so a definition of administrative data that encompasses information from reporting systems is appropriate. Second, Definition 1.1.2. may be too general to be useful. It includes, for example, trade figures and satellite data, which are not of primary interest for our work.

An alternative definition identifies administrative data with the distinguishing characteristic that “… the specific identity of the respondent or data source is central to the use of the data” (Sen, undated). Taxation data, for example, are collected to determine the tax payable by each individual; but in the context of statistical surveys or censuses, groups of individuals are randomly or deterministically selected from populations of interest, and aggregate characteristics are of interest. As Sen explains: “the distinction between statistical and administrative data is based on the focus on the general against the particular”, as reflected in Definition 1.1.3:

1.1.3. DEFINITION
Administrative sources are data holdings containing information collected for a purpose that is intrinsically linked to the identity of the individual unit in the population.

This definition includes the administrative sources in Definition 1.1.1. and also sources that it would not classify as administrative. Definition 1.1.3. would include electronic data on credit card transactions, unlike Definition 1.1.1, and it is also an improvement on Definition 1.1.2 because remotely sensed information is excluded. But Definition 1.1.3 is still inadequate for our purposes because it excludes certain sources of information that could provide “administrative” data on agriculture in developing countries. In particular, we want to include aggregate statistics from reporting systems or private corporations in the class of “administrative data” even though they are not “administrative” according to Definition 1.1.3.

In framing Definition 1.1.4, we observe that many of the administrative sources mentioned above are selective processes that define specific populations. Taxation data, for example, results from the process of gathering taxes and applies to the population of taxpayers; information from farm-subsidy programmes is produced from participation in such schemes and applies to the population of participating farms. With regard to reporting systems, the information provided will often depend on the professional experience of the reporter – a selection process that differs fundamentally from sample surveys, in which the target population is defined in advance and the sample of
participating units is randomly selected. Wallgren and Wallgren (2010) describe this distinction between administrative and survey data in a context in which administrative data are based in “registers”:

To design a register-based survey is a different statistical task than designing a sample survey. In the case of sample surveys, the first step is to determine the population and which parameters are to be estimated for which domains of study. This in turn determines the character of the survey with regard to sampling design and estimation. Thus the definition of population and parameters comes first, then data collection. As a rule one survey at a time is considered, with a limited number of parameters.

In the case of a register-based survey a different approach is taken, since data have already been collected and are available in different administrative registers that are not tailored for a particular statistical application. With the aid of available registers, a selection is made of objects and variables that are relevant to the issue addressed by the register-based survey. It may be that, on the basis of available registers, new variables – and possibly new objects as well – have to be derived. Thus, the data come first, and then the determination of population, parameters and domains of study. Sample errors do not restrict the possibilities of selecting domains of study for the coming analysis and reporting of results.

With this in mind we offer Definition 1.1.4:

**1.1.4. DEFINITION**

Administrative sources are data holdings containing information such that the processes governing the data collection are selective or define specific populations.

The following sections show that selectivity is a distinguishing characteristic underlying the benefits and challenges of using administrative data. Because administrative data arise naturally through selective processes, the costs of collection and maintenance are minimized. – but in using administrative data a statistical office sacrifices control over the definition of the population of interest and the data-collection protocols. In principle, the administrative file contains information based on all units in the defined population, but in reality the data is likely to be incomplete as a result of reporting errors or refusals to participate. In an ideal situation the population defined by the administrative concept would align with the population of interest to the statistical agency, but in practice these populations may differ and the effects of the selection process on the estimates of interest must be evaluated.
A drawback of Definition 1.1.4 is lack of specificity. Although it characterizes administrative data and implicitly covers data collected by government offices for specific administrative purposes (i.e., taxation, regulation, registration, and social assistance), Definition 1.1.4 does not identify such sources explicitly. We therefore adopt the following, which is based on the Administrative Data Liaison Service (2015), as the definition of administrative data for this research:

1.1.5. DEFINITION

Administrative sources are data holdings containing “information collected primarily for administrative (not statistical) purposes by government departments and other organizations usually during the delivery of a service or for the purpose of registration, record-keeping, or documentation of a transaction.” In the context of agriculture these sources include: i) registers, for example of farms, livestock, farmers, farmers’ associations and parastatal and other institutions that generate administrative data for commercial or cash crops; ii) transaction data such as imports and exports and data generated during transactions at customs posts; and iii) routine data collected by agricultural extension workers.

Definition 1.1.5 identifies the types of administrative data of interest for our work. It covers data from private corporations and reporting systems and includes examples of the types of administrative data relevant to agricultural statistics. Nordbotten (2008) provides examples of administrative data used to produce statistics for various sectors:

- taxation data used for income statistics;
- registration for public services used for social statistics;
- unemployment registration used for employment statistics;
- medical registration data used for health-related statistics; and
- registration in schools used for education statistics.

These relate primarily to population and social statistics rather than agricultural statistics. Section 2 provides examples of administrative data sources relevant to the estimation of core agricultural data items identified by the Global Strategy for Improving Agricultural Statistics.

1.2. WHY USE ADMINISTRATIVE DATA?

Technical Report 1 explains the benefits of using administrative data as discussed in United Nations (2011). Section 3 extends the discussion to include material from Carfagna and Carfagna (2010), Brackstone (1987) and experiences from statistical offices in developed nations. Two of the primary
incentives for using administrative data are low costs and a reduced burden on respondents, resulting because statistical offices do not need to contact units of interest for information that exists in administrative files (Carfagna and Carfagna, 2010). The use of administrative data can also improve the quality of statistical products, for example by reducing errors related to measurement or missing data (Australian Bureau of Statistics, 2011). Consolini (2008) discusses the use of tax data to reduce measurement error arising from under-reporting in the Survey of Income and Living Conditions in the European Union. The integration of several administrative databases can improve consistency across sectors (Brackstone, 1987), and where administrative data are updated and released on a frequent basis they can show changes and trends (Australian Bureau of Statistics, 2011).

1.3. WHAT ARE THE CHALLENGES?

The potential benefits of using administrative data are not automatic, and the statistician will have to find ways to overcome deficiencies in quality. Several challenges are discussed in section 4 of Technical Report 1. A fundamental reason why the quality of administrative data may not be suitable for statistical purposes is that the objectives underlying the creation of the administrative source differ from the aims of the statistical office (Brackstone, 1987). This can limit standardization and lead to under-coverage, misrepresentation of the population of interest, inconsistencies over time and errors resulting from non-participation (Wallgren and Wallgren, 2010; Carfagna and Carfagna, 2010) (see sections 6 and 7).

1.4. OVERVIEW OF USES AROUND THE WORLD

Administrative data are used in various ways to improve the collection of agricultural statistics and related statistical products. Data of sufficiently high quality are often tabulated directly to construct official statistics, and even quality is inadequate for independent use, administrative data are used in conjunction with surveys and censuses. Many of these applications of administrative data in agricultural statistics are discussed in detail later in this review.

The Agricultural Program at Statistics Canada uses “over 200” administrative datasets (Smith et al., 2013) in processes involving data collection, processing and analysis. Such data are fundamental to estimates of farm cash receipts, net farm income and food available for consumption. Taxation data are used extensively, and various segments of industry and government supply the Agriculture Division of Statistics Canada with information to support official statistics. The division continues to pursue new avenues for the use of
administrative data (Dion, Chartrand and Murray, 2010): as a consequence of a study vetting the reliability of tax data, for example, financial data will not be collected from most farms in the 2016 agricultural census in Canada but will be drawn from administrative data (Smith et al., 2013).

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) uses data from administrative agencies such as the Farm Services Agency (FSA) and the Agricultural Marketing Service to maintain and construct list frames, to evaluate and improve estimates, and for imputation (Harris, Beckler and Beranek, 2014). To finalize livestock estimates, for example, it supplements survey data with records of slaughter, shipment, import and export (Harris and Clark, 2013). It has also incorporated information on imports from the Economic Research Service and information on country of origin from the Foreign Agricultural Service in weighting for estimates of pesticide use (Kott, 2006).

The USDA Natural Resources Conservation Service (NRCS) supports the National Resources Inventory (NRI), a longitudinal survey of trends on non-federal United States land. As estimation controls, it uses data on federal land ownership, large bodies of water and enrolment in the government Conservation Reserve Program, which provides a monetary incentive to farmers to preserve rather than farm land. The NRI applications of calibration (Deville, Sarndal, and Sautory, 1993) seek to improve consistency in publications and to maximize the efficiency of estimators of other land-cover types and uses such as cropland, pasture, forest and rangeland (Nusser and Goebel, 1997).

The system of land records in state and union territories in India, an administrative source of agriculture statistics, covers 88 percent of the crop area (Goel, 2002; Sen, undated). Village accountants record information on areas used for purposes such as cultivation, orchards and irrigation, which are aggregated to the village level to produce reports of the crop area for each agricultural season (Goel, 2002) and are also used as sampling frames for surveys of yields and production (Sen, undated).

In Europe, administrative databases contribute to statistical registers used in the design and planning of surveys (Wallgren and Wallgren, 2007). The Integrated Agricultural Control System (IACS), for example, contains information on crop area of farms applying for subsidies. In Denmark, the IACS register is linked to the database of the Agricultural Census (FAO, 2010), and in Italy it is used to detect outliers and errors in census data (Reale et al., 2013). Utilizing aggregated totals of planted areas reported in IACS as calibration controls (Deville, Sarndal and Sautory, 1993) has improved the efficiency of crop area estimators for sub-domains of Italy (Carfagna and Carfagna, 2010).
and cattle movements drawn from tracing systems inform register-based statistics in several countries: examples include the Cattle Tracing System (CTS) of Great Britain and the Animal and Public Health Information System (APHIS) of Northern Ireland (Elliott and McDonnel, 2007). In view of the evidence of reliability of these systems, the Department of Environment Food and Rural Affairs in the United Kingdom requested that Eurostat replace statistical surveys and censuses with CTS and APHIS data (Elliott and McDonnel, 2007).

The Australian Bureau of Statistics (ABS) recognizes the various potential uses of administrative data. Australia uses data from the Australian Taxation Office to maintain list frames (Australian Bureau of Statistics, 2011), and in response to a 2011/12 review, the ABS expanded research into new ways of substituting or augmenting survey data with administrative data in the production of agricultural statistics (Henderson and Pitchford, 2013).

A review of metadata associated with agricultural censuses (FAO, 2010) cites other uses of administrative data in agricultural statistics programmes. A unique case is Kuwait, where no agricultural census is carried out because agriculture accounts for a small fraction of the economy; information is instead obtained from a registration database covering 3,000 to 4,000 farms. In Mongolia, the frame for the livestock census is the information base created by local offices responsible for maintaining data on their domain; crop data not covered by the livestock census are derived directly from administrative records (FAO, 2010).

1.5. OUTLINE

Section 2 cites examples of administrative data sources used by statistical offices in developed nations, with an emphasis on agriculture. Section 3 discusses the advantages of incorporating administrative data in statistical processes. Section 4 reviews the mechanisms through which these benefits are achieved. Section 5 identifies the required quality dimensions, and Section 6 discusses challenges to achieving them. Section 7 sets out approaches to quality control and mechanisms for overcoming the challenges discussed in Section 6. Sections 8 and 9 discuss probabilistic record linkage and register systems, respectively. Section 10 discusses lessons learned for developing countries. The examples in this review reflect various areas of application such as business and population statistics as well as agricultural statistics because many of the general concepts are potentially applicable to agricultural statistics and may be particularly useful in applications in developing countries.
Sources of Administrative Data for Agriculture

Section 2.2 of Technical Report 1 gives examples of administrative data relating mainly to household and business surveys drawn from United Nations (2011). Here, we provide examples of administrative data sources used to produce statistics related to the core data for the Global Strategy – land cover, crop production and livestock.

First we review three types of non-statistical data sources for the production of agricultural statistics – trade data, meteorological data and satellite imagery. Because these are covered extensively elsewhere, our discussion is limited.

Import and export information is an important source of non-statistical data for the production of agricultural statistics: control totals for imports and exports can, for example, serve as benchmarks for other estimates. The NASS and the Agriculture Division of Statistics Canada use trade data to improve the efficiency of estimators based on surveys and censuses (USDA, 2014b; Trant and Whitridge, 2000).

Statistical offices in many countries use satellite data to forecast crop yields, monitor crop conditions and classify land cover. Research is ongoing in Europe and in Brazil, Canada, China and the United States (Statistics Canada, 2012). Summaries of satellite data on greenness – the normalized difference vegetative index – can be used to model crop yields, for example (Johnson, 2014; Bellow and Lahiri, 2011).

Information on precipitation and temperature can support forecasts of crop production and yield (Kantanatha, Serban, and Griffin, 2010; Nandram, Berg, and Barboza, 2013).
2.1. SOIL INFORMATION

Topographical maps and maps of soil characteristics are often maintained through administrative processes. The NRCS, for example, maintains the Soil Data Mart database of soil characteristics on United States land. Information about land characteristics from topographical and soil maps can be used for stratification in surveys (i.e., Goebel, 2009) and as auxiliary information in constructing estimates.

2.2. CROP INSURANCE AND SUBSIDY PROGRAMMES

Government assistance programmes generate administrative data. Subsidies may be offered to encourage farmers to adopt conservation practices, for example, and farmers in developed countries often enrol in insurance programmes for protection against weather disasters or pest infestations. Administrative lists from such programmes can provide information such as areas planted with particular crops (Carfagna and Carfagna, 2010). Access to the administrative databases of government subsidy and insurance programmes requires good working relations with the administrative agency (Prell et al., 2009). Discussions with expert reviewers raised the issue that such sources are scarce in developing countries, but some reports note (Roberts, 2005 and Clay, 2013) increase in insurance programmes in developing countries.

Example 2.2.1. The IACS contains information on crop areas on farms in subsidy programs. Statistical offices in Denmark, Germany and Italy utilize the IASC database for various purposes (FAO, 2010).

Example 2.2.2. Beckler (2013): “The FSA is another agency within USDA and is tasked with administering a variety of agricultural assistance and conservation programs that provide price-support, disaster assistance, loans, and other services to agricultural producers. The omnibus United States Farm Bill, generally renewed every five years, provides authorizing legislation to FSA for the programs it administers. FSA collects an abundance of information from agricultural producers on the various application forms required to participate in the programs. Some of these data and FSA’s geographical information system data are used by NASS as administrative data (also called administrative records). NASS uses these administrative data in a variety of ways, including: (1) building and maintaining sampling frames, (2) as ground truth data for remotely sensed data, and (3) to supplement data collected on NASS’s censuses and surveys.”

Example 2.2.3. Smith et al. (2013): “AgriInvest helps to manage income decreases that are less than 30 percent of their historical reference margin. After submitting income taxes, participating producers receive an AgriInvest deposit notice stating the maximum amount they can deposit into their AgriInvest account, to be matched by government contributions. AgriStability is a program designed to assist producers when they experience larger decreases in their margins (greater than 30 percent of their historical reference margin) due to falling market prices, rising input costs or production fluctuation. Producers must submit an application and pay a fee to be eligible to receive an AgriStability payment. (AgriInvest and AgriStability are programmes operated at the federal and provincial levels to assist farmers with managing risk.)

Example 2.2.1. The IACS contains information on crop areas on farms in subsidy programs. Statistical offices in Denmark, Germany and Italy utilize the IASC database for various purposes (FAO, 2010).

Example 2.2.2. Beckler (2013): “The FSA is another agency within USDA and is tasked with administering a variety of agricultural assistance and conservation programs that provide price-support, disaster assistance, loans, and other services to agricultural producers. The omnibus United States Farm Bill, generally renewed every five years, provides authorizing legislation to FSA for the programs it administers. FSA collects an abundance of information from agricultural producers on the various application forms required to participate in the programs. Some of these data and FSA’s geographical information system data are used by NASS as administrative data (also called administrative records). NASS uses these administrative data in a variety of ways, including: (1) building and maintaining sampling frames, (2) as ground truth data for remotely sensed data, and (3) to supplement data collected on NASS’s censuses and surveys.”

Example 2.2.3. Smith et al. (2013): “AgriInvest helps to manage income decreases that are less than 30 percent of their historical reference margin. After submitting income taxes, participating producers receive an AgriInvest deposit notice stating the maximum amount they can deposit into their AgriInvest account, to be matched by government contributions. AgriStability is a program designed to assist producers when they experience larger decreases in their margins (greater than 30 percent of their historical reference margin) due to falling market prices, rising input costs or production fluctuation. Producers must submit an application and pay a fee to be eligible to receive an AgriStability payment. (AgriInvest and AgriStability are programmes operated at the federal and provincial levels to assist farmers with managing risk.)
2.3. LAND REGISTRATION AND CADASTRAL RECORDS

A cadastre consists of records defining the “extent, value and ownership of land” (Bins and Dale, 1995). They are used for purposes of taxation, and to provide precise descriptions and continuous records of land ownership. Many European countries maintain them.

Cadastres have potential as contributors to statistical farm registers in the European Union (Turtoi, Akyildirim and Petkov, 2012), and the cadastral survey of Romania has provided the basis for sample surveys (NMR, 2008). Defining a farm register, Turtoi, Akyildirim and Petkov (2012) explain: “… [the Farm register] is based on the holder of the land. In order to have a subsidy, each individual holder has to register himself/herself by bringing land ownership documentations from the Cadastre Office. In the register, both physical entities and enterprises are recorded.”

India’s land registration system has characteristics in common with European cadastres. In India, land revenue administration by state governments is a source of administrative data for agricultural statistics (Sen, undated; Goel, 2002). This resource consists of information on land use and crop management gathered by village-level accountants. Examples include crop areas, fruit orchards, irrigated areas and irrigation sources (Goel, 2002). The land registration system covers 88 percent of the crop area (Goel, 2002). Data are tabulated directly from land records, and registration information is used as a sample frame for surveys of crop yields and production (Goel, 2002; Sen, undated). The aim of the Timely Reporting System, a process whereby village heads collect data for a 20 percent sub-sample instead of all crop areas, is to accelerate data collection (Government of India, 2013).

2.4. GOVERNMENT REGULATION AND MONITORING PROGRAMMES

Government regulation and monitoring programmes, whether voluntary or mandatory, produce substantial administrative data. Regulatory activities include the monitoring of production processes, financial institutions and insurance practices, and the resulting administrative data are used by statistical offices in a variety of ways.

Sen (undated) discusses the role of regulatory processes in India’s statistical system: “… the emergence of regulators in some areas of industry has led to a new source of administrative data i.e. regulatory records. These have been found particularly useful in areas such as pharmaceuticals and power…Indeed, practically all the data for the banking and financial sectors, including
insurance, is derived almost exclusively from regulatory demands. This is also true of practically the entire transport sector including road, air, and marine transport.” India’s Central Statistical Office receives production data on milk, meat, eggs and wool from the Department of Animal Husbandry; state governments provide data on other livestock items (Government of India, 2013).

Data from regulation and monitoring of agricultural production and consumption have a significant role in agricultural statistics: in some countries, landowners are required to register their land, and information is derived from farm food-safety and health inspections and vaccination records. Systems that monitor births, deaths and movement of registered livestock are becoming increasingly important as sources of administrative data: Sweden and Uruguay, for example, have tracing systems that provide good coverage of animal populations.

An example of the usefulness of such systems for agricultural statistics stems from a discussion in the United Kingdom: “The CTS records the identification, births, deaths and movements of individual animals. By tracking individual animals throughout their lives the system makes it possible to determine which animals are present on a given holding and also the size and composition of the cattle population at any given time” (Elliott and McDonnell, 2007). With information from APHIS in Northern Ireland, details of cattle populations for the whole of the United Kingdom are obtained.

The Forestry Commission of the United Kingdom makes use of administrative data from government monitoring programmes when constructing its statistics. Sources include recreation facilities, grant schemes, sales records and tree-felling licenses, and the data are often used in conjunction with surveys or specific data collections. Statistical publications that utilize this administrative information include Timber Price Indices, Forestry Statistics and Woodland Area. (Ward, 2014)

**Example 2.4.1.** SANITEL is a relational database that contains a permanent inventory of animals in Belgium. It provides a complete inventory of counts and movements of cattle and pigs, and contains information on health status and the detection of antibiotics, hormones or contaminants. The database is managed by the Central Association for Animal Health, not by a statistical office. The information in the database is supplied by regulatory activities: “Every keeper of pigs is required to complete a health certificate showing the capacity of his holding. Subsequently, every three or four months, approximately, he has a visit from an approved veterinarian so that he can declare the type and number of animals actually present” (European Communities, 2003). Since 2002, Belgium has reduced the number of pig surveys from four to two with a view to replacing the survey data with information from SANITEL to compile its gross indigenous production forecasts (European Communities, 2003).
2.5. PRIVATE SOURCES OF DATA

Private organizations concerned with agriculture such as licensing or regulatory bureaux, grain associations, commodity associations, cooperatives, factories, slaughterhouses, distributors of agricultural inputs and agricultural extension workers affiliated to universities gather regular agricultural information that may be used in official statistics (USDA, 2011). In Canada, price information is obtained from sources such as the Ontario Wheat Producers Marketing Board and the Nova Scotia Grain Marketing Board (USDA, 2011). Sources used in maintaining the National Agricultural Statistics Service frame include: “… licensing bureaux, grain associations, commodity associations, cooperatives, extension crop specialists at universities, auction facilities, factories, mills, buyers, feeders, brewers, ginners, processors and distributors. NASS also uses administrative data in establishing mid-month prices, rather than using survey information from buyers and sellers. The data are used for setting national prices for fruit, vegetables, livestock, poultry, feeder livestock, and fuel” (USDA, 2011). In Australia, information from growers’ associations is used to establish benchmarks for dairy production (Boreo, personal communication).

Example 2.5.1. Keita and Chin (2013) cite a study in Cape Verde in which information from private organizations was critical because there were no consistent survey or census data. They explain that Cape Verde “… is an island country with irrigated agriculture and cash crops concentrated in a limited number of well-known zones …” and that the culture of agricultural production fosters a system of farmers’ organizations and cooperatives for certain cash crops.

A multi-disciplinary team investigated the feasibility of filling data gaps in time series for the production of primary horticulture crops, irrigated crops and cash crops. It was soon evident that with the exception of non-irrigated agriculture there was no centralized survey system for estimating agricultural production: primary information came from decentralized surveys, administrative data and expert knowledge.

The investigation involved: i) a review of documentation, surveys and studies; and ii) collection of data from administrative sources, regional offices, research institutions and experts. Several methods were compared to determine the range of estimates possible from the available information, and a specific method was developed to estimate horticulture production. A primary input to the horticulture production estimate was information on seeds sold by the two main seed marketing companies. Other sources included average yield from a crop-cutting survey by the statistical service of the Ministry of Agriculture, and estimates of seed quantity per hectare by research institutions. The final production estimates were cross-checked with data from consumption surveys.

2.6. REPORTING SYSTEMS AND EXPERT OPINIONS

Individuals who regularly participate in agricultural production and research processes naturally gain considerable expert knowledge. Subjective assessments
by expert reporters provide information for statistical offices in many countries (Keita and Chin, 2013; Galmes, 2013; Hamer, 2013). Such reporters include individuals involved in agribusiness, university research and administrative agencies (Hamer, 2013), who often have expert knowledge of a particular domain of interest.

Systems for expert reporting are of particular interest to this project because of their prevalence in developing countries. In Africa, agricultural reporting systems in ministries of agriculture can provide weekly, monthly, semi-annual or annual reports of plantings, production, crop conditions and weather (see section 10.2).

Limitations of this kind of information include lack of objectivity and uncertainty as to representativeness. Galmes (2013) notes that survey samples are the only “… scientific way to provide measures of sampling errors in terms of precision and accuracy of estimates.” Nonetheless, administrative data from reporting systems and expert opinions are valuable, particularly when timely data or data at detailed levels of spatial aggregation are needed (Keita and Chin, 2013). To compensate for the limitations of information based on expert assessments alone, it is advisable to combine administrative data from assessments by qualified informants with objective data from surveys (Galmes, 2013).

Example 2.6.1. NASS publishes weekly crop-progress and crop-condition reports (Hamer, 2013) based on variables such as percentage of land planted, percentage in good condition and percentage harvested; there are specific questions for particular crops. The report is published each Monday after 16:00 EST during the primary United States growing season from the beginning of April to the end of November.

The reports contain inputs from 4,000 reporters, who interact with farmers and agricultural operations as part of their jobs. They include employees of USDA agencies involved in agriculture, farmers and participants in agribusiness.

A criticism of the reports is that the estimates are based on subjective interpretation – but they are timely, and the format is standardized. Interestingly, the NASS crop-progress and crop-condition reports are requested more frequently than any other NASS publication (Hamer, 2013).

2.7. TAXATION DATA

Taxation data have been used for a long time in statistical processes (Nordbotten, 2008), providing information on individual and household incomes, business types and sizes, and changes such as migrations and the start or end of commercial operations in years when no census is carried out. The various roles of tax data in producing short-term business statistics are described in OECD (2015). Statistical offices in Europe and in Australia,
Canada and the United States make extensive use of taxation data in producing agricultural statistics: Canada, for example, plans to use tax data to measure farm expenses for most farms in the 2016 Census of Agriculture (Smith et al., 2013) (see also Section 4).

2.8. CONNECTIONS TO CORE DATA ITEMS

The sources discussed above cover many of the core data items from the Global Strategy. Subsidies provide information on crop areas, tax data provide information on farm expenses, slaughter and vaccination records provide information for forecasting and estimating hog inventories, and data from distributors provide information about dairy production (see Table 1).

<table>
<thead>
<tr>
<th>Core data items</th>
<th>Administrative data type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crops</td>
<td>Farm subsidies</td>
<td>IACS contains crop areas for crops enrolled for subsidies</td>
</tr>
<tr>
<td></td>
<td>Grower associations</td>
<td>The Ontario Grain Association provides information on prices</td>
</tr>
<tr>
<td>Livestock</td>
<td>Animal health regulations</td>
<td>SANITEL in Belgium is populated with data from animal health regulations and supplements surveys</td>
</tr>
<tr>
<td></td>
<td>Cattle tracing systems</td>
<td>Cattle tracing systems populate the European Union Bovine Register</td>
</tr>
<tr>
<td>Forestry</td>
<td>Forest cover area</td>
<td>Forestry Commission records complement statistical surveys in estimations of forest area and woodland prices in the United Kingdom</td>
</tr>
<tr>
<td>Land cover</td>
<td>Land registration and cadastral records</td>
<td>India’s land registration system supports estimates of areas in various land-cover categories</td>
</tr>
</tbody>
</table>
Benefits of Administrative Data

Section 3 of Technical Report 1 discusses the benefits of administrative data in producing official statistics: these include reductions in cost and the burden of response, improved timeliness and efficiencies derived from collaboration among government agencies. This section extends the discussion.

Figure 1 shows perceptions by “key informants and focus group discussions” of the benefits of administrative data in various areas on a scale of 1 (low) to 4 (high). It is evident that the participants in the study viewed administrative processes as sources of timely data at low cost and a reduced burden on respondents.

Figure 1. Characteristics of Administrative and Survey Data

Source: Pangapanga et al. (2013)
Carfagna and Carfagna (2010) note the lower costs and burden on respondents from using administrative data. Recent cuts in government budgets have motivated statistical agencies to use all available data, and increasing non-response rates in surveys lead to concerns about bias and options to reduce the burden on respondents. Administrative data could be better than those from surveys and censuses in terms of timeliness, accuracy and small-area statistics. These issues are discussed in detail below.

3.1. COST SAVINGS

Lower costs make administrative data attractive to official statisticians, especially when budgets for government statistical agencies are reduced. Collecting data by mail, telephone or interview is expensive, and considerable savings are achieved by tapping into information in administrative databases. Administrative data are not entirely free because there is a cost associated with quality assurance, but this is generally less than the cost of a survey and the saving is spread across data collection and data maintenance (United Nations, 2011). The cost benefits of using administrative data in official statistics are evident in censuses conducted by European Union countries (see Example 3.1.1).

**Example 3.1.1.** A comparison of the costs of population censuses in 2000/01 in 16 European countries illustrates the savings obtained by using administrative data. Finland achieved the lowest cost per head of population – €0.2 – by basing its census entirely on administrative registers (United Nations, 2011). The highest cost per head was €13.6.

In 1981, Denmark became the first country in the world to conduct an entirely register-based census; Finland followed in 1990. Studies of the advantages and disadvantages of register-based censuses show that the cost savings outweighed the drawbacks in Finland and Denmark. The lower expenditures did not significantly compromise data accuracy in the Nordic countries, which are known for their sophisticated data-sharing infrastructure, prompting the decision to use registers exclusively for subsequent censuses (United Nations, 2011).

Many developing countries do not have the resources to conduct register-based surveys and censuses. This may be a result of factors such as prevalence of small-scale subsistence farms and nomadic populations that are not enrolled in government programmes (Himelein, Eckman and Murray, 2014). As statistical offices in developing countries expand, they could adopt the practices in Nordic countries as a goal. Current surveys and censuses may form the foundation for future register-based tabulations if resources are invested in record-keeping and maintenance of personal identification numbers. Statistical offices in developing countries would also benefit from the development of positive working relations with government agencies with a view to establishing comprehensive administrative lists.
3.2. REDUCED BURDEN ON RESPONDENTS

Obtaining information from an administrative source rather than a survey reduces the burden on respondents. This is an attractive characteristic of administrative data, particularly in view of recent increases in non-response rates. Guigo (2008), for example, cites the report by Al and Bakker (2000) of a decrease in the response rate for the Netherlands Labour Force Survey from 90 percent in 1977 to 60 percent in 1995. Filling forms and providing required data are routinely expected of businesses and individuals, but if such information is already in a government database the exercise becomes counter-productive. Similarly, if information such as change of address has been filed with one agency, it should ideally be updated automatically in other government databases. Data sharing and collaborative dataset construction and maintenance would enhance the efficiency and transparency of governance, and would also support the effective use of administrative data for statistical purposes.

Example 3.2.1. Experience of using tax data for business statistics in Lithuania shows how the burden on respondents can be minimized. Statistics Lithuania uses 110 administrative data sources. Use of tax data has reduced the burden on respondents in that 40,000 enterprises do not need to fill out annual questionnaires on earnings; and because in 2007 administrative sources replaced short-term labour indicators from surveys, 16,000 enterprises received abbreviated questionnaires (Lapeniene, 2007).

Example 3.2.2. In the Conservation Effects Assessment Project (CEAP) supported by USDA/NRCS, administrative data substitute survey with a view to reducing the number of questions that farmers must answer. “CEAP is a multi-agency effort… to quantify the environmental effects of conservation practices and programs and develop the science base for managing the agricultural landscape for environmental quality” – see: http://www.nrcs.usda.gov/wps/portal/nrcs/main/national/technical/nra/ceap/

To understand the dynamics of crop-management and conservation practices in watersheds, CEAP utilizes detailed information from farmers. Some of this is available from the NRCS field office records database, which is maintained for planning purposes. To reduce the number of questions put to farmers and thereby save time, many of the data elements in the database are excluded from the questionnaire.

Example 3.2.3. Administrative data are combined with survey data in the production of business statistics in Austria (Dinges, den on2008) with a view to reducing the burden on respondents. Under a cut-off strategy, survey questionnaires are administered only to businesses above a specified size; for businesses below the threshold administrative data are used instead. This strategy has potential in the agricultural context: the size cut-off approach, for example, has been researched at USDA in the context of collecting information on hog producers.

Example 3.2.4. The Red Tape Reduction Commission of the Canadian government is intended to reduce the burden of federal regulatory requirements on Canadian enterprises, including agricultural operations. Statistics Canada was an agency judged to impose a significant burden of regulatory obligations, and in response to the commission it evaluated the feasibility of using administrative and remotely sensed data instead of survey or census data. In another case, the Tax Replacement Study found that taxation data provided reliable and detailed information on expenses, and tax data will replace questions on expenses in the 2016 Census of Agriculture, reducing the length of the questionnaire by 7 percent (Smith et al., 2013).
3.3. IMPROVEMENTS IN MICRO-DARA QUALITY

Because the collection of administrative data is often mandated in government programmes, the information may be more complete and more accurate than survey data. It follows that substituting or augmenting survey information with administrative data can improve the accuracy and completeness of the resulting micro-data, and evaluations of discrepancies between survey data and administrative data can be used to identify outliers and correct errors in either source. Statistics Canada uses administrative data to “… confront, validate, and certify estimates from surveys or censuses (e.g. comparing estimates from administrative sources to those produced from the survey) …” (Smith et al., 2013). Information from taxation records is a significant source of administrative data for improving the quality of micro-data. Examples 3.3.1 and 3.3.2 illustrate the use of tax data to improve micro-data related to business and agriculture.

Example 3.3.1. The Italian team of the European Union Statistics on Income and Living Conditions survey integrated survey and administrative data to improve the quality of survey data on disposable income (Consolini, 2008). Tax data were available for sampled units in addition to information from the survey. In the tax data, disposable income was the sum of net taxable income and deductions. It was reported directly in the survey, but there was a concern about under-reporting (see Figure 2). Because the analysts suspected under-reporting in both sources, the maximum of the two versions of disposable income was used. This imputation process increased estimates of disposable income by 15 percent (Consolini, 2008).

Example 3.3.2. The Canadian Agriculture Taxation Data Program has provided statistics on farm revenues, expenses, incomes and families for 25 years. A 2009 census test involved merging census and tax records: the tax data were judged to have high reliability for important domains, and modest differences indicated that “… information prepared for tax data might be more thorough and complete …” than information reported on the census form (Smith et al., 2013). In related work, Statistics Canada has observed that farmers tend to over-state expenditures and understate sales, but the integration of tax data with farmer interview data, has reduced bias arising from reporting error (Trant and Whitridge, 2000).
3.4. IMPROVEMENTS IN THE EFFICIENCY OF MACRO-LEVEL ESTIMATORS

Even if information at the micro-data level is sparse or difficult to obtain, administrative data may improve the efficiency of estimators at larger levels of aggregation. If released on a regular time schedule, administrative data can improve the consistency of published time series (Carfagna and Carfagna, 2010). One thoroughly studied procedure to improve efficiency of large domain estimates is called calibration. This technique is discussed in more detail in Section 4.4. In concept, the total from the administrative source is assumed known at the level of the domain of interest. If the quantity observed on the administrative source is correlated with the quantity of interest in the survey, calibration can lead to improvements in efficiency of estimators. The example below illustrates the potential reductions in mean squared error that are possible from calibration.

**Example 3.4.1.** As mentioned previously, the IACS contains information on the area of subsidized crops in Europe. Carfagna and Carfagna (2010) reported the results of using IACS data as controls in calibration estimators; the improvements in the estimated coefficients of variation are shown in the table.
3.5. SMALL – AREA STATISTICS

Because the number of records in administrative databases is often larger than in surveys, administrative data have the potential to improve small-area estimates and refine the detail in reports. Holt (2007) comments that statistical offices in many countries increased their use of administrative data in response to increasing demand for small-area statistics, and the Asian Development Bank (2010) states: “Compared with censuses and surveys, administrative data…can be presented at various levels of disaggregation, such as by geographical location or by age and gender.”. Statistics New Zealand (undated) reinforces this in the context of business surveys: “The use of individual business tax and other administrative data within the official economic statistical program has major benefits in terms of the range and depth of statistical outputs that can be efficiently produced without imposing additional compliance costs on providers.”

Galmes (2013) advocates the use of information from expert informants in conjunction with surveys as a mechanism for obtaining reliable small-area estimates, and notes that surveys have the advantage of providing a scientific measure of sampling variation; cost constraints, however, often prohibit the construction of geographically detailed survey-based estimates. He also states that a combination of sampling surveys with administrative data derived from expert assessments can meet decision-makers’ demands for agricultural data in small administrative areas without sacrificing the scientific integrity of traditional survey methods.

Keita and Chin (2013) cite increasing demand for small-area statistics as a reason why alternatives to surveys will continue to be important elements in agricultural statistics. They state: i) that increasing decentralization of government, planning and monitoring systems increases the need for information at low geographic levels; ii) that monitoring poverty and implementing poverty-reduction polices “… require an understanding of local conditions for implementing poverty relief measures”; and iii) that because sample sizes in surveys are often too small to provide reliable information for small areas, alternative methods of generating agricultural data are needed:

<table>
<thead>
<tr>
<th>Domain</th>
<th>Estimated coefficient of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without calibration to IACS</td>
</tr>
<tr>
<td>Puglia</td>
<td>4.8%</td>
</tr>
<tr>
<td>Sicily</td>
<td>5.9%</td>
</tr>
</tbody>
</table>
these include “administrative sources, eye estimates by local informants, expert opinion or assessment, …village surveys, or other subjective assessment.”

One way of integrating survey data with information from administrative sources to produce small-area statistics is a small-area model (Rao, 2003; see also Section 4.6). In this approach the administrative data are covariates in a model, with the survey information as the response. If the administrative data are reliable at the small-area level and correlated with the response variable of interest, the estimates for the small-areas that incorporate the extra information from the administrative source are likely to be more reliable than estimates from the survey information alone.

**3.6. TIMELINESS**

Administrative data sources are often updated more frequently than surveys. The USDA Census of Agriculture, for example, is conducted every five years, whereas agricultural operations submit tax returns and apply for subsidy programmes at shorter intervals. The Asian Development Bank (2010) comments on timeliness as an advantage of administrative data: “Frequent and quick compilation of statistics is possible so that these can be released earlier than through data collected through censes and surveys.”

Keita and Chin (2013) consider that demands for timely data are a reason for using administrative sources rather than surveys: “Even in countries where censuses and surveys are conducted, these activities are likely to be implemented with multi-annual frequency. Governments are increasingly demanding more frequent data on a yearly or even sub-annual basis. This is particularly the case for certain areas such as prices, where economic volatility has resulted periodically in rapid and wide ranging fluctuations from past trends.”

Galmes (2013) suggests that expert opinion assessments are a means of obtaining timely estimates; the NASS crop-progress and crop-condition reports are examples of this. “The power of the weekly report lies in the flexibility of asking over 3,000 potential questions each week, the panel of experts who provide data, and the fast processing time” (Hamer, 2013).

Rao (2003) discusses the role of administrative data in obtaining timely estimates of population changes in non-census years: “In the absence of population registers maintained over time (as in some Scandinavian countries), it becomes necessary to develop suitable methods of population estimation in non-census years, exploiting administrative files that contain valuable demographic information related to population change.” He also notes that administrative files used in population estimates include registers of births and
deaths, school enrolment and income tax returns. Observing that when these methods are used to obtain small-area estimates, he notes that they often rely on the assumption that the dynamics of the small area match those of the larger area and states that to reduce the need for such assumptions, methods of small-area estimation have evolved that use explicit assumptions based on mixed-effects models. In such cases the administrative data are often incorporated as a covariate in the model. Model-based small-area estimation is discussed in Section 4.6.
4

Uses of Administrative Data

The benefits of administrative data – saving money, reducing the burden on respondents and improving quality – are usually achieved together. It is therefore difficult to assign the mechanisms for achieving these benefits to distinct categories, and even a single application of administrative data often involves multiple uses. The following section sets out ways in which administrative data can be used to achieve the purposes discussed in Section 3.

The extent of the contribution of administrative data to statistical products varies: at one extreme the data may be used directly as published estimates, but in other contexts they may provide a partial substitute for survey data or improve the efficiency of estimators through integration with survey data.

4.1. DIRECT TABULATION

If administrative data are of adequate quality they may be used directly as statistical products (Brackstone, 1987; Wallgren and Wallgren, 2010). On the basis of practices at Statistics Canada, Brackstone (1987) describes direct tabulation as the processes of counting units in files, cross-classifying them by attribute and aggregating quantitative variables associated with each unit. Estimates of vital events such as births, deaths, immigration and emigration and business developments are often obtained from administrative sources (Trant and Whitridge, 2000). An agricultural example is the start of a new farm operation. Customs documents for imports and exports can serve as a basis for statistics on agricultural production (Trant and Whitridge, 2000); NASS routinely publishes information on imports and exports of agricultural products (Harris and Clark, 2013).

Direct publication of administrative data is often based on register systems, defined in Carfagna and Carfagna (2010) and United Nations (2011) as systematic collections of uniquely identifiable unit-level data with an updating mechanism (see Section 9). A register populated from several administrative sources may be more comprehensive and provide better coverage than a single source. Population and business registers are frequently used. Wallgren and
Wallgren (2010) note that business registers have potential as a basis for agricultural statistics because farm operations are a type of business enterprise. And in discussing the use of the IACS for direct tabulation they point out: i) that for some crops receiving government subsidies the IACS is very reliable and therefore directly tabulated to obtain aggregated area statistics; and ii) that combining the IACS database with the business register leads to further improvements.

The availability of administrative databases for direct tabulation of agricultural statistics is different in different countries. Clay (2013) observes that in the last 12 years many developing countries have created their own agricultural subsidy programmes in order to compete with prices in the United States and the European Union, and that such programmes have grown fastest in Brazil, China, India, Indonesia and Russia. Roberts (2005) noted that farm insurance programmes are another potential source of administrative data in developing countries, and that crop insurance is expanding as agriculture becomes more commercial, with new products based on weather indexes, and as international trade policies. The coverage of such programmes in developing countries may not currently be sufficient for direct tabulation, but if they continue to grow they may be leveraged in future register-based agricultural statistics.

Example 4.1.1. NASS publishes administrative information on hog slaughter (Harris and Clark, 2013) obtained from inspections by federal and state officials. The data from NASS hog and pig inventories should align with published slaughter data.

Example 4.1.2. Statistics Canada uses tax records to estimate farm expenses with a view to reducing the burden on respondents. The use of administrative data instead of survey data could improve data quality because farmers are thought to over-state expenses and under-state sales in surveys (Trant and Whitridge, 2000).

Example 4.1.3. Statistics Canada conducts 39 surveys on a sub-annual or annual basis. In some cases only part of the population is surveyed, and the data for the remainder are obtained from administrative sources. Two examples of this are the Maple Products Survey, where only producers in the provinces of Ontario and New Brunswick are surveyed and administrative data are used instead of conducting surveys in Quebec and Nova Scotia; and the Honey Production, Value and Colonies Survey, where only producers in the provinces of Prince Edward Island and New Brunswick are surveyed, since administrative data are obtained from the other provinces (Smith et al., 2013).

4.2. FRAME CONSTRUCTION AND IMPROVEMENT

A sampling frame is a representation of a target population: it may be in the form of a list, called a “list frame”, or a map, called an “area frame”. Frames provide the basis for sample surveys and censuses. Statisticians desire an administrative database – which may be based on several lists – to be complete, accurate, rich in variables and up to date. Breadth and depth are needed to
support sample surveys with different goals and to provide auxiliary information for designing samples. It may be tempting to build a frame from a single reliable administrative dataset to save on cost, but it must be borne in mind that an administrative file judged sufficient for administrative purposes may not cover the full population of interest to a statistical agency (see Section 6). A better approach than using a single data source to define the frame directly involves using the administrative file for frame improvement: this results in improved coverage for sample surveys and censuses (Carfagna and Carfagna, 2010).

Statistical offices may use tax data to maintain sampling frames for establishment surveys. Statistics Sweden, for example, uses taxation information to analyse the coverage of their business register: the various sources include administrative data from value-added tax payments, gross pay and preliminary tax based on statements of income and gross pay, payroll taxes and preliminary tax from employers’ monthly tax returns (Berg and Hall, 2007). In determining eligibility for specific surveys, examination of which combinations of sources contain a particular unit will show whether the unit is a member of the sub-population of interest (Berg and Hall, 2007). Sampling frames for many of Statistics Canada’s business surveys are derived from the business register: when a new business is formed, the Canadian Revenue Agency sends Statistics Canada a description of the type of activity and a business number, which is incorporated into the business register and assigned an industry classification (Yung, Rancourt and Hidiroglou, 2007). Tax data is the primary administrative data source used to maintain the business frame at Statistics New Zealand (Statistics New Zealand, undated).

### 4.3. SURVEY DESIGN

Efficient sample designs rely on information about the structure of the population of interest, which is often available from administrative sources. Two examples are probability-proportional-to-size (pps) sampling and stratified sampling.

---

**Example 4.2.1.** NASS maintains a list of farmers and ranchers for the United States Agricultural Census and constantly improves it by obtaining external agriculture-related lists kept by bodies such as state and federal governments, producers’ associations, seed growers, pesticide applicators, veterinarians and marketing associations. NASS also uses the United States Postal Service’s National Change of Address Registry and Locatable Address Conversion System to improve the accuracy of names and addresses in the frame; a national telephone database is used to replace missing or invalid phone numbers from the list. (USDA, 2014 a).

**Example 4.2.2.** At Statistics Canada, administrative lists have helped in the development of frames covering farms with small land area such as chicken, egg, pig, fruit and vegetable farms. Such farms are difficult to capture in the absence of administrative lists (Trant and Whitridge, 2000).
sampling. In pps sampling, a size measure is defined for all units in the frame and the selection probability is proportional to it; if the size measure is correlated with the response of interest, pps sampling is more efficient than simple random sampling. In stratified sampling, the population of interest is divided into groups called strata; if the strata boundaries explain variation in the population, efficiency can be increased. Fuller (2009) and Sarndal, Swensson, and Wretman (2005) discuss the role of auxiliary information in sample designs.

Example 4.3.1. The Australian Bureau of Statistics uses taxation office data to maintain sampling frames for several business surveys such as industry, state, public or private sector, the number of paid employees in a business and turnover to help to stratify the business frame. This can improve the efficiency of estimators from resulting survey samples (Australian Bureau of Statistics, 2011).

Example 4.3.2. In changing the Italian AGRIT from an area frame to an unclustered point frame, a candidate stratification scheme is based on the IACS, whose records permit classification of each unit in the frame into one of the following classes: arable land, permanent crops, permanent grass, forest, isolated trees and rural buildings and other (Carfagna and Carfagna, 2010). This classification provides one stratification option for stratified sample designs.

Example 4.3.3. Statistics Sweden uses tax data to define strata for a survey of the shares and assets of businesses. The population of interest is highly skewed, with a small number of units accounting for a large percentage of population totals of the variables of interest. The stratification of the survey is according to the total amount of shares and assets recorded on the tax data (Berg and Hall, 2007).

Example 4.3.4. All active units on the business register of Statistics Canada have an associated gross business income; for most of the units in the frame, this is based on tax data or is modelled from employment data. In the past, most surveys used gross business income as a measure of size during stratification; more recently, business surveys have started to use data from sources such as reported survey data from a previous time period or data obtained from the annual tax programme (Yung, Rancourt and Hidiroglou, 2007).

4.4. MODEL - ASSISTED CALIBRATION ESTIMATORS

Auxiliary variables with encapsulated information from administrative data are often used in estimation and design. In estimation, administrative data are used because they have zero sampling variance and are often highly correlated with the quantity of interest in the survey, even though they may be less accurate in some respects. In calibration, the weights for sampled units are modified so that the weighted sum of the auxiliary variable is equal to the administrative control. The term “control” is used in the sense that estimates of sub-categories must match a predetermined total when combined, and the predetermined total is from sources external to statistical surveys. The stronger the correlation between the variable recorded on the administrative file and the survey variable, the greater the efficiency gain from calibration (Deville, Sarndal and Sautory,
Thomsen and Holmoy (1998) discuss the use of administrative data in calibration at Statistics Norway.

In some cases administrative data do not provide information on exact quantities, but they provide information about ranges and inequalities. An administrative total that represents a combination of more detailed categories, for example, provides an upper bound for the total of any one of the contributing categories. In such instances the survey weights can be constructed to preserve inequality constraints or range restrictions determined by the administrative source. Example 4.4.1 illustrates the use of customs data to define an inequality restriction.

Example 4.4.1. In Canada customs based data on imports and exports are integrated with surveys to improve estimates of the level of agricultural production required to meet domestic demand. Grain estimates, for example, are restricted so that exports do not exceed domestic production, and livestock estimates are constrained so that reported slaughter levels do not exceed reported animal counts. These measures not only improve statistical efficiency but also ensure that the published estimates are internally consistent (Trant and Whitridge, 2000).

Example 4.4.2. The United States NRI exercises different levels of control so that its estimates are consistent with data from other federal and state authorities. It classifies land into 18 mutually exclusive and exhaustive categories such as cultivated cropland, non-cultivated cropland, pasture, federal, public road and large urban. For each state, the sum of the estimated areas in all 18 categories must equal an administrative value for the total acreage of the state (Nusser and Goebel, 1997).

Example 4.4.3. The United States Bureau of Land Management partners the USDA/NRCS to obtain estimates of rangeland conditions through rangeland surveys. The 2012 survey aimed to assess the conditions of greater sage grouse habitat on bureau rangeland under three domains: greater sage grouse priority habitat, ecoregions and Western Association of Fish and Wildlife Agencies zones. At the estimation stage of the survey, administrative data on the area of rangeland in 13 western states were used as calibration controls in constructing weights.

The administrative data came from:

- Geographic information system (GIS) layers from the Bureau of Land Management, which define the boundaries of bureau-managed land in survey-eligible states;
- GIS layers representing the joint work of the NRCS and the Bureau of Land Management by combining information on the spatial distribution of greater sage grouse breeding densities with the NRCS Common Resource Area geographic database;
- the United States Environmental Protection Agency’s designation of ecoregion classes, based on Omernik (1987) level II and III ecoregions; and
- GIS layers delineating sage grouse management zones developed by the Western Association of Fish and Wildlife Agencies that reflected ecological and biological issues and similarities, not political boundaries.

The estimation procedure began with the construction of weights for all points in the sample to obtain estimates of the acreage of bureau-managed rangeland in each combination of state, sage grouse habitat and non-habitat, ecoregion and zone. Subsequent weighting involved the application of raking and successive ratio adjustments to preserve the three sets of control totals – state-by-type strata, ecoregions and Western Association of Fish and Wildlife Agencies zones. At the end of the calibration, the final analysis weights sum to the administrative acres of bureau-managed rangeland in each Western Association of Fish and Wildlife Agencies zone.
4.5. NON-RESPONSE ADJUSTMENTS AND IMPUTATION

Surveyed units may complete only part of the questionnaire or may refuse to respond. If the characteristics of non-respondents are systematically different from the characteristics of respondents, estimators constructed with only the complete data can be biased for underlying population parameters of interest. Consider, for example, a survey intended to provide information on average erosion rates on cropland. If farmers who apply conservation practices have higher response probabilities, estimates of mean erosion based only on the complete data are likely to be biased.

Administrative data may be available for respondents and non-respondents. If a variable from an administrative database is observed for both and is related to the response variable of interest, the auxiliary information from the administrative source may be used to reduce the bias arising from non-response. Comparisons of means of the auxiliary variable for respondents and non-respondents can indicate the nature of the non-response: if the quantity recorded on the administrative source is correlated with the outcomes of interest, the administrative data may be used as auxiliary information in constructing estimators that account for the non-response bias. Two methods for adjusting for nonresponse bias, imputation and weighting, are discussed in Sections 4.5.1 and 4.5.2.

4.5.1. IMPUTATION

One mechanism to adjust for non-response is to impute missing data. This is particularly useful for item non-response, where units complete only part the survey. Kim and Shao (2013) and Sarndal and Lundstrom (2005) describe the theory and methods of imputation.

To describe the imputation method, we use the following notation:

- Units. We use $i$ to index the unit in the population of interest. In an agricultural context, for example, $i$ may index an agricultural holding. We use $N$ for the number of units in the administrative data source, so $i=1,...,N$ in the administrative data file. We use $n$ for the number of units in the sample, so $i=1,...,n$ in the sample.
- Response indicators. In many practical applications, not all sampled units respond to the survey. We define the response indicator $\delta_i$ by $\delta_i = 1$ if unit $i$ responds, and $\delta_i = 0$ if unit $i$ does not respond. The number of respondents in the sample – $n_r$ – is the sum of the $\delta_i$ for sampled units $i=1,...,n$.
- Variables. Let $y_i$ be the response variable of interest. We assume that $y_i$ is measured in the survey but not in the administrative source. The
variable $y_i$ is observed only for sampled respondents – units in the sample for which $\delta_i = 1$. Assume that a vector of auxiliary variables – $x_i$ – is available from the administrative source. Because the vector $x_i$ is measured in the administrative file, $x_i$ is available for respondents and for non-respondents. The assumption that the auxiliary variable $x_i$ is known for the full sample is critical to the use of $x_i$ in analyzing and adjusting for non-response bias. We refer to the collection $(y_i, x_i, \delta_i)$ for sampled units, with $\delta_i = 1$ as the complete data.

- Models. Suppose a relationship between the response and the auxiliary variable is specified. To be precise, suppose that this relationship is described as a conditional distribution and let $f(y_i \mid x_i)$ denote the conditional distribution of $y_i$ given $x_i$ in the population. For the purposes of imputation, we assume that the response mechanism satisfies the assumption of “missing at random” (Rubin, 2004; Kim and Shao, 2013) in the population. This assumption asserts that the distribution of the response variable $y_i$ is independent of the response mechanism $\delta_i$ given the auxiliary information. Mathematically, the assumption of missing at random in the population can be expressed as:

$$f(y_i \mid x_i, \delta_i = 1) = f(y_i \mid x_i, \delta_i = 0) = f(y_i \mid x_i)$$

- The assumption of missing at random is critical for imputation because it enables prediction of $y_i$ for non-respondents with $\delta_i = 0$ given an estimate of the conditional distribution and an observed covariate $x_i$ for non-respondent $i$ (see next).

- Imputation. Estimate $f(y_i \mid x_i)$ using the complete data $\{(y_i, x_i, \delta_i): \delta_i = 1\}$ and denote the estimated conditional distribution by $\hat{f}(y_i \mid x_i)$. For methods to estimate $f(y_i \mid x_i)$ under complex sample designs, see Kim and Shao (2013). By the assumption that the response mechanism satisfies missing at random in the population, $\hat{f}(y_i \mid x_i)$ is an estimate of $f(y_i \mid x_i, \delta_i = 0)$. This justifies obtaining an imputed value $y_i^*$ based on the estimated distribution $\hat{f}(y_i^* \mid x_i)$. Estimates are then constructed using the union of the complete data $\{(y_i, x_i, \delta_i): \delta_i = 1\}$ and the imputed data for nonrespondents $\{(y_i^*, x_i, \delta_i): \delta_i = 0\}$.

In practice, accounting for the variance associated with imputation is important for accurate accounting for uncertainty arising from the use of the imputed $y_i^*$ in place of the true, but unobserved, $y_i$. One class of approaches for estimating the variance associated with imputation involves the generation of several imputed values, which can also reduce the variability of the imputed estimator.
Fractional imputation (Kim and Shao, 2013) enables consistent variance estimates in the frequentist sense. Multiple imputation of Rubin (2004) is a Bayesian approach.

4.5.2. WEIGHTING

An alternative non-response adjustment involves modifying weights for respondents. Calibration and propensity scores are two techniques for determining weights to adjust for non-response. In calibration, the weight is determined so that the mean of the auxiliary variables across sampled units is equal to the mean based on the administrative data. A propensity score is an estimate of the probability that unit “i” responds. Both methods require auxiliary information (Lundstrom and Sarndal, 2005), which may be derived from administrative sources.

Example 4.5.2. Geuzinge, Rooijen and Bakker (2000) describe the use of administrative records to construct calibration weights to reduce non-response bias in household surveys. In one application, administrative registers of jobs and social security benefits were used to weight the respondents to the 1995 Netherlands Health Interview Survey. The theory was that individuals with greater health problems and greater use of medical resources were more likely to respond to the survey as a result of greater interest in health care processes. The concern of the statistical agency was that without adjustment an estimate of medical use based on the complete data would over-estimate the true health-care cost. Estimates of days in hospital that incorporate the administrative data were smaller than corresponding estimates based only on the unweighted complete survey data. The weights were also applied to obtain estimates of education levels. The weighted estimates of the proportion of individuals educated beyond higher secondary level was smaller than the unweighted estimate. The reason for this result was thought to be that individuals with higher education levels had higher response probabilities because they had a better understanding of the usefulness of the survey and greater trust in the Government.

4.6. MODEL BASED SMALL-AREA ESTIMATION AND FORECASTING

Many statistical procedures for obtaining estimates for small areas or to forecast a future outcome are based on models. In the case of small-area estimation, population information at the level of the small domain of interest is critical for improving efficiency. If the objective is forecasting or improving the timeliness of estimates, auxiliary information that reflects a more recent time period or changes over time can reduce the mean squared error of forecasts. In the construction of the 1997 NRI estimates, administrative data on transport was used to create small-area estimates of road area (Nusser and Goebel, 1997; Wang and Fuller, 2003). Section 4.6.1 gives a general model framework for constructing small-area estimates using administrative data as covariates, and discusses an application of the method to the Small Area Income and Poverty Estimation project of the United States Census Bureau.
4.6.1. ADMINISTRATIVE DATA AS AUXILIARY INFORMATION IN SMALL-AREA ESTIMATION

Rao (2003) gives a comprehensive account of methods and applications of small-area estimation; Pfeffermann (2013) discusses developments in the first decade of this century with regard to handling statistical issues such as unequal variances and outliers. Both emphasize the importance of reliable auxiliary information. Rao (2003) states: “Availability of good auxiliary information and determination of suitable linking models are crucial to the formation of indirect estimators.” Pfeffermann (2013) echoes this: “A common feature of design- and model-based SAE is the use of auxiliary covariate information, as obtained from large surveys and/or administrative records…Some estimators only require knowledge of the covariates for the sampled units and the true area means of these covariates. Other estimators require knowledge of the covariates for every unit in the population. The use of auxiliary information for SAE is vital because with the small sample sizes often encountered in practice, even the most elaborated model can be of little help if it does not involve a set of covariates with good predictive power for the small area quantities of interest.”

Rao (2003) classifies small-area estimation methods as unit-level models and area-level models. The former was initially introduced by Batteese, Harter and Fuller (1988), the latter by Fay and Herriot (1979). We consider the area-level model, introducing the following notation:

- Areas. Let $k$ denote the small areas of interest, where $k=1,...,m$, and $m$ is the number of small areas.
- Parameters. Let $Y_k$ denote the parameter of interest in area $k$. The $Y_k$ typically has the interpretation of a population mean of a response variable $Y$ for units in the population of area $k$.
- Direct estimators. Let $\hat{Y}_k$ be a survey-based estimator of the quantity of interest $Y_k$. In the simplest case, $\hat{Y}_k$ is a sample mean of the respondents in area $k$. For unequal probability samples, $\hat{Y}_k$ may be a Horvitz-Thompson or Hajek mean (Sarndal, Swenson and Wretman, 2005). If calibration procedures (see Section 4.4) are used, the weights used to construct the weighted sum defining $\hat{Y}_k$ are adjusted to preserve population-level control totals.
- Auxiliary information. Assume that external information such as data on tax or crop insurance is available, and suppose that the auxiliary information is obtained from an administrative source external to the survey. The auxiliary variable may be known for all units in the administrative data file, but for this development we only require the population mean or total of the auxiliary variable for area $k$. We use
\(x_{Nk}\) to denote the vector of auxiliary variables aggregated to the level of the population for area \(k\).

- Models. The small-area predictor is based on the following area-level mixed effects model. Assume

\[
\hat{Y}_k = Y_k + e_k,
\]

where \(e_k \sim N(0, \sigma^2_{ek})\), and \(\sigma^2_{ek}\) is the known sampling variance. Efficiency gains are possible through specification of a model relating \(Y_k\) to the auxiliary controls. Suppose:

\[
Y_k = x'_{Nk}\beta + u_k,
\]

where \(u_k \sim N(0, \sigma^2_u)\). An estimator of the best linear unbiased predictor of \(Y_k\) is:

\[
\hat{Y}_k = x'_{Nk}\hat{\beta} + \hat{\gamma}_k(Y_k - x'_{Nk}\hat{\beta}),
\]

where \(\hat{\gamma}_k = (\sigma^2_{ek} + \sigma^2_u)^{-1}\sigma^2_u\), and \(\hat{\beta}\) and \(\hat{\sigma}^2_u\) are estimates obtained, for example, from maximum likelihood or restricted maximum likelihood. Under the model assumptions and assuming the number of areas is large, the mean squared error of the predictor \(\hat{Y}_k\) is approximately the product of the variance of the survey estimator and \(\gamma_k\), where \(\gamma_k = (\sigma^2_{ek} + \sigma^2_u)^{-1}\sigma^2_u\). The ratio of the mean squared error of the direct estimator \(\hat{Y}_k\) to the mean squared error of the predictor \(\hat{Y}_k\) is approximately equal to \(\gamma_k\). The smaller the \(\gamma_k\), the larger the efficiency gain from the prediction model.

**Example 4.6.1.** Maples and Bell (2007) discuss the role of small-area models in developing estimates related to income for counties and school districts. The United States Small Area Income and Poverty Estimates (SAIPE) provide timely estimates of income and poverty at school district levels with a view to allocating state and federal funds for additional resources in low-income areas. The Annual Social and Economic Supplement of the Current Population Survey provides survey-based information for constructing the desired estimates. The survey is designed, however, to produce estimates at the national level, and sample sizes for counties and school districts are often too small to produce reliable estimates. Auxiliary data from various administrative sources are integrated into the programme to obtain additional information for estimating the poverty and income levels in the small geographic domains of interest. Maples and Bell (2007) explore the use of income tax data and data on food stamp participation in constructing model-based estimators of small-area income and poverty.

This example is not specifically related to agriculture, but it contains lessons applicable to an agricultural context in developing countries: it demonstrates in particular the value of forming relationships with administrative agencies to facilitate the use of administrative data for non-statistical purposes and it illustrates the importance of utilizing administrative data sources that are correlated with the quantity of interest in a survey.
The USDA has investigated the use of several sources of administrative data to construct small area estimates and forecasts. Nandram, Berg and Barboza (2013) explore the use of meteorological and satellite data to improve the precision of NASS crop-yield forecasts, and Bellow and Lahiri (2011) use satellite data to improve small-area estimates of crop yields. Satellite data are not considered an administrative data source for the purposes of this project, but a discussion of its use in small-area estimation is informative. The method applies to a situation in which the auxiliary information is obtained from an administrative data source instead of remote sensors.

4.7. CUT-OFF SURVEYS

In a cut-off survey, where part of the questionnaire is administered to a portion of the population, the required information is obtained from an external source such as a survey or administrative database. Examples of the kinds of external data sources used in cut-off surveys include tax data and information from private corporations. Statistical agencies in many countries apply cut-off strategies in surveys related, for example, to business establishments or energy use.

The design of cut-off surveys relies on an auxiliary variable known for the full population. A common approach begins with ordering the population of interest with respect to a measure of size, because the size measure associated with a unit is often indicative of the importance of the unit to the overall estimate. In surveys of business establishments or agricultural operations, for example, the size measure may be related to total employment or to farm area. In typical applications of cut-off sampling, units with a size measure below the cut-off value are not included. Cut-off sampling may be viewed as a relative of

| Example 4.6.2. Satellite data are used for small-area estimation and for crop area and yield. In the Battese, Harter and Fuller (1988) estimation of the area of corn and soybeans in 12 Iowa counties the auxiliary information in the small-area model was the number of pixels in the county classified as corn or soybeans; because the satellite data were highly correlated with the survey data the small-area models led to reductions in mean squared error in the small-area predictors. Bellow and Lahiri (2011) employed the normalized difference vegetative index of greenness as a covariate for estimating crop yields at the county level. Johnson (2014) extends the use of the normalized difference vegetative index to forecast crop yield at the county level. Statistics Canada (2012) states: “In China remote sensing is used to monitor crop area change, crop yields, production and growth, drought and other agriculture-related information for five main crops. In Europe, it is used to monitor crop vegetation growth (seven crop types) and to provide annual crop production forecasts.” The Remote Sensing and Geospatial Analysis section at Statistics Canada developed an experimental model to estimate yields of spring wheat, durum wheat, barley and canola in western Canada. |
probability proportional to size sampling, where the size measure associated with certain units in the population is zero.

Reliable auxiliary information is essential at the estimation stage. Surrogates for the responses to target questions of interest for units in the population that were not part of the data collection are required. They are often derived from administrative sources that collect similar or related information. Concepts measured in the external data source may differ from the target variable of interest to the survey as a result of differences in reference periods, coverage or definitions. In such cases, models may be needed to calibrate the variables available in the administrative file to the target concepts of interest in the survey.

The primary motives for using cut-off surveys are to reduce costs and the burden on respondents. A cut-off survey can also improve the efficiency of data collection if units in the portion of the population that is not surveyed are difficult to contact. Where the collection of accurate data is a challenge and administrative data are reliable, the data obtained from a cut-off survey will be of better quality than data from other sources. Yorgason et al. (2011) state: “In general, cut-off sampling is an effective sampling method when one has a good working knowledge of the characteristics of the total population and subpopulations of interest and is willing to take some calibrated risks with regard to the lack of information from the non-sampled subpopulations. Under many quite varied sets of circumstances, it can provide the required information at minimal cost.”

Cut-off sample designs are useful in surveys of businesses: the population of interest is often highly skewed, with a relatively small fraction of the units in the population contributing to a large percentage of the overall estimator. Tax data may provide the necessary auxiliary information for sample design and estimation in cut-off surveys. The tax data may contain variables used to identify the size of the units in the population and variables correlated with the quantity of interest, which are then used in estimation (see examples 4.7.1–4.7.3).
4.8. USES IN DATA COLLECTION

Administrative data can also facilitate data-collection processes, especially in cases where the characteristics or identities of the sampled units are unknown until contact is established. A significant case, especially in agricultural contexts, is surveys based partly or entirely on area frames. In the examples 4.8.1 and 4.8.2, administrative sources provide lists of names and addresses that are useful in contacting units originally sampled from an area frame rather than a list frame.

Example 4.8.1. The Swedish Business Survey of Investments uses a cut-off sample design to reduce the burden on respondents. Data are collected for all enterprises with a tax record indicating that they made an investment in excess of €500,000. Enterprises with investments below the threshold are not surveyed, and a model is used to estimate survey variables from the available tax data (Berg and Hall, 2007).

Example 4.8.2. Statistics Canada used a cut-off survey design to reduce the burden on small business respondents in the late 1990s. Businesses that were too small to contribute substantially to the overall estimate were placed in a “take-none” stratum, and tax data were used to produce estimates for these units (Yung, Rancourt and Hidiroglou, 2007).

Example 4.8.3. The statistical office of the Republic of Slovenia uses a cut-off survey to improve the timeliness of estimates of monthly turnover indexes. In the population of businesses, the largest 3 percent of units account for more than 50 percent of total turnover. A classical questionnaire is administered to estimate turnover for the 3 percent, and tax data are used to estimate monthly turnover in the remaining units in the population (Seljak, 2007).
Defining the Quality of Administrative Data

The quality attributes considered in this section are classified in the “performance” category of the Administrative Data System for Agricultural Statistics (ADSAS) framework in Table 2.1 in Part B of Technical Report 1, but “structural” or “conduct” issues are also relevant. The sustainability of an administrative source, a structural example, depends on the extent to which it is used by the agricultural sector: if the sector depends on the source, its reliability and stability can be improved over time. Conduct issues such as data collection for raw materials and processes for transferring data can affect the accuracy and timeliness of administrative data.

Groves (2004) establishes a quality framework that distinguishes errors arising from random sampling from errors arising from other aspects of statistical production; his model for total survey error is illustrated in Figure 3. One of the benefits of this representation is the depiction of the ways in which errors arise at various steps in statistical processing.

Figure 3. The Total Survey Error Paradigm

Bakker (2010) discusses the advantages of applying the Groves (2004) framework, and Zhang (2012) considers extending it to quality assessments of administrative data. Bakker (2010) notes that many of the quality metrics that apply to survey data also apply to administrative data. Zhang (2012) elaborates on Groves’ representation of total survey error with administrative data in mind, identifying two phases in the life-cycle – data acquisition and the combination of data from different sources. Figure 4 in Zhang (2012) illustrates the quality attributes and error sources that arise during the development of statistical micro-data: the ovals indicate the errors that may arise in transferring a data concept from one rectangle to the rectangle beneath.

Figure 4. The Two-Phase Life Cycle of Administrative Data

In converting a “target concept” such as a farm to a “target measure” such as a “unit that gains at least 1,000 dollars in agricultural production each year”, for example, the “validity error” associated with the risk that the target measure may not accurately represent the concept of interest must be avoided. The
quality frameworks in Groves (2004) provide the foundation of quality assessments of administrative data.

Daas and Fonville (2007) define administrative data quality as “the capability of [administrative] data to be used effectively, economically, and rapidly to inform and evaluate decisions.” They distinguish two types of quality: i) source-specific quality, which relates to the quality of the input data; and ii) product-specific quality, which refers to the usability of the data for statistical purposes. They define quality indicators such as clarity, specificity of administrative concepts, coverage, reference time, data freshness, errors in data, completeness, record matching ability, confidentiality and privacy, compatibility between file formats, comparability of administrative datasets and anticipated use. These metrics encompass the Eurostat quality indicators: relevance, completeness, punctuality, coherence, comparability and accuracy.

United Nations (2011) defines quality as the degree to which an administrative source meets the demands of the user. As discussed in Section 2 administrative data have various roles in the production of official statistics, so an evaluation of the quality of such sources must be specified in relation to the intended use. Laitila, Wallgren and Wallgren (2011) identify three classes of use according to the extent to which administrative data are used directly to form statistical products. But regardless of intended use, specific quality dimensions must be identified.

5.1. QUALITY DIMENSIONS

Several lists of criteria for evaluating the quality of a statistical product have been developed by statistical agencies and in academic journals. The following list is based on United Nations (2011), and is also discussed in Section 6 of Technical Report 1.

- Relevance – the extent to which an administrative source aligns with current and future uses and the “target concept” in Figure 4. Because administrative data sources have many potential uses, relevance must be considered with respect to specific applications (see Section 5.2).
- Accuracy – the closeness of the match between the measurement or estimate and the value of interest. This applies to estimates of aggregate quantities such as means or totals and also to micro-data. Reducing the types of error source shown in the ovals can help to improve accuracy (see Section 5.3).
- Timeliness and punctuality – the time between the occurrence of an event of interest and the date on which the administrative data are available for use in official statistics. It is important to factor editing,
modelling and processing times into the equation (see Section 7). An administrative source should not be considered ready for use until these processes are complete. Punctuality reflects the ability of an administrative agency to meet deadlines: it is the time between the date on which data are released and the announced release date.

- **Accessibility** – the ease with which users can obtain administrative data. Confidentiality conditions and policy and legal constraints can hinder access to administrative data.
- **Clarity and interpretability** – the ease with which users can understand administrative data. Information about the administrative data source can improve its clarity and interpretability of the source, and metadata such as maps, summary tables or documentation on collection procedures can improve users’ ability to interpret the data.
- **Coherence and consistency** – the compatibility between administrative data sources or between administrative and survey data. Several data sources may measure a target concept (see Figure 4), but different processes can lead to differences among means. Differences in definitions and data collection can adversely affect coherence (see Sections 6 and 7).
- **Comparability** – the validity of comparisons among estimates based on related administrative sources; this is related to coherence and consistency and can be affected by methodological differences in administrative or survey sources and differences in definitions. Assessing comparability involves determining how observed differences result from differences between the true underlying concepts of interest or the differences in observed values arising from measurement processes. Laitila, Wallgren and Wallgren (2011) and United Nations (2011) distinguish comparability over time, space and statistical domains such as strata.

### 5.2. FITNESS FOR USE: EVALUATING QUALITY WITH RESPECT TO INTENDED USE

As discussed in Section 2, administrative data have various uses in official statistics, and a single administrative source may have numerous applications; the dimensions discussed in Section 5.1 may have different interpretations depending on intended use. Laitila, Wallgren and Wallgren (2011) define three forms of quality in relation to three classes of use: i) in “output quality” the administrative source supplies a statistical product of interest directly; direct tabulation is an example; ii) in “input quality” the administrative data source is used only after refinement or in conjunction with other sources; using
administrative data as a partial substitute for survey data is an example; iii) in “production process quality” the administrative data source is used to improve the quality of estimates based on sample surveys or other administrative sources; calibration and small area-estimation are processes in which production process quality measures should be considered.

5.2.1. OUTPUT DATA QUALITY

If an administrative source is to be used for direct tabulation of official statistics, the output quality is pertinent, and the quality indicators listed in Section 5.1, which generally apply to surveys and censuses, will apply to the administrative source.

Example 5.2.1. The annual pay register of Statistics Sweden is delivered by the Swedish tax agency and used directly to estimate wage sums. The source is considered to have output of high quality, partly because of coherence between the definition of wages used by the administrative office and the definition required by users of the statistical product. Comparability with the business register is an advantage in that merging the pay register with the business register enables the production of estimates of wage sums by sector of economic activity (Laitila, Wallgren and Wallgren, 2011).

5.2.2. INPUT DATA QUALITY

Administrative data may only be suitable for generating final statistical products after refinement or integration with other data. Laitila, Wallgren and Wallgren (2011) consider two cases: i) the administrative source is used independently for the statistical process after cleaning and editing; and ii) micro-data from the administrative source are combined with micro-data from other registers and surveys, and the data from the completed register system are used to form the administrative database. The input data quality of the administrative source should be evaluated in both cases.

The criteria for evaluating input data quality are less restrictive than those for output data quality because fewer demands are placed on the administrative source. Coordination with the statistical agency may be needed to improve accessibility. Linking variables may have to be converted to a form compatible with the statistical production system, and units may have to be re-defined or aggregated, a process known as “profiling” in the literature on business registers (see Section 7). If the coverage of an administrative source is thought to be incomplete, the source may be added to a register system of administrative files from other agencies, censuses or geographic information systems. A critical component of input data quality is “linking variables”: a unique personal identification number, for example, can improve the feasibility of linking processes (Wallgren and Wallgren, 2010).
5.2.3 PRODUCTION PROCESS QUALITY

Section 4.1 describes ways in which an administrative source may be used to improve estimates based on surveys. For calibration and area-level small-area models the requirement for micro-data is reduced, which in turn reduces the need for compatible identifying variables. Because relationships between means based on administrative sources and means obtained from surveys are estimated, administrative data do not need to provide an unbiased measure of the true underlying quantity of interest. This reduces the amount of coherence and accuracy needed in the data source relative to direct tabulation (Laitila, Wallgren and Wallgren, 2011). Timeliness, punctuality and accessibility are important in process quality, because obtaining auxiliary information in a timely manner is essential to give time for model building and evaluation. Analysts often need to choose between correlated variables from administrative sources, even for model-assisted estimators. Evaluations of trade-offs between bias and variance also require attention.

**Example 5.2.2.** Quality dimensions for the construction of small-area income and poverty estimates at the United States Census Bureau (see Section 4.6). Accessibility and timeliness are important because logistics systems are needed to enable the bureau to access tax information efficiently without compromising confidentiality. Coherence and consistency are evaluated primarily through correlations between means based on tax records and means based on the current population survey: high correlation indicates potential for efficiency gains in estimation using the administrative source to estimate the proportion of school-age children in poverty in small areas (Maples and Bell, 2007).
Challenges and Limitations of Administrative Data

Sections 3 and 4 explain that administrative data have many uses and the potential to improve the quality of statistical products. But statisticians must be aware of the challenges and limitations of most administrative data sources, many of which arise from collection processes. Unlike statistical surveys and censuses, administrative data are gathered for purposes other than those of a statistical operation, and data may be entered by individuals who are not governed by the protocols of surveys. Challenges arising from the nature of administrative data can affect quality and limit their usefulness; challenges associated with lack of standardization, lack of control over suppliers’ methods, resistance from the public, incomplete coverage of populations and concepts of interest, and difficulties associated with computation and methods are discussed below.

6.1. LACK OF STANDARDIZATION

Several applications of administrative data in the production of official statistics involve integration of the administrative data with other such data or with surveys or censuses. Challenges arise when different data sources employ different definitions or coding systems. Simple differences in labels identifying units in micro-data, for example, can make it difficult to link disparate data. This section discuss ways in which differences between definitions of units, definitions of variables and differences in coding systems can create challenges.

6.1.1. DEFINITIONS OF UNITS

The unit, in this context, refers to the smallest reporting entity in a micro-data file. Definitions of units in administrative files are generally driven by the function of the administrative agency, so its definition of a unit may differ from the definition of a unit by the statistical office. According to the Farm Services Agency, an individual farm producer may, for example, be associated with
several farms as defined by NASS. As a consequence, studies show that inconsistency among definitions of units results in increases in the burden of response when the Farm Services Agency database is used as a sampling frame instead of the NASS list frame (Barboza and Harris, 2009; Beckler, 2013). Because numerous individual farms may comprise a single operation, the unit in the administrative database may differ from the statistical unit. The unit-related challenge facing statisticians tops the list of concerns in Benedetti et al., 2010, particularly with regard to the large choice of units – family, agricultural holding, household, parcel of land or point – and dependence on the availability of a quality frame of units. Differences between definitions of units can limit the utility of the administrative source, especially for purposes that involve linking micro-data.

Example 6.1.1. Wallgren and Wallgren (2010) give an example of different legal units posing a problem in combining the strength of the value-added tax (VAT) register and the IACS register. Husbands and wives, for example, both report income but if one of them applies for a subsidy the treatment is different from the case in which both apply. An enterprise with two agricultural units and holdings is treated differently from a company having one agricultural unit and another that is only 40 percent involved in agriculture. Techniques to harmonize units include averaging and splitting in clerical or model-based operations. In any case mismatches would result in missing values or misleading values.

Example 6.1.2. In Australia, statistical offices and administrative offices have different definitions of a dairy cow (Boreo, personal communication).

Example 6.1.3. The Italian team of the European Union Statistics on Income and Living Condition project investigated the use of administrative data from a pensions register, which is the only administrative source that enumerates all recipients of pensions and includes classifications according different kinds of pensions. A limitation of the register is that monthly pensions before tax reported in the register do not coincide with the three target variables of the project (Consolini, 2008).

Example 6.1.4. Statistics Canada uses customs data on imports and exports to improve estimates of agricultural production. Customs documents provide information on agricultural imports to and exports from the United States. The United States customs provide information about imports to and exports from Canada. Even after ten years of attempting to reconcile the two sources the statistics, differences remained. Ultimately, Statistics Canada adopted the United States figure for exports to Canada and the United States Department of Commerce accepted the Canadian figure for exports to the United States (Trant and Whitridge, 2000).

6.1.2. DEFINITIONS OF VARIABLES

Differences in the definitions of related concepts can produce inconsistencies between administrative and statistical sources. The definition of income for tax purposes, for example, may differ from the definition of needed for a survey. Such differences can cause systematic deviations between quantities derived from administrative sources and corresponding quantities obtained from surveys (United Nations, 2011; Wallgren and Wallgren, 2010; Carfagna and Carfagna, 2010; Brackstone, 1987). Definitions and content in administrative
records are sometimes changed without notice or a grace period in which new and old definitions are reported simultaneously.

In the absence of an overlap period in which data are collected with both definitions, real change cannot be separated from the effect of the revised definitions. The effects of changes in definitions are more pronounced when files such as registers are updated continuously (United Nations Economic and Social Council, 2007).

### Example 6.1.5
At the beginning of 2005, the law providing for a reduced tax rate on incomes below 50 percent of the average income was changed to provide for a reduced rate on incomes below 60 percent of the mean wage. The Israeli Central Bureau of Statistics, unaware of the change, continued to publish the data without modifying the definitions, hence raising concerns about a new trend in the economy – an increase in posts for low-wage employees. Several months passed before the discrepancy was detected; and even when it was detected there was no way to prevent a break in the series (United Nations Economic and Social Council, 2007).

### Example 6.1.6
Statistics Slovenia uses tax data to estimate turnover in businesses. Turnover indexes based on survey and tax data tend to move in the same direction. Discrepancies occur when relevant events for tax purposes differ from those of interest in surveys; in particular, sales of real estate are reported to the tax agency but should not be included in the definition of turnover (Seljak, 2007).

### 6.1.3. DIFFERENCES IN CODING SYSTEMS
Challenges that arise as a result of using different coding systems are related to those associated with differences in definitions of variables and units: a statistical office, for example, may require a more granular coding system than the system needed by an administrative agency. In merging sources with different coding systems, inconclusive situations arise when one code in the administrative source maps to several codes in the system used by the statistical office (United Nations, 2011) (see Section 7).

### Example 6.1.7
The NASS cropland data layer classifies 30x30 meter pixels in a mutually exclusive set of land-cover categories. In the National Resources Inventory, a longitudinal survey of United States land, data collectors classify sampled locations similarly. Model-based estimation methods for the National Resources Inventory incorporate the Cropland Data Layer as auxiliary information. Because the coding systems used in the National Resources Inventory and the Cropland Data Layer are not identical, exploratory data analysis of relationships between codes in the two systems is used to develop a system for aligning their codes (Demuth, 2015).

### 6.2. LACK OF CONTROL
Because statistical offices do not collect administrative data, they have limited control over data-collection methods. The schedule of collection and release of administrative data may be subject to legal and policy-related processes that are difficult to manage or change. Changes in policy may alter the availability, characteristics or quality of administrative data, which can produce
inconsistencies over time. Many kinds of administrative data are entered by untrained individuals in the population, which can lead to errors and variability. This section discusses ways in which changes in policy, legal and political constraints and uncontrolled data collection can affect the quality and usability of administrative data.

6.2.1. CHANGES IN POLICY

Changes in the policies and funding levels of administrative agencies, which are usually outside their control, can affect the quality of administrative data, change the way they are collected and affect their availability and accessibility (United Nations, 2011). Brackstone (1987) explains: “Discontinuities in time series based on administrative records can be caused by simple changes in the coverage of a program, the introduction of an incentive to join or leave a program, or procedural changes that affect quality or completeness of records.” Statistical offices are advised to work with administrative agencies to anticipate such changes and mitigate their effects (Brackstone, 1987; Iwig et al., 2013) (see Section 7).

In some cases, a change in policy may improve the quality of administrative data. Brackstone (1987) cites an example in which a change in the method of collecting tax data improved the coverage of the tax register: the introduction of the child tax credit in Canada required low earners to file a tax return to claim the credit, thereby improving the coverage of the tax register and the resulting estimates.

Example 6.2.1. In a report on the use of the Austrian Activity Register in the register-based census, Reiner and Schodl (2008) report the effects of legal constraints on the timeliness of available data. One potential database, the personal tax database, contains “basic personal data of all tax paying Austrians, including self-employed” (Reiner and Schodl, 2008). Legal and technical constraints cause a lag of up to three years in the availability of the database, so use of the personal tax database for the Austrian Activity Register is judged impractical.

6.2.2. LEGAL AND POLITICAL CONSTRAINTS

Access to administrative data may be limited by legal and political constraints, for example to protect the confidentiality of individuals in the population of interest (Brackstone, 1987). A memorandum of understanding is often needed to establish a flow of data from an administrative agency to a statistical office by explaining the objectives of the statistical office and the data requirements of the administrative office to meet them. Developing such a memorandum can be expensive and time-consuming (Prell et al., 2009). Guidelines for writing a memorandum of understanding are given in Section 7.
Statistical agencies are committed to preserving the confidentiality of their data. The complexity of the requirements increases when administrative data collected and maintained by other agencies is involved. Statistics New Zealand (undated) elucidates the complexities of adhering to legal and political requirements: “Consideration of the legal, confidentiality and privacy issues associated with administrative data are also very important for Statistics New Zealand. Not only must Statistics New Zealand meet the requirements of the Statistics Act 1975, it must ensure that the security and other requirements of legislation such as the Tax Administration Act relating to access and use of individual administrative data are also met.”

6.2.3. DATA COLLECTIONS PROCEDURES

The way in which administrative data are collected is largely outside the control of statistical agencies. Tax forms, for example, are generally completed by individuals, and agricultural returns are completed by agricultural extension staff or even chiefs in many developing countries. These forms of data collection lack standardization and can lead to reporting errors and inconsistencies (United Nations, 2011). In some cases, bias arises from programme-induced incentives (Brackstone, 1987; Carfagna and Carfagna, 2010). Casual data-collection procedures may be acceptable for the purposes of an administrative office, but users of the statistical product usually demand a higher level of consistency and accuracy.

Example 6.2.2. The crop areas in the IACS register are reported in order to determine subsidies. The administrative agency conducts checks and audits, with severe consequences if over-reporting of crop areas is detected. To protect against the financial consequences of inadvertent over-reporting, farmers are likely to under-report crop areas. In Sweden, comparisons between administrative and survey data indicate that the bias resulting from deliberate under-reporting can be as large as 20 percent (Carfagna and Carfagna, 2010).

6.3. COVERAGE OF POPULATION AND CONCEPTS

An administrative database may not cover the full population of interest and may not include all necessary data items; it may, for example, consist only of participants in a government programme (Carfagna and Carfagna, 2010). The administrative process may only collect data on certain items or may aggregate detailed categories that are individually of interest to a statistical office (United Nations, 2011; Wallgren and Wallgren, 2010). Information may be missing for sub-domains of the population of interest: in evaluating the possibility of a register-based census for New Zealand, for example, one consideration was that educational qualifications gained abroad would not be included for new immigrants and that emigrants might fail to de-register. This issue is
particularly problematic in countries with substantial international mobility (Bycroft, 2010). Because of these limitations, it may be necessary to combine several administrative databases to create a register system with sufficient coverage. A related solution involves integrating administrative and survey data. Because of the importance of pooling data sources to overcome under-coverage problems, methods and techniques for linking records and datasets feature largely in the literature. Difficulties with linking records are discussed in Section 6.4; associated solutions are overviewed in sections 7 and 8.

6.4. COMPUTATIONAL AND METHODOLOGICAL CHALLENGES

Administrative data often arrive at statistical offices in large volumes and require editing, imputation and integration with other sources before use in statistical products. Managing such large quantities of data leads to significant computational challenges. Because there is rarely a one-to-one correspondence between identifying variables in administrative and survey datasets, linkage is not straightforward. Procedures for cleaning data and merging data from different sources often rely on explicit or implicit model assumptions. Errors that arise in data cleaning can propagate to the final statistical product.

6.4.1. COMPUTATIONAL CHALLENGES

Increases in the capacity and speed of computing environments make the idea of utilizing massive quantities of administrative data increasingly appealing. Although computational resources continue to expand, managing large volumes of data from disparate sources still poses significant IT and computing challenge.

**Example 6.3.1.** Administrative data from a farm-subsidy programme provides Statistics Sweden with information on areas planted with various crops. Because of the timing of data collection and release, the administrative database does not contain information on crop yield and so surveys and censuses are still needed to estimate crop yield (Wallgren and Wallgren, 2010).

**Example 6.3.2.** Data from the Department of Animal Husbandry are used for livestock statistics in India. Under-coverage is a concern because the database does not include unregistered farms (Government of India, 2013).
6.4.2 METHODOLOGICAL CHALLENGES

Cibella and Tuoto (2008) report a survey of national statistical institutes in Europe that found that most applications of administrative data to the production of official statistics involved harmonizing several data sources. Lack of standardization across data sources, for example, and the use of different identifying variables leads to challenges in merging different data sources. Editing and imputation are needed to improve consistency across data sources or improve the coverage of a register, and developing models to support these processes can be difficult.

Example 6.4.1. Reiner and Schodl (2008) discuss computational challenges in a report on an unemployment database used for the Austrian register-based census: “We are facing difficulties regarding the performance of low-level queries. This is a result of the huge data base: for one single reference day there are about 11 million observations. The problem arises from two requirements. Firstly a great flexibility on the reference time (different reference days, reference weeks, and reference months) is needed by the teams of analysts, which enlarges the underlying data base enormously. Secondly the counting on a personal level is often needed, while data in general is provided on employment or insurance case level, so there is more than one observation per person.”

Example 6.4.2. Dinges (2008) discusses issues related to integrating disparate data sources in the production of structural business statistics in Austria. A cut-off survey was used to obtain information from large and medium sized enterprises; data on small businesses were obtained from the business register and tax register. Because of differences between variables in the registers and variables of interest for the survey, model-based techniques were used to construct a single harmonized dataset. Additional challenges arose from missing tax declarations, incomplete links and differences in turnover definitions across data sources. A subjective assessment by domain specialists of impacts of model assumptions indicated that bias could not be avoided. Nonetheless, the study concluded that the effects of modeling are small for many industry categories and that the approach of combining survey and administrative data worked well.

When records are linked through probabilistic linkage models or from incomplete linkage, the estimation procedures should account for linkage error. Quantifying the effect of this source of uncertainty on the mean square error of subsequent inferences is a methodological challenge (Carfagna and Carfagna, 2010). Methods such as those proposed by Kim and Chambers (2012), which extend the regression analysis techniques of Chambers (2009) to applications with more than two linked datasets, can be employed.
6.5. CONFIDENTIALITY AND PUBLIC PERCEPTION

The issue of confidentiality is complex, especially with regard to administrative data. Consider two situations: i) some information in the administrative database is not protected by laws that ensure secrecy: in this case the issue of confidentiality is not an obstacle to data sharing; and ii) the administrative information is confidential: in this case an agreement between the national statistical office and the administrative office is needed for access by the statistical office; permission may come in the form of a memorandum of understanding, a re-definition of the statistical system or a government decision (see Section 7).

Example 6.5.1. Restrictions related to confidentiality can exacerbate difficulties associated with record linkage. In Italy, for example, administrative records are linked with interview surveys to improve the quality of data on income. In view of regulations to preserve confidentiality, the Italian tax agency performs record linkage and then provides the statistical institute with the desired tax information for each sampled individual (Consolini, 2008). This constraint further limits the statistical agency’s control over the administrative database.

Example 6.5.2. Experiences at Statistics New Zealand illustrate the complexities associated with preserving confidentiality when data from several agencies are being used. The tax administration act permits administrative agencies to transmit their data to Statistics New Zealand, but it also prevents Statistics New Zealand from sharing unit-level data about businesses or people with other government departments, including those involved in judiciary issues (Statistics New Zealand, undated).

In the second case in paragraph 117, individuals and enterprises may give information to the administrative agency on the understanding that it will remain confidential. As a result the use of administrative data for statistical purposes can be met with public scepticism (Brackstone, 1987). To address public concerns about privacy, the statistical office may be advised to take robust measures to ensure the confidentiality of administrative data: at Statistics Canada, for example, administrative tax data are housed in a secure area. The need to accommodate public concerns about privacy and confidentiality can increase the cost of managing administrative data and limit accessibility.

6.6. METHODOLOGICAL CHALLENGES IN THE CONSTRUCTION OF THE AUSTRIAN BUSINESS REGISTER

Austria maintains a comprehensive business register based on at least four administrative sources; it is used primarily to reduce the burden on respondents. Haslinger (2008) highlights many of the challenges associated with administrative data discussed in sections 6.1–6.5; each of the four main contributing registers provides an example of one of the drawbacks discussed
above. The ideas related to the Austrian structural business statistics apply generally and are relevant for agriculture because an agricultural operation is a form of business.

A challenge associated with maintaining the business register is the length of time between the collection and release of the administrative data. The most comprehensive administrative register used to update the business register is the tax register: it is judged to have broad coverage, but the time between the fiscal year and the time by which all units have received their tax assessment is two or three years; as a result a realistic value for total turnover in 2000, for example was not available until 2003.

The Register of the Economic Chamber, the oldest contributor to the business register, is obtained from applications for trade licences; these are often required for entrepreneurs. This register suffers from under-coverage because not all economic units need to become members of the Economic Chamber.

A third contributor is the social-security register. Austrian employers are required to register employees in one of many social-security insurance institutions, and a single employer may have employees registered at two or more institutions. A difficulty associated with this register is that the units are different from the units in the business register: a unit in the business register may correspond to an arbitrary number of units in the social-security register, which makes it difficult to link operations.

Despite the limitations of individual registers, the register system as a whole provides useful information for the statistical agency. In particular, the use of register-based statistics reduces the burden on respondents. The next section discusses strategies for making use of administrative data.
Mechanisms to Improve the Quality of Estimates Derived from Administrative Data

Statistical agencies in developed countries have mechanisms to overcome the challenges discussed in Section 6. Engaging with administrative offices and with the public can help to align definitions needed for statistical purposes with those used by administrative agencies, mitigate the effects of administrative changes on the usability of the data for statistical purposes and address concerns about privacy and security (Brackstone, 1987; United Nations, 2011). Approaches that combine administrative databases with information from surveys can reduce problems associated with measurement error, enable reconciliation between definitions in different sources and improve coverage (Wallgren and Wallgren, 2010). Audits and sampling of administrative data are used to check for errors and evaluate coverage (Carfagna and Carfagna, 2010). Adoption of best practices for quality control and assurance can help to manage errors in administrative data. These four techniques are discussed in sections 7.1–7.4; section 7.5 describes tools for implementing the techniques.

7.1. PROACTIVE INVOLVEMENT IN ADMINISTRATIVE PROCESSES AND ENGAGEMENT WITH THE PUBLIC

Engagement with the public and participation in decision-making as to the nature and accessibility of administrative data can help to improve usability. Legislation permitting statistical offices to influence and access administrative data can promote standardization and prevent time lapses between collection and release (United Nations, 2011; Brackstone, 1987). Bilateral committees
have fostered productive working relationships between administrative and statistical offices. Publicizing the measures used to ensure the confidentiality of administrative data and the public benefits derived from utilizing administrative data can help to minimize scepticism.

### 7.1.1. FRAMEWORKS FOR LEGISLATION AND POLICY

United Nations (2011) discusses the value of laws and policies enabling statistical offices to access administrative data. In many countries, legislation includes explicit provisions for access to administrative data: Ireland and Norway are examples. The 1993 Irish Statistics Act states: “For the purpose of assisting the [statistical] Office in the exercise of its functions under this Act, the Director General may by delivery of a notice request any public authority to – (a) allow officers of statistics at all reasonable times to have access to, inspect, and take copies of or extracts from any records in its charge, and (b) provide the Office, if any such officer so requires, with copies or extracts from any such record, and the public authority shall, subject to subsection (2) of this section, comply with any such request free of charge” (United Nations, 2011). Because opportunities to establish or change legislation are scarce, statistical offices are advised to propose legislation with a long-term strategy in mind (United Nations, 2011). Brackstone (1987) notes that a tax reform by the Government of Canada provided an opportunity for the statistical office to engage with policy makers and shape collected data to comply with its needs.

When influencing legal frameworks is impractical, policies may be established to facilitate access or changes to administrative data. Policies are easier to change than laws and tend to evolve over time (United Nations, 2011): an example of a policy framework involving administrative data is Principle 5 of the United Nations *Fundamental Principles for Official Statistics*, which emphasizes the cost-effectiveness of administrative data and promotes the use of such sources in the interest of making efficient use of information.

**Example 7.1.1.** The Austrian register-based census provides an example of a policy that permits improvement of the administrative database: “If an entry can only be found in the central population register without any entries in other registers, a special procedure called "residency analysis" is used which is a signs-of-life analysis. Statistics Austria is authorized to ask the registrar for the names and addresses of the people concerned, to whom a letter is addressed that they are obliged to answer. If there is no reaction to the letter and there are no other signs of existence, the person is supposed to be deleted from the census records” (Lenk, 2008).

**Example 7.1.2.** The Canadian statistics act provides the legal framework for many of the activities of Statistics Canada; an example is the census of agriculture. The act also permits the statistical office to access administrative data; Section 13 of the act enables Statistics Canada to obtain information from the AgriInsurance risk-management programme (Smith *et al*., 2013).
7.1.2. BILATERAL COMMITTEES AND ORGANIZATIONAL ARRANGEMENTS

Agreements are often needed to facilitate the transfer of data (United Nations, 2011): they may be in the form of memoranda of understanding specifying the objectives of the statistical office in using the administrative data and the information needed to achieve them (Prell et al., 2009). Brackstone (1987) cites the success of Statistics Canada in forming bilateral committees involving the statistical office and the administrative agency to develop organizational and technical infrastructures.

Prell et al. (2009) analyse seven cases involving written agreements establishing or expanding relationships with administrative agencies. They identify four characteristics of successful memoranda of understanding: i) vision and support by agency leaders; ii) precise but flexible goals; iii) infrastructure; and iv) mutual interest. These “elements for success” enable agencies to work through the challenges that arise when a memorandum of understanding is being established.

Iwig et al. (2013) provide a guide to interactions between statistical and administrative offices when forming a data-sharing relationship. Their “data quality assessment tool for administrative data” is based on relevance, accessibility, coherence, interpretability, accuracy and institutional environment, and they recommend questions that statistical offices should ask administrative agencies. To ensure coherence of concepts, classification and data-collection methods over time and across geographic domains, for example, they recommend the question: “Describe any classification systems used for categorizing or classifying the data. For example, do questions asked about race use the current race and ethnicity categories defined by the Office of Management and Budget? Do questions asked about industry classifications use the current North American Industry Classification System?”

Organizational agreements can also overcome restrictions associated with confidentiality. If statistical and administrative data are deemed to be confidential, a memorandum of understanding can provide the legal mechanism for data transfer (Prell et al., 2009). Expansion of the definition of the national statistical system can promote circulation of administrative data among government offices and statistical agencies. Statistics Sweden, for example, receives regular deliveries of administrative data from agencies responsible for government programmes and regulations (Wallgren and Wallgren, 2007).
Example 7.1.3. A case study discussed in Prell et al. (2009) involves an agreement between the USDA Farm Services Agency and NASS, which share common stakeholders and have a long history of engagement. The agency provides NASS with administrative data for frame maintenance in exchange for unpublished survey results. The case study aimed to identify avenues for expanding this relationship, and it exemplifies the value of well defined specifications when administrative data are involved: “It must be understood that a large relational database is not an Excel spreadsheet. There are millions, billions of rows of data. There are dozens of variables in each table . . . There are constant changes. There are new code developments and pre-existing code changes . . . It takes a long time before a database is released, which can involve going from 400 million records down to a first release of summarized totals of 60 million. FSA pulls a lot of data from different databases. There are accounting codes. There are transactions codes. It seems to be a simple request to provide data on “payments” to [agricultural] producers . . . It is not obvious what “payment” data is requested. There are different payments.”

Example 7.1.4. One of the challenges associated with the use of administrative data for official statistics in India stems from differences between requirements of administrative and statistical offices. Sen (undated) describes the challenge and an approach to resolving it: “[There] is often divergence between the nature of data required for administrative purposes, especially when the objective is to monitor programmes, and the nature of the data that would be required for statistical purposes. Since the data collection machinery is generally under administrative control of the programme authorities rather than the statistical authority, the introduction of appropriate questions and indicators quite often becomes a victim of the need to keep the record keeping process manageable. In recognition of this issue, recently the Indian Government has raised the status of the statistical officers in the line Ministries significantly and, hopefully, over time their voice would be heard more prominently while designing the administrative data collection system.”

Example 7.1.5. Several examples of inter-agency collaboration in using administrative data to generate statistics about human populations and health issues in Asia are cited in Asian Development Bank (2010). In Mongolia, the national statistical office collaborated with the Ministry of Health and the authority for state registration to reconcile their data on population, migration and vital events. In Palau, the ministries of education, health and finance combined their administrative data to produce statistics related to education and health. The Sri Lanka Ministry of Education coordinated with the departments of the Registrar General and census and statistics to produce annual indicators of progress in improving education: estimates based on different types of data provided by the three agencies were produced for detailed geographic domains by sex. The process revealed related data that contributed to differences between the three sources and led to recommendations for improving data quality.

7.1.3. PUBLIC ENGAGEMENT

Brackstone (1987) and United Nations (2011) advise statistical offices to demonstrate the cost savings and reductions in the burden on respondents to generate public support for using administrative data for statistical purposes. In providing feedback for administrative offices, it would be useful to explain the potential utility of statistical processes in evaluating the effectiveness of programmes; Prell et al. (2009) cite an example in which the combination of
survey and administrative data was applied to evaluate the effectiveness of Medicare programmes.

Explaining the measures used to ensure privacy can mitigate public concerns related to confidentiality. United Nations (2011) suggests: “… publication of clear limits and rules regarding the use of data, ensuring that people and businesses understand that sensitive data used or collected for statistical purposes will not be fed back to other parts of government (particularly tax and benefits agencies).” This is consistent with principle 6 of the United Nations *Fundamental Principles of Official Statistics*, which establishes the “one-way flow of data to the statistical office” stating: “… individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.”

**Example 7.1.6.** Brackstone (1987) describes how Statistics Canada improved public trust in the confidentiality of administrative data used for statistical purposes: “Public communications to respondents and users should continually stress the importance attached to confidentiality of all individual micro-data acquired by the statistical agency… The physical security that surrounds the use of sensitive administrative records should be clearly visible… Exemption of statistical files from examination by security or intelligence services is an important element in maintaining public trust in the absolute confidentiality of data provided to the statistical agency. An exemption for Statistics Canada data (the sole institutional exemption within government) was provided when the new Security and Intelligence Service was formed in 1983.”

### 7.2. COMBINING MULTIPLE DATA SOURCES

As discussed in Section 6, statistics and numerical summaries resulting from different data collection systems often differ substantially even if they are intended to measure related concepts. Use of a single administrative data source may lead to biased estimators as a result of under-coverage or reporting errors that may derive from incentives in the administrative process.

Combining data from disparate sources can help to compensate for errors in a single source, and using multiple data sources can improve coverage of a population or increase the available variables. Combining information from several sectors can improve the coherence of estimates in different subject areas: combining a business register with a farm register is an example. Synthesizing several data sources with different error properties can lead to estimators with better properties than an estimator based on a single source. Figure 5 illustrates a statistical system of data sources in which census data available every five years are augmented with administrative and survey data that are available annually.
Several of the uses of administrative data discussed in Section 4 support surveys or censuses. An administrative file may be used to update a list frame that is later used in combination with an area frame in a multiple-frame survey. External data may be used to reduce non-sampling errors through error checking or imputation of missing data, and administrative data may be used as controls in calibration or as auxiliary information in model-based estimation.

The implications of differences between administrative and survey data depend on the intended use of the administrative data source. If it is to be used to tabulate statistics directly, and the resulting statistics contradict parallel statistics based on a survey, the disagreement between the two sources can cause confusion. But if the administrative data source is to be used as an input in a model-based or model-assisted estimator, lack of consistency among the sources is not necessarily a problem. This is because if part of the model-fitting process involves estimation of bias associated with the administrative data, the process of constructing the final statistic accounts for the differences between the administrative and survey data.

In combining multiple data sources, it is often necessary to designate one source as the standard unbiased measurement. This source may still have errors such as sampling errors if it is a survey, but designation as the unbiased source means that the average of this source aligns with the target concept or quantity of interest.

It is important to understand the nature of each data source when determining data-reconciliation procedures. Understanding input data sources guides the choice of data source as the unbiased standard source. As discussed in Section 7.4, incorporating sufficient metadata is one way of ensuring that the properties of the administrative data source are transparent to the data user.
Example 7.1. Experiences at Statistics New Zealand exemplify the opportunities created by combining administrative data with other sources. Before the 1980s, Statistics New Zealand used tax data exclusively for direct tabulation of incomes. In the mid-1980s, it reconciled the business frame with the goods and services tax, a form of VAT with nearly complete coverage. The process involved creating a link between the business frame and the tax data, and maintenance of the link was the central facilitator of expanded use of tax data.

Tax data are now used in various ways to improve the statistical efficiency of estimators and the economic efficiency of the statistical production process. Goods and services tax data replace survey data for businesses that contribute less than 10 percent to 15 percent of the total in several cut-off surveys such as the monthly retail trade survey, the quarterly wholesale trade survey and the quarterly manufacturing survey. Goods and services tax data and employment counts from the employer monthly schedules are used in sample design and estimation for the annual enterprise survey. Tax data are used for validation by comparing the total profit recorded in surveys with total tax paid by businesses.

Source: Dion, Chartrand, and Murray (2010).
7.2.1. REGISTER SYSTEMS

Wallgren and Wallgren (2007 and 2010) discuss the value of combining administrative registers to form register systems, which consist of data from different sources linked together form a single dataset in which each record appears once. In developing register systems, high quality information on micro-data is essential: in particular, accurate identifying variables are critical for efficient linking processes. Register systems may consist of data from surveys as well as administrative sources. Section 9 discusses this topic further.

7.2.2. MICRO-INTEGRATION

Different registers and surveys often contain different “response variables.” When data are linked at the unit-level, the existence of several versions of related variables provides an opportunity for quality improvement and expansion. Creating a single complete dataset in which each record appears one is called “mass imputation”. Data are then imputed for all records in the resulting register system. Mass imputation involves complex modelling techniques, and computational challenges arise as a result of the enormous volume of data (Guigo, 2008).

Three challenges in micro-integration, discussed in Section 6, involve differences in units, identifying variables and definitions. Methods exist to cope with these difficulties: in particular “profiling,” probabilistic record linkage and model-based editing (see Section 7.4).

Example 7.2.1. Statistics Canada’s survey of employment payroll and hours provides monthly estimates of status and trends in 10,000 establishments. Statistics Canada also has access to the complete file of payroll deductions remittance forms from the customs and revenue agency. These administrative sources provide the number of employees and gross monthly payroll variables. Regression models predict survey variables from the administrative records. These models are used for mass imputation of the survey response variables for all units in the administrative file (Grondin and Lavallée, undated).

Statistics Canada’s goods and services tax data project replaced survey data for 50 percent of “simple establishments” with goods and services tax data. Because the definitions and reference period for the tax data do not align perfectly with the survey data, models are needed. Two models – a micro approach and a macro approach – were developed to predict survey response variables from the goods and services tax data. The model-based predictions were applied to the monthly restaurant, caterers and taverns survey and the monthly survey of manufacturing. The micro approach was judged superior, primarily because the data structure permitted continued use of existing software. Replacing the survey data with goods and services tax data for 50 percent of simple establishments reduced data collection costs and the burden on respondents (Pierre and Brodeur, undated).
7.2.3. MACRO-INTEGRATION

Differences between estimators do not necessarily prevent the combination of different sources to obtain final synthesized statistics. Synthesizing estimators with different error properties can reduce bias associated with each source. The eventual statistic based on the combination of the various estimators may have better statistical properties such as low bias and MSE than either of the individual estimators, even if the input estimators have different characteristics.

When estimators are combined, one source may have to be designated the unbiased measurement; this involves detailed understanding the properties of the various data sources. The unbiased source is often assumed to be a sample survey because statistical offices have control over data-collection procedures.

Given an unbiased source of information, statistical models can be developed to integrate estimators from administrative data sources with an unbiased sample survey estimator. The biases of the alternative estimators are parameters to be estimated in the model. If the model assumptions are reasonable approximations to the data, the model-based estimator that combines the alternative data sources is likely to have a smaller mean squared error than any of the input estimators alone.

Example 7.2.1 illustrates the use of a statistical model to combine an unbiased survey estimator with estimators from administrative data. The data are from two USDA agencies and the objective is to estimate areas planted with different crops. The survey estimator is from a NASS survey, and one of the administrative sources is from the Farm Services Agency.
The model-based approach of Nandram, Berg and Barboza (2013) provides a related way to address the problem of combining information from multiple surveys. It was not developed specifically for administrative data, but it is applicable to the problem of combining survey and administrative data: one survey is identified as the unbiased estimator, and the estimators from the other surveys are assumed to have bias and variance depending on a fixed number of model parameters. Inference for the bias and variance parameters uses hierarchical Bayesian methods.

7.2.4. MULTIPLE-FRAME SURVEYS AND CENSUSES

The quality of identifying variables may be insufficient to support the linking operations needed to form unified register systems. In this case, collections of distinct but possibly overlapping databases may be used as the basis of multiple-frame surveys and censuses. In multiple-frame surveys, independent samples are drawn from different data sources to improve coverage without complex reconciliation of different databases with inconsistent identification variables (Carfagna and Carfagna, 2010).

In agricultural surveys, using a list frame in conjunction with an area frame may be a way of obtaining complete coverage of the target population of farms. A list frame may contain detailed variables useful for stratification, but may not include all farms. An area frame may cover the complete population, but
maintaining detailed information for use in stratification may be too expensive. This approach is advantageous when a list frame contains large and probably more variable operations (Carfagna and Carfagna, 2010).

In constructing estimates from multiple-frame censuses or surveys, it is essential to account for possible overlaps between different data sources. Consider a dual-frame survey in which two samples are drawn independently from two different frames. Assume that the two frames, $A$ and $B$, are each incomplete but together cover the entire population. The frames produce three mutually exclusive and exhaustive sub-domains defined by the units in frame $A$ but not frame $B$, the units in frame $B$ but not frame $A$, and the units in the intersection of the two sets. If the three domain sizes are known, an estimate of the population total is:

$$Y = N_a \bar{y}_a + N_{ab} (p \bar{y}_a^A + q \bar{y}_a^B) + N_b \bar{y}_b,$$

where $N_d$ is the population total for domain $\bar{y}_d$ ($d = a, b, ab$), $\bar{y}_d$ is the estimate of the mean for domain $d$, $\bar{y}_a^K$ is the estimate of the mean for domain $ab$ based on the sample selected form frame $K$, and $p$ and $q$ are two non-negative numbers that sum to 1 (Carfagna and Carfagna, 2010).

**Example 7.2.2.** NASS uses an area frame and a list frame in the census of agriculture. The list frame, composed partly from administrative data from the Farm Services Agency, is thought to have incomplete coverage for small farms, so to prevent bias the list-frame census is supplemented with information from the June area survey, an area-frame based operation. Estimates of the number of farms based on the dual-frame procedure are thought to have smaller bias than estimates based on the list frame alone, and smaller variance than estimates based solely on the June area survey (USDA, 2014a).

**7.2.5. USE OF SAMPLE SURVEYS TO IDENTIFY AND CORRECT ERRORS**

Comparisons of aggregate estimates based on an administrative source with comparable estimates obtained from a survey can validate an administrative source or indicate possible biases. Sen (undated) discusses the use of surveys in India to check and correct administrative data: “…in the case of civil registration, India operates a sample registration system which provides reasonable estimates of demographic indicators for the inter-census period.”. He also advocates an approach that would: “involve the use of limited surveys based on strong statistical principles to provide validation and corrective factors for the data generated on a regular basis through administrative accounts. In an important sense, this would be akin to implementing a sample audit system, where perhaps the purpose would not be to find fault but to provide information.
which would be used to correct the inherent biases that may occur in administrative reporting.”

Many ways of combining data sources with conflicting properties have been explored by statistical offices in developed countries. One approach involves manual review of discrepancies to correct errors in administrative and survey data sources that measure the same concept, as in example 7.2.3.

Example 7.2.3. In the production of the national resources inventory (see Section 1), information on large bodies of water is obtained from administrative data sources, primarily the national hydrography database. Each large water area is represented as a unique geographic information system polygon. If a sampled segment contains a large body of water, the data collector determines its area; if a secondary sampling unit falls in the body of water, the data collector designates the water feature as a lake, reservoir, estuary, large stream or bay.

In many cases the information collected in the national resources inventory conflicts with the administrative information, partly because the data collection protocols for the survey and the administrative database are different. A common type of conflict – when a geographical information system polygon classified as an estuary contains a point classified as large stream – results from differences between definitions used in the survey and administrative processes. In this situation, experts will determine that the estuary is the correct classification.

Conflicts between geographical information system and national resources inventory data are often resolved through manual review. An automated procedure is used to create a database of conflicts between the different data, and data collectors manually review the conflicting records and correct one or other set of data as appropriate. This is time consuming and costly, but it improves the quality of final statistical products. Manual review is preferred to an automated approach because of the complexity of the data items that require editing.

7.2.6. DATA COMBINATION IN REPORTING SYSTEMS AND EXPERT ASSESSMENTS

One limitation of information from subjective assessment of qualified informants is the potential for biased reporting or lack of complete representation. Combining information from several experts can minimize the effects of biased reporting, and incorporating the views of reporters from different areas of the agricultural sector, for example, will provide a more representative set of opinions.

Galmes (2013) gives guidelines for minimizing biases in reports of expert assessments. The first two are: i) select more than one qualified informant for each geographical domain; and ii) as far as possible, select informants with conflicting interests. Examples of such informants could be a farmer or farmers’ group and the agricultural bank official responsible for approval of credit to
farmers, or an agricultural extension officer and the buyers of agricultural products.

Keita and Chin (2013) note the need for reconciliation when there are several conflicting assessments: “… the differences must be explained, a judgment made, and the synthesis of the figure carried out. The final synthesis is essential and should be carried out by the statistical office."

7.3. AUDITS AND EMBEDDED SAMPLES

Audits and embedded sample surveys are recommended to correct for errors in administrative data. Carfagna and Carfagna (2010) explain that in the IACS database: “… sample surveys must be performed or the statistical system will produce biased results.” A survey or audit that flags unusual records may be conducted to identify erroneous cases. A slightly different design is needed if the objective is to estimate the extent of error in an administrative source.

Example 7.3.1. Carfagna and Carfagna (2010) provide two examples related to the IACS database that illustrate the differences between correcting erroneous cases and estimating error properties. Quality control is carried out on the IACS database every year: because the goal is to “detect irregularities” rather than estimate the error without bias, the audit sample is not a probability sample. For major crops such as durum wheat, the error detected through the audit is 3.5 percent of the area checked. For minor crops such as legumes, the relative error is of the order of 16 percent of the area checked. In a study in the Italian regions of Puglia and Sicily to estimate the error in IACS-based estimates for the area in durum wheat, an area frame based sample of segments was selected and the area of durum wheat in each sampled segment recorded. The sampled segments were then linked to the IACS database and the declared area in durum wheat was computed. Assuming that the survey led to an unbiased estimate, the comparison of the IACS values with the areas measured in the survey suggest a relative error of omission of 13.9 percent in Puglia and 23.3 percent in Sicily.

7.4. SYSTEMS AND BEST PRACTICES FOR QUALITY ASSURANCE AND CONTROL

Standard best practices for quality control and assurance also apply to administrative data; standardization of variables, for example is valuable for data integration and interpretation. This section reviews mechanisms for quality management related to data management, collection and dissemination (see also: Dion, Chartrand and Murray (2010); Daas et al. (2012b); Laitila, Wallgren and Wallgren (2011); Karr, Sanil and Banks (2006); Biemer and Lyberg (2003); Biemer et al. (2014); Lyberg and Biemer (2008)).

Establishing conceptual and technical infrastructures for maintaining administrative data can help to ensure data quality. Technical Report 1 proposes the Administrative Data System for Agricultural Statistics for maintaining and
verifying the quality of administrative data related to the core agricultural data items in the FAO Global Strategy for Improving Agricultural Statistics. An alternative is to borrow from the Health Matrix Network, a tool designed to assist developing countries in planning health information systems (African Development Bank et al., 2007). The reader is referred to Technical Report 1 for further detail. Biemer et al. (2014) describe a framework for quality evaluation called A System for Product Improvement, Review and Evaluation, which “provides a comprehensive framework for systematically evaluating all dimensions of quality with the primary focus on accuracy.” The concept, which covers a range of quality dimensions, was developed to evaluate the collection of statistical products generated by national statistical offices.

Detailed monitoring of the collection of administrative data can help to improve their quality. Galmes (2013) provides guidelines for improving the quality of collected data. Though stated in the context of expert assessment, the guidelines apply to data collection and include the following advice: i) use a standard format for collecting information; ii) prepare manuals containing clear definitions of activities; iii) train data collectors periodically; and iv) supervise all operations. Implementing such quality controls in data collection can improve the transparency of data-collection procedures and help statistical analysts and consumers of statistical products to use the data appropriately.

To apply quality controls it is important to have sufficient information about administrative data. Information about collection processes can indicate the reliability and accuracy of the resulting administrative data. Accurate definitions of administrative data sources are essential for understanding the data items and utilizing them appropriately. When transferring administrative data between administrative and statistical agencies, administrative data offices should provide sufficient information to statistical agencies to ensure appropriate use and interpretation of the data (Dion, Chartrand and Murray, 2010).

In their approach to evaluating administrative data, Daas et al. (2012b) define quality as “statistical usability”, which is different from usability as a final product. They found that Statistics Netherlands had a well developed approach to evaluating metadata, but that there were no corresponding guidelines to evaluate “the quality of the data in administrative sources” (Dion, Chartrand and Murray, 2010). The approach to routine evaluation of administrative data for statistical purposes developed indicators for five quality dimensions (see Example 7.4.2). The tools recommended for evaluating the quality dimensions include graphical techniques, which are useful for visualizing relationships between data sources and identifying errors.
Daas and Fonville (2007) emphasize the value of metadata in evaluating and improving an administrative register. They recommend that before an administrative data source is used the metadata are evaluated in a quality checklist in which scores determine whether a source can be used and the extent to which it can be used. They advocate a sequential approach: i) review the metadata; ii) evaluate coverage and overall reliability; and iii) evaluate the data in terms of timeliness and continuity.

Statistical offices apply established best practices to verify the quality of incoming administrative data. The 2013 special issue of the Journal of Official Statistics was devoted to “Systems of Architectures for High-Quality Statistics Production.” Concepts from quality-control and assurance frameworks developed for survey data apply to administrative data, and frameworks for evaluating and maintaining the quality of data used and produced by statistical offices contain concepts valuable in the Administrative Data System for Agricultural Statistics, as outlined in Technical Report 1. Examples 7.4.3–7.4.5 illustrate best practices for checking and improving the quality of administrative data.
Example 7.4.1. Before using administrative data to evaluate survey data and construct estimators, NASS evaluates and documents the reliability, coverage and definitions of the external data source. The analysis of administrative data has four dimensions:

i) Universe. The universe represented by the administrative source is defined and evaluated, and differences between the administrative universe and target population defined by NASS are identified and documented. Under-coverage or over-coverage of the source, possible duplication of records and problems related to the reference period are evaluated.

ii) Data-collection methods. This involves questions such as: Are the data supplied voluntarily or by law? Are data collected using standard forms with defined questions and terms? Are non-respondents and refusals pursued in follow-up investigations?

iii) Data validation and summary. This involves a review of the way data are handled after collection, including checking for and editing errors, imputing missing data and identifying unusual values.

iv) Data quality and consistency. The quality of the administrative source is assessed by examination of indicators such as non-response rates, fluctuations of reporting periods and quality-control processes used by the administrative agency. (USDA, 2011)

Example 7.4.2. Daas et al. (2012b) propose five quality dimensions for administrative data:

i) “Technical checks” determine how easily the administrative data file and the data in the file can be used, using indicators such as how easily the file can be accessed and read into subsequent programmes, the clarity of metadata and the ease with which the file can be converted into standard formats used by the statistical office.

ii) “Accuracy of objects” measures the degree to which the data are correct, reliable and certified. Indicators of accuracy include internal inconsistencies and outliers.

iii) “Completeness” means the extent to which the administrative data file encompasses the collection of population units and conceptual variables of interest to the statistical agency. Issues related to completeness include under-coverage, over-coverage and alignment between the population of the administrative process and the population of interest for statistical purposes.

iv) “Time-related dimensions” refer to the temporal frequency and reference period of the administrative source. Time dimensions include timeliness, punctuality overall time lag and delay. Timeliness is defined as the length of time between the end of the reference period of the source and the moment of receipt by the statistical office. Punctuality is the time between the date when the administrative agency agrees to deliver the data set and the actual delivery date. The overall time lag is the difference between the end of the reference period and the date at which the administrative agency is prepared to produce statistics based on the source. Delay refers to the time between an actual change in the physical population of interest and the date when the change is captured in the administrative file.

v) “Integrability” is the ease with which the administrative source can be combined with other datasets maintained by the statistical office. Related aspects include alignment of definitions of units and variables, the quality of identifying variables used for linking operations and the closeness of values of variables in alternative sources of data.
As discussed in Section 6, the operation of data integration presents computational and methodological challenges. Software has been developed to address such issues, and methods of profiling and probabilistic record linkage have been developed to handle lack of standardization in definitions of units or identifying variables.

### 7.5. METHODOLOGICAL AND TECHNICAL TOOLS FOR DATA INTEGRATION

As discussed in Section 6, the operation of data integration presents computational and methodological challenges. Software has been developed to address such issues, and methods of profiling and probabilistic record linkage have been developed to handle lack of standardization in definitions of units or identifying variables.

#### 7.5.1. PROFILING

One challenge associated with integrating multiple sources of information is that different data sources have different definitions of units. United Nations (2011) explains: “… converting administrative units to statistical units can be quite difficult conceptually and often involves some form of modelling.” The term “profiling” is used in business surveys to describe this process, but the concept applies in other contexts (United Nations, 2011). Chapter 15 of the United Kingdom Business Register Recommendations Manual defines profiling as: “… a method to analyse the legal, operational and accounting structure of an enterprise group at national and world level, in order to establish the statistical units within that group, their links, and the most efficient structures for the collection of statistical data” (United Nations, 2011).
Profiling may be manual or automated (United Nations, 2011). Standard rules based on attributes or the nature of links between units can help to overcome differences between administrative and statistical units. Statistical households, for example, can be derived on the basis of relationships among the individuals living in a building; this approach is a component of the register-based population census method used in Nordic countries. Even with clerical profiling the disaggregation of units may require subjective determinations, and a single correct solution may not exist. In automated processes, which are cheaper and faster than clerical profiling, standard rules regarding the nature of links are applied.

An alternative to rule-based profiling involves the specification of statistical models. Relationships between administrative and statistical units may be established for a subset of a population, for example through a survey, and parameters of models describing relationships are then estimated and applied to the full population. United Nations (2011) provides a simplified example in which the administrative unit is a “job” and the statistical unit is a “person”. In an estimate based on a survey, on average each person has 1.15 jobs; this estimate can be used as a global adjustment factor to determine estimates of employment from the number of jobs. The variability in the survey-based estimate of the ratio would have to be incorporated in subsequent employment analyses.

### 7.5.2. PROBABILISTIC RECORD LINKAGE

Probabilistic record linkage is a model-based procedure that is often used to merge two files or to remove duplicates from a single file and can be used for determining whether sets of data define the same underlying concept. Fellegi and Sunter (1969) establish a statistical framework for probabilistic record linkage. Probabilistic record linkage and related computational tools are useful for applications involving administrative data. For detailed discussion, see Section 9.

Many authors extend the Fellegi and Sunter (1969) procedure. Larsen and Rubin (2001), for example, develop a procedure that involves alternating between inference based on conditional distributions and manual review; incorporating a clerical review reduces the degree of uncertainty. When datasets are linked through probabilistic linkage models or from incomplete linkage, the estimation procedures need to account for linkage error. Methods such as those proposed by Kim and Chambers (2012) can be employed, which extend the regression analysis techniques of Chambers (2009) to applications with more than two linked datasets. Berka et al. (2012) examine the effectiveness of the
Dempster-Shafer theory to quantify uncertainty in each datum in registers used for the register-based Austrian census.

7.5.3. AUTOMATED AND MODEL-BASED EDITING

When an administrative database is delivered to a statistical office, it is not automatically ready for statistical use in that records may be incomplete or contain errors. Calculations may be needed, for example to convert the variables available in the administrative file to variables of interest in a survey, but because administrative files are often large, manual editing and imputation are impracticable. Automated methods, which may be based on explicit or implicit statistical models, will have to be applied to process administrative files and convert them into a format that can be used for statistical purposes.

Adaptations of solutions to the problems of lack of standardization of units and identifying variables can be applied to problems arising from differences in definitions and classification systems. In cases involving differences in classifications, conversion matrixes are constructed to assign codes relevant for statistical purposes to codes in the administrative database (United Nations, 2011). If the correct assignment of codes is uncertain, probabilistic selections may be determined from weights (United Nations, 2011), which may be obtained from a combination of subjective and empirical sources of information. Figure 7 is an example of a conversion matrix modified from United Nations (2011): the pairs with a weight of 100 correspond to certain matches; pairs with weights less than 100 reflect a one-to-many relationship. The code is selected with probability proportional to weight.

Table 2: Example of comparison matrix, adapted from United Nations (2011)

<table>
<thead>
<tr>
<th>Code 1</th>
<th>Code 2</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>C</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>70</td>
</tr>
</tbody>
</table>
Example 7.5.1. An example of the role of automated editing and imputation procedures in converting a complex administrative data file into a format usable for the production of official statistics is provided by the Italian statistical office. Each month, businesses submit a DM10 form with declarations of social security contributions to the social security office, which transmits the data to the statistical office 45 days after the end of the reference quarter.

Baldi et al., 2003 note that before statistical analysis “… the DM10 form passes through a preliminary check procedure to verify the validity of each code and of the relative information recorded. In particular, errors in the ID of the units, duplications and formal coherence with the current legislation are checked and, where possible, corrected.” A preliminary check identifies incoherence between number of employees, wage bills and social contribution bills; a more complex selective editing procedure is then applied to the survey data. This process consists of a combination of manual and automated steps: it first selects from among 500,000 units the sub-sample to be examined on the basis of estimated probability of an error determined from functional relationships between variables and over time for the same variable. The number of edits is generally small enough to indicate that the microdata are of high quality.

A form arriving at the social security office 14 months or more after the target date is considered a non-respondent. To complete the administrative data file, variables for non-respondents are imputed using models based on analyses of activity patterns in the DM10 forms in quarters neighbouring the quarter to be imputed. Approximately 12% of the units that do not submit a DM10 form are subject to the imputation process.

Preliminary and final estimates are subject to manual review. The manual inspections examine simple measures such as comparisons over time and checking whether the new observation is an overall maximum or minimum. Time constraints often limit the extent of the manual review, and automated error-detection algorithms based on time series are also applied (Baldi et al., 2003).

Example 7.5.2. The use of goods and services tax data at Statistics Canada demonstrates the value of automated editing in improving the usability of administrative data for statistical purposes. A project at Statistics Canada aimed to replace survey data for a subset of business establishments in sub-annual surveys with goods and services tax information. The data were not immediately ready for statistical use because they contained errors and missing values. But because of the tax data collection schedule and the time required by the tax agency to process the data, the goods and services tax data for the month prior to the reference month of the survey were available for statistical estimation. The frequency – monthly, quarterly or annual – with which businesses submit these data depends on the size of the business. The aim of processing the data was to transform the database submitted by the tax agency to a format with a consistent monthly time series for each unit.

The processing involved several steps: one component involves identifying bounds for outlier detection and subsequently identifying outliers; a second set of operations determines which business transactions require imputation, selects an imputation model from a class of possible models and performs the imputation, and a third component called “calendarization” converts the time series to uniform monthly intervals. The calendarization benchmarks the goods and services tax data to a published monthly indicator series (Pierre and Brodeur, undated).

Example 7.5.3. Reale et al. (2013) used administrative data to detect outliers in the Italian agricultural census. A forward search algorithm was developed to identify large differences between census and integrated agricultural control system data. Identified outliers were reviewed manually. The search was based on selected variables such as utilized area, total area, vineyards and olive plantations. The outlier identification operation not only reduced errors but also narrowed gaps between provisional and final figures.
Probabilistic Record Linkage

Many applications of administrative data to the production of official statistics involve multiple sources of information such as administrative and survey files. In many cases, it is desirable to link records from at least two files at the level of the individual unit in a population. Consider, for example, the use of administrative data to check for errors in survey data. A comparison of marginal distributions of the administrative file to corresponding marginal distributions from the survey may be informative, but a comparison of the alternative data sources at the unit level opens greater possibilities: a unit-level linking operation, for example, enables the evaluation of records with relatively large differences in the values recorded in the two different sources.

Merging the files at the record level presents many challenges. The identifying variables may differ across datasets, and even if a unique identifier exists the identification variable may be missing or incorrectly recorded for some units. Duplicate records may exist in one or more files, and large datasets may demand substantial computational effort. Probabilistic record linkage is a statistical procedure for determining the probability that two sets of identifying variables represent the same unit in the population.

Fellegi and Sunter (1969) developed a widely used probabilistic record linkage procedure, in which the latent match status of interest is represented as a latent binary variable, $\delta$, that is 1 if a given pair is a match and is 0 otherwise. The observations are vectors of comparison variables, $Y = (y_1, \ldots, y_K)$, where $y_j = 1$ if variable $x_j^A$ from data set A is equal to variable $x_j^B$ from data set B, and is zero otherwise. The distribution used for inference is the conditional distribution of $\delta$ given the observed vectors of comparison variables.

Many extensions to the Fellegi and Sunter (1969) procedure have been developed. Larsen and Rubin (2001), for example, developed a procedure that involves alternating between inference based on conditional distributions and manual review. Incorporating clerical review reduces the degree of uncertainty. When datasets are linked through probabilistic linkage models or from
incomplete linkage, the estimation procedures must account for linkage error. Methods such as those proposed by Kim and Chambers (2012) can be employed: these extend the regression analysis techniques of Chambers (2009) to applications with more than two linked data sets. Berka et al. (2012) examine the effectiveness of Dempster-Shafer theory to quantify uncertainty in each datum in each of several registers used for the register-based Austrian census.

8.1. STEPS INVOLVED IN DATA CLEANING AND RECORD LINKAGE

Preparing a dataset for record linkage and performing the linkage algorithm involve several related components. One involves selection of variables to use for matching, and then, given a set of matching variables, an operation is often needed to convert representations in different data sets to a standard form. Because comparison of all pairs of records in two files is often computationally prohibitive, methods are needed to reduce the dimension of the comparison space. In comparing two vectors of identifying variables, strict equality is often too restrictive a metric, but record linkage algorithms permit the use of different comparison metrics that determine the extent to which two vectors of matching variables agree. The Fellegi-Sunter probabilistic record linkage paradigm contains several options that users can adapt to their needs: these involve estimating matching probabilities and the decision rule that determines which records to review manually. ESSnet (2008c) and Day (1994) discuss methods for several aspects of record linkage.

The selection of matching variables often involves manual and automated steps. Analysts often have prior knowledge about logical matching variables: two examples of useful variables for matching the files of individuals are social security numbers in the United States and personal identification numbers in the Netherlands. Automated algorithms known as profiling procedures have been developed to assist in the selection of variables: these measure the correlation between variables across datasets as well as the quality of the potential matching variables within a dataset. Profiling procedures are particularly helpful in applications with large numbers of potential matching variables with varying degrees of reliability.

Before linking records, the formats of the matching variables must be standardized. Two ways to store the date December 8 1952, for example, are “12-8-1952” and “12/8/1952.” A standardization algorithm would convert these into a format that would enable a matching algorithm to recognize them as equivalent. In practice, standardization algorithms have to operate on more complex character strings that represent attributes such as telephone numbers,
addresses, names of people, names of businesses and names of farming operations. Character strings that represent unique entities may differ as a result of subtle differences in capitalization, spelling and punctuation for example. A way of standardizing character strings is to use a phonetic coding scheme; such algorithms convert strings that “sound the same” into a unique character format. Soundex and NYSIIS are two examples.

Comparing all pairs of records in two files is often computationally prohibitive. Two methods for reducing the number of comparisons are called “blocking” and “sorted neighbourhood”. In “blocking,” the files are divided into subgroups called blocks, and pairs of records are only compared within each block. To illustrate this, suppose that two files each with 5,000 records are split into 10 blocks of 500 records: the number of comparisons is reduced from 5,000×5,000 in the unblocked structure to 10(500×500) in the blocked structure – ten times fewer. In the sorted neighbourhood dimension reduction procedure, records are only compared if they fall in a window that traverses the sorted records.

Because exact matching is often too restrictive, different metrics have been developed to measure the extent to which two vectors of matching variables differ. In the record linkage literature, different metrics are often referred to as “comparison functions”. Alternatives to strict equality include the Levenshtein metric for comparing two strings and the Jaro-Winkler metric, which is specific to comparing names.

In probabilistic record linkage, users may choose between different methods of parameter estimation and decision rules. Match probabilities, for example, may be estimated by maximizing a likelihood, which is often based on an assumption of conditional independence, and implemented with the expectation-maximization (EM) algorithm. Alternatives to maximum likelihood include frequency-based matching or algorithmic procedures that evaluate patterns of agreement and disagreement. Record-linkage algorithms often result in decisions about which records to review manually. Fellegi and Sunter (1969) prove that a particular decision rule minimizes the number of pairs to review for a given error rate. An alternative to the Fellegi-Sunter (1969) decision procedure is a threshold-based rule, where any pair with a probability in a specified range is reviewed.

8.2. SOFTWARE FOR MICRO-INTEGRATION AND RECORD LINKAGE

The challenges of combining disparate data sources are conceptual and computational. Record linkage often involves managing large quantities of data, and algorithms for data cleaning and standardization are needed. A variety of
software packages have been developed to clean data and combine multiple data sources. This section first reviews the technical capabilities of existing software for record linkage discussed in Section 8.1 and then compares them in terms of cost, extendibility and transparency. A summary of our interpretation of the primary strengths and weaknesses of administrative data and a recommendation for developing countries conclude the section.

The discussion below is primarily based on ESSnet (2008c) and includes ideas from Day (1994), Sariyar and Borg (2010) and da Silva et al. (2011). ESSnet (2008c) reviews several record linkage software packages from the standpoint of producing official statistics of business data. Day (1994) reviews record linkage software with a view to determining the most appropriate tool for the National Agricultural Statistics Service. Day (1994) contains a useful list of questions and criteria for analysts to consider when selecting a record linkage tool. The computational tools discussed in Day (1994) may be outdated, but the suggested criteria and questions remain relevant. Da Silva et al. (2010) reviews probabilistic record linkage software with a view to integrating data from the Brazil census with data from a post-enumeration survey.

The first nine packages are reviewed in ESSnet (2008c). The last is an R package discussed in Borg and Sariyar (2010). The following software packages are considered:

1. AutoMatch, developed at the United States Bureau of Census, now under the purview of IBM (Herzog et al. 2007).
2. Febrl – Freely Extensible Biomedical Record Linkage – developed at the Australian National University.
4. LinkageWiz, commercial software.
5. RELAIS, developed at the Italian statistics office.
6. DataFlux, commercialized by SAS.
7. The Link King, commercial software.
8. Trillium, commercial software.
9. Link Plus, developed at the United States Centre for Disease Control and Prevention Cancer Division.
10. RecordLink, an R package developed by Murat Sariyar and Andreas Borg.

The evaluation relies on our interpretation of the review in ESSnet (2008c) rather than experience of using the software packages. The reader is advised to view the conclusions and interpretations of software capability as a guide. In deciding which tools to employ for a particular application, an independent
evaluation of software may be useful. The overview below is intended as a starting point.

8.2.1. TECHNICAL CAPACITY

Table 3 summarizes the capabilities of alternative software packages; they are reviewed in detail in ESSnet (2008c). The “standardization” column indicates pre-processing and standardization capabilities; the “profiling” column indicates whether the software has options for automated profiling; the “space reduction” column indicates the blocking methods available; the “estimation and decision rules” column provides information on the procedure used to estimate matching probabilities and the type of rule used to decide if a pair of records is classified as a “match”, a “non-match” or a “possible match.” Because all software packages contain comparison functions, the column “comparison functions” indicates the extent of the available functions based on our understanding of the information provided in ESSnet (2008c).

8.2.2. USABILITY

Following ESSnet (2008c), the following indicators of usability are considered (see Table 4):

- Cost. Is the software free or commercial? Does it require licenses for particular data management or statistical analysis tools?
- Domain specificity. Can the tool handle different languages, or is it specific to English? Is the tool developed for a specific class of applications or objects such as business data, human subjects or health services?
- Transparency. Are the procedures well documented? Can the analyst build an understanding of how the record linkage and data management tools work?
- Extendibility. Can the analyst modify and adapt the procedures to suit specific needs?
- Output Reports. Is the output in a convenient format? Are linked files easy to use and transport to a different system?
<table>
<thead>
<tr>
<th>Package</th>
<th>Standardization</th>
<th>Profiling</th>
<th>Space reduction</th>
<th>Estimation and decision rules</th>
<th>Comparison functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>AutoMatch</td>
<td>NYSIIS, Soundex, other</td>
<td>None</td>
<td>Blocking</td>
<td>Frequency weighting with threshold</td>
<td>Standard</td>
</tr>
<tr>
<td>Febrl</td>
<td>Hidden Markov models and rule-based methods</td>
<td>None</td>
<td>Blocking and sorted neighbourhood</td>
<td>Several unsupervised classifiers</td>
<td>Wide variety</td>
</tr>
<tr>
<td>GRLS</td>
<td>NYSIIS, Soundex, other</td>
<td>None</td>
<td>Blocking</td>
<td>Agreement/disagreement patterns</td>
<td>Standard</td>
</tr>
<tr>
<td>LinkageWiz</td>
<td>NYSIIS, Soundex</td>
<td>None</td>
<td>Not Specified</td>
<td>Few details on estimation and decision method</td>
<td>Standard</td>
</tr>
<tr>
<td>DataFlux</td>
<td>Tools for business data</td>
<td>Yes</td>
<td>Not Specified</td>
<td>Simple, deterministic decision</td>
<td>Wide variety</td>
</tr>
<tr>
<td>RELAIS</td>
<td>None</td>
<td>Yes</td>
<td>Blocking and sorted neighbourhood</td>
<td>Maximum likelihood estimation, manual review of many-to-many links</td>
<td>Standard</td>
</tr>
<tr>
<td>The Link King</td>
<td>None</td>
<td>Yes</td>
<td>Blocking</td>
<td>Ad hoc iterative procedure for estimation, and probabilistic and deterministic decision rules</td>
<td>Wide variety</td>
</tr>
<tr>
<td>Trillium</td>
<td>Extensive</td>
<td>Yes</td>
<td>Not specified</td>
<td>Probabilistic, not Fellegi-Sunter, procedure not specified</td>
<td>Standard</td>
</tr>
<tr>
<td>Link Plus</td>
<td>None</td>
<td>None</td>
<td>Blocking</td>
<td>Maximum likelihood, and probabilistic decision</td>
<td>Wide variety</td>
</tr>
<tr>
<td>RecordLink</td>
<td>None</td>
<td>None</td>
<td>Blocking</td>
<td>Maximum likelihood with EM algorithm and numerous classifiers for decision rules</td>
<td>Standard</td>
</tr>
</tbody>
</table>
Table 4: Summary of Usability of Alternative Software Packages

<table>
<thead>
<tr>
<th>Package</th>
<th>Cost and requirements</th>
<th>Domain specificity</th>
<th>Transparency</th>
<th>Adoption</th>
</tr>
</thead>
<tbody>
<tr>
<td>AutoMatch</td>
<td>Commercial</td>
<td>English only</td>
<td>Rich documentation</td>
<td>High</td>
</tr>
<tr>
<td>Febrl</td>
<td>Free</td>
<td>English only</td>
<td>Source code available</td>
<td>Medium</td>
</tr>
<tr>
<td>GRLS</td>
<td>Requires ORACLE</td>
<td>English only</td>
<td>Free, bilingual training course</td>
<td>Medium</td>
</tr>
<tr>
<td>LinkageWiz</td>
<td>Commercial but low price</td>
<td>English and French</td>
<td>No precise description</td>
<td>Medium</td>
</tr>
<tr>
<td>RELAIS</td>
<td>Open source, free</td>
<td>No specific domain</td>
<td>Full availability of source code</td>
<td>Low</td>
</tr>
<tr>
<td>DataFlux</td>
<td>Requires SAS, but low cost</td>
<td>No specific domain</td>
<td>Documentation available</td>
<td>High</td>
</tr>
<tr>
<td>LinkKing</td>
<td>Free</td>
<td>Health and human subjects</td>
<td>Well documented</td>
<td>Medium</td>
</tr>
<tr>
<td>Trillium</td>
<td>Commercial</td>
<td>Almost any language or country, but specific to marketing applications</td>
<td>Algorithms not precisely defined</td>
<td>Medium</td>
</tr>
<tr>
<td>LinkPlus</td>
<td>Free</td>
<td>Cancer registries</td>
<td>No source code but good documentation</td>
<td>High</td>
</tr>
<tr>
<td>RecordLink</td>
<td>Free</td>
<td>English and German</td>
<td>Source code and documentation available</td>
<td>Unknown, relatively new R package</td>
</tr>
</tbody>
</table>

8.2.3. SUMMARY

This section summarizes our understanding of the primary strengths and weaknesses of the software discussed above and provides guidance for developing countries. Table 5 summarizes the strengths and weakness of the software tools. The table is based on ESSnet (2008c) and our interpretation of the evaluations provided there.
Table 5: Primary Strengths and Weaknesses of Record Linkage Software Packages.

<table>
<thead>
<tr>
<th>Package</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>AutoMatch</td>
<td>User-friendly pre-processing</td>
<td>No error rate estimate, English only</td>
</tr>
<tr>
<td>Febrl</td>
<td>Open source, good pre-processing and tools to compare solutions from different record linkage algorithms</td>
<td>No profiling, requires ORACLE</td>
</tr>
<tr>
<td>GRLS</td>
<td>Good pre-processing and documentation</td>
<td>Only English</td>
</tr>
<tr>
<td>LinkageWiz</td>
<td>Speed, pre-processing standardization</td>
<td>No profiling or space reduction; black box</td>
</tr>
<tr>
<td>RELAIS</td>
<td>Allows user-specified combinations of linkage options; adaptable to a range of situations</td>
<td>Low adoption (new); relatively untested</td>
</tr>
<tr>
<td>DataFlux</td>
<td>Flexible pre-processing</td>
<td>Deterministic decision</td>
</tr>
<tr>
<td>LinkKing</td>
<td>Easy to use</td>
<td>Not flexible, non-standard estimation</td>
</tr>
<tr>
<td>Trillium</td>
<td>User-friendly and language flexibility</td>
<td>Specific to commercial applications rather than official statistics, limited documentation</td>
</tr>
<tr>
<td>LinkPlus</td>
<td>User-friendly and free</td>
<td>No pre-processing, specific to cancer registries, poor handling of non-showing characters in input data</td>
</tr>
<tr>
<td>RecordLink</td>
<td>Free, open source, numerous decision procedures, good documentation</td>
<td>No pre-processing, requires standardized input data in format compatible with package</td>
</tr>
</tbody>
</table>

In developing countries, a package with characteristics of RELAIS may be particularly useful; the software is free and the source code is completely available. The software allows the user to fix preferred combinations of linkage options and it is not specific to a particular subject domain.
Registers

Formal administrative processes often result in comprehensive listings of all participating units. If landowners are required to register acquisitions, for example, the resulting cadastral register contains a complete list of all owned land. Such registers are specific to a particular administrative process and the resulting register may hence only cover a portion of the population of interest to the statistical office, so combining registers and updating administrative lists with additional data from surveys, censuses or other administrative sources is essential when using administrative registers for statistical purposes.

Wallgren and Wallgren (2007) define registers and provide guidelines for creating the various components of a register, they discuss estimation techniques appropriate for register-based surveys and methods for imputing missing data and they provide guidelines for ensuring the quality of the register system, with discussion of the need for metadata and of the need to preserve the confidentiality of the data in the register; they also summarize many of these issues in the context of agricultural statistics in Sweden. This section considers the distinction between statistical registers and administrative registers, discusses the value of combining registers to form register systems and concludes with a case study of the statistical farm register in the European Union.

9.1. STATISTICAL AND ADMINISTRATIVE REGISTERS

Wallgren and Wallgren (2007) define three concepts:

- Register. A register aims to be a complete list of the objects in a specific group or population, but data on some objects can be missing as a result of quality deficiencies. Data on an object’s identity should be available so that the register can be updated and expanded with new variable values for each object. Complete listing and known identities are thus the important characteristics of a register.
• Administrative register. An administrative register aims to store records on all objects to be administered, and the administrative process requires that it is possible to identify all objects.
• Statistical register. A statistical register is based on data from administrative registers that have been processed to suit statistical purposes.

9.2. COMBINING ADMINISTRATIVE REGISTERS

Converting a single administrative register into a statistical register requires ensuring that the data in the administrative source cover the population of interest. This often requires that multiple administrative registers be combined to form a system of registers. Wallgren and Wallgren (2010) identify seven components of a register system:

• Base registers – statistical registers of objects or statistical units of fundamental importance for the system.
• Other statistical registers with statistical variables.
• Linkages between the objects or statistical units in different base registers and between base registers and statistical registers.
• Standardized variables – classifications of fundamental importance for the system.
• Metadata – definitions of objects, object sets and statistical variables and information about quality and comparability over time should be easily accessible.
• Register-statistical methods, including routines for quality assurance.
• Routines for the protection of integrity.

They argue that the use of register systems as opposed to a single source “… creates new opportunities to improve quality factors such as coverage, consistency, and coherence” and explain that incorporating tax information, monthly payroll reports and yearly income declarations can improve the business register. They advocate use of administrative information on crop subsidies and cattle inventories to improve the agricultural component of the Swedish business register.
9.3. STATISTICAL FARM REGISTER IN THE EU ACCEDING COUNTRIES

Turtoi, Akyildirim and Petkov (2012) discuss a conceptual approach to developing a statistical farm register with a view to producing coherent agricultural statistics. The primary purposes include provision of reliable sampling frames, a basis for inter-censal updates and more efficient use of existing administrative data.

Several administrative sources would ideally contribute to a statistical farm register. Information on crop areas may be obtained from farm subsidies; regulatory processes may provide information on animal health; land ownership information may be acquired from cadastral records. Registers specific to agriculture may be connected to registers containing information from other sectors to ensure consistency, for example between business and agricultural statistics.

One challenge in creating a harmonized farm register is that the basic unit varies across administrative sources. A unit in the farm register arising from subsidies is based on the holder of the land because “… each individual holder has to register himself/herself by bringing land ownership documents from the cadastral office.” The bovine register, maintained for veterinary purposes, relies on animal traceability and is therefore animal-based; linking the farm register with the bovine register would require linking each animal with the identity of the animal’s owner or keeper.

To remain current, the statistical farm register would have to be updated once or twice per year. Transactions captured in the contributing administrative registers must to be transmitted to the statistical farm register so that estimates
resulting from the final register are up to date. Figure 8, taken from Turtoi, Akyildirim and Petkov (2012), illustrates the complexity of the updating process: in addition to the farm and bovine registers, the information from surveys, census and the business register would inform the statistical farm register.

Figure 8: Structure of the conceptual statistical farm register for the European Union.

Source: Turtoi, Akyildirim and Petkov (2012)
Summary, New Directions and Lessons for Developing Countries

With increasing computing capacity and government budget cuts, statistical offices face mounting pressure to utilize non-statistical sources of data. Sources of administrative data used for agricultural statistics in developed countries include information from taxation records, regulatory processes, animal-tracing systems and land-ownership records. Uses range from direct tabulation to frame improvement. The quality of administrative data varies depending on the degree of formalization in the data-collection process and coverage of the administrative programme. Mechanisms to improve quality include combining multiple data sources, integration with survey data and quality control and assurance. The list below summarizes the primary points from Sections 2–7.

Sources

- Tax registers. Tax data are used as partial or complete supplements to survey data. In some cases concerns about under-reporting in surveys exist, and tax data are judged more reliable than expenditures reported in surveys. In Canada, tax data substitute survey data on agricultural expenditures because they are considered more reliable.
- Land registers and cadastral surveys. Complete enumeration of land ownership history is an important source of administrative data in India and in Europe.
- Sources of information on animal populations. Cattle-tracing systems, records from veterinary visits and information from the enforcement of government regulations provide thorough sources of information on animal populations.
Uses

- Sample development and estimation. If an administrative source or register is not of sufficient quality to support estimates independently, the source may be used in sample design and estimation. Administrative data can be used to improve the quality of a sampling frame, as in Canada where administrative sources improve coverage of specialty crops and farms with small land areas. An administrative list can be used as one frame in a multiple-frame survey. Data items from the administrative list may be used to improve the efficiency of stratification. Administrative data can also improve the efficiency of survey estimators. An administrative total may be used as a control in calibration.

Quality

- Performance dimensions. Many of the basic principles that underlie quality evaluations of surveys and censuses apply to administrative data. The ADSAS framework, outlined in Part B of Technical Report 1, adapts a quality framework to the context of administrative data. Performance measures such as timeliness, consistency and coherence are important dimensions to consider when evaluating administrative data. Because operations involving administrative data often involve multiple sources, maintaining standardized and accurate identification variables is especially important in the context of administrative data.

- Fitness for use. The choice of quality evaluation depends on the intended use of the administrative source. If the administrative source is to be used directly for the statistical product, complete coverage and accuracy are essential. If the administrative data source is to be combined with other administrative sources, surveys and censuses, maintaining high-quality identification variables may be at least as valuable as complete coverage of a single source. If the administrative data are to be used for small-area estimation or calibration, high correlation with the items of interest leads to potential gains in efficiency.

Potential Limitations

- Changes in administrative processes. The processes for generating administrative data are often outside the control of statistical agencies. Unexpected changes can lead to inconsistencies in estimates over time or reduced data availability.
• Lack of standardization. Differences between the purpose underlying the administrative data and the statistical objective can lead to differences in the definitions of variables, units and identifiers. This can make synthesizing of multiple sources a challenge. Developing countries are advised to pay attention to definitions at several levels when using and evaluating administrative data.

• Under-coverage. It may be that not all members of a target population participate in the administrative process. An example is the IACS database: because not all farmers participate in subsidy programmes and not all crops are relevant for the programmes, the IACS database does not cover the full population.

• Reporting error. Incentives underlying administrative processes can lead to reporting error. The IACS database provides an example: farmers are found to over-report crop areas to avoid the consequences of inadvertent under-reporting.

• Computational demands. Using administrative data often involves processing and cleaning large volumes of data, which can create substantial computational demands. Statistical offices in developing countries may benefit from investments in methods and infrastructure for processing large amounts of data. In planning a project involving administrative data, the statistical agency should reserve sufficient time and analytical resources to handle computational demands.

**Mechanisms for Improvement**

• Relations with administrative agencies. Understanding administrative data sources is an important component of managing the challenges associated with administrative data. Understanding the definitions, variables and units of the administrative database is important for appropriate use. Forming good relations with agencies can mitigate potential changes and improve the standardization of items collected through administrative processes and items of interest in a survey.

• Audits and sample surveys. Carfagna and Carfagna (2010) note that audits and sample surveys are valuable tools for checking and improving the quality of administrative data.

• Combine multiple data sources with complementary strengths and weaknesses to overcome problems associated with under-coverage and measurement error. An example is the Swedish farm register, which supplements the IACS database with information from registers such as the business register.
• Develop good identifying variables. Maintaining high-quality identifying variables can be helpful when combining multiple data sources. As developing countries set up register systems, they should seek to create standardized identification variables.

10.1. NEW DIRECTIONS

The explosion of “big data” is creating more administrative data sources that could be used in official statistics. Statistics Netherlands has evaluated traffic loop detection data, social media messages and mobile phone data (Daas et al., 2012a; Buelens et al., 2012), and NASS applies classification techniques to convert satellite data into land-cover maps of the United States. Many of the methodological challenges associated with using new “big data” sources for producing official statistics parallel the challenges in traditional sources of administrative data. The magnitude of the data is encouraging, but analysts need to appreciate the limitations. Daas et al. (2012a), Buelens et al. (2012) and Nordbotten (2008) review new sources of administrative data and challenges relevant to high-volume data.

The routine operations of public and private organizations are creating new sources of electronically observable data. Examples of mechanisms leading to such data include credit card transactions, radio frequency identification, electronic travel tickets and electronic passports. Data generated and collected from precision agriculture may contain information relevant for agricultural statistics. Daas et al. (2012a) consider two sources: traffic loop detection signals and twitter messages: the former provide the number of vehicles passing a monitored location per minute and information on vehicle length and speed; twitter messages are analysed to obtain an indicator of monthly consumer sentiment, which correlates strongly with corresponding indictors obtained from a monthly survey by Statistics Netherlands. Buelens et al. (2012) also consider mobile telephone data as a source of administrative data for use in producing official statistics.

The challenges anticipated in this context parallel the difficulties associated with administrative data in general, but there are subtle differences when considered in the context of “big data”. Sensitive issues can arise associated with data-sharing rights and the privacy of personal information. Computational techniques and electronic infrastructures are needed to handle large volumes of data. New statistical methods or adaptations of recent data-mining techniques may be required to process data into usable formats (Nordbotten, 2008; Daas et al., 2012a).
BueLens et al. (2012 and 2014) emphasize that participants in many processes leading to “big data” are a selective group: they choose to participate in particular data-collection processes and may have characteristics that are different from those of the general population. If the objective of a statistical office is to make inferences about the general population, the potential for selection bias should be evaluated. BueLens et al. (2012) consider model-based and algorithmic methods to account for selection bias under the assumptions that: i) auxiliary information is available for the full population; and ii) the selection mechanism is a function of the auxiliary information. These assumptions parallel the common assumption of “missing at random” (Kim and Shao, 2013) in the literature on missing data.

10.2. LESSONS FOR DEVELOPING COUNTRIES

Because of the importance of agriculture in the economies of many developing countries, agricultural statistics have implications for policies related to issues such as reducing poverty, improving public health and reducing food security. In some African countries, for example, agriculture contributes to 30 percent of the GDP, 50 percent of exports and 75 percent of employment. Agricultural data are therefore required by various stakeholders and can affect policies, with far-reaching implications (Kiregyera, 2002).

Developing agricultural monitoring and planning programmes requires sound data on the characteristics of the agricultural sector. Reliable, consistent and comprehensive statistics are needed to ensure that policies and monitoring programmes address the actual dynamics of a particular domain: work on reducing poverty, for example, must focus on the appropriate areas and sub-populations. Policy-making requires information about the demographic and socio-economic characteristics of targeted populations. PARIS21 (2002) notes: “Decision-makers need quantitative information to (i) analyse constraints, (ii) identify benchmark situations, (iii) set quantified objectives, (iv) monitor implementation and (v) measure the impact of policies and programmes.” Given that two-thirds of people exposed to poverty or food insecurity live in rural areas, agricultural statistics are an integral part of the development of such policies (Kiregyera, 2002; PARIS21, 2002).

The Millennium Development Goals (MDGs) established at the United Nations Millennium Declaration in September 2000 support the development of reliable statistics and agricultural statistics. The eight MDG targets are listed in Table 5. The declaration of the MDGs added 20 specific targets and 60 quantitative indicators and encouraged participating countries to monitor their progress,
partly by establishing programmes for data collection and analysis (Asian Development Bank, 2010).

Table 6: Summary of the Millennium Development Goals

<table>
<thead>
<tr>
<th>Goal 1:</th>
<th>Eradicate extreme poverty and hunger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal 2:</td>
<td>Achieve universal primary education</td>
</tr>
<tr>
<td>Goal 3:</td>
<td>Promote gender equality and empower women</td>
</tr>
<tr>
<td>Goal 4:</td>
<td>Reduce child mortality</td>
</tr>
<tr>
<td>Goal 5:</td>
<td>Improve maternal health</td>
</tr>
<tr>
<td>Goal 6:</td>
<td>Combat HIV/AIDS, malaria and other diseases</td>
</tr>
<tr>
<td>Goal 7:</td>
<td>Ensure environmental sustainability</td>
</tr>
<tr>
<td>Goal 8:</td>
<td>Develop a global partnership for development</td>
</tr>
</tbody>
</table>

Since the establishment of the MDGs several international initiatives to improve the reliability of statistics in developing countries have been implemented. Some African countries have published poverty reduction strategy papers, and in 2002 FAO/AFRISTAT held a workshop on integrated agricultural statistics systems in support of food security. A PARIS21 workshop in 2002 urged that agricultural policy decisions be based on reliable data.

The 2004 Marrakech Action Plan for Statistics established a “… timeline for low-income countries to develop and implement national statistical data systems in time to produce better statistics for national and international use for the 2010 millennium review” (African Development Bank et al., 2007) with a view to helping countries to achieving the MDGs. One of the guidelines involves supporting the integration of statistics from sources such as surveys and administrative databases.


In developing countries, administrative data are often the primary and sometimes the only source of information (Keita and Chin, 2013). Statistics South Africa, for example, noted that it “… produces less than 10 percent of the statistics required to inform national development, while the rest of the data are administrative data produced by other organs of state …” (Kiregyera, personal communication). This is unlike the situation in developed countries, where
administrative data are used mainly in conjunction with surveys and censuses to reduce the burden on respondents, reduce data-collection costs or support more efficient estimators (Carfagna and Carfagna, 2010; Brackstone, 1987).

A challenge in making effective use of administrative data involves overcoming quality deficiencies. Asian Development Bank (2010) asserts that administrative data have the potential to provide cost-effective solutions, but that “… in most developing countries, however, administrative data remain underutilized because of their poor quality as a result of incomplete coverage, biased reporting, and other data quality issues.” Kiregyera (2002) reiterates this concern in the context of administrative data from agricultural reporting services. As discussed in Section 2.6, the ministries of agriculture in some African countries have agricultural reporting services that maintain administrative data, which is often collected by field extension staff. The resulting administrative registers provide regular information on crops but are “… not well maintained and data based on them are notoriously incomplete …” Kiregyera (2002). Reasons for the quality problems include poorly trained staff and lack of standard data-collection systems. And “… given the serious manpower, financial resource constraints, and lack of supervision many countries face, it is common for data not to be collected but to be ‘made up’ by extension staff…” (Kiregyera, 2002).

The structures and constraints of statistical and administrative organizations in developed countries are different from those in developing nations. Carfagna and Carfagna (2010) and Brackstone (1987), for example, show that in developed countries a desire to reduce data-collection costs and the burden on respondents are factors driving the use of administrative data, and that administrative data are used to support, enhance or sometimes replace elements of census or survey programmes. In developing countries, consistent survey or census programmes may not exist and administrative data may be the best available source of information (Keita and Chin, 2013), even though they may be of questionable quality as a result of the lack of standard processes and insufficient resources (Asian Development Bank, 2010; Kiregyera, 2002).

Developing countries face two challenges in expanding the use of administrative data for agricultural statistics: i) improving the quality of the data so that they can be used in formal processes; and ii) developing the analytical resources and technical infrastructure to make use of administrative data, possibly in combination with surveys. The experiences of developed countries in creating processes to use administrative data to produce official statistics carry important lessons: combining different administrative data
Combining Disparate Data Sources

Combining several data sources can be a useful mechanism for quality improvement. The data sources may be administrative databases or may contain information from surveys as well. Using multiple data sources can improve coverage of target populations and can provide a mechanism for imputation and editing. Asian Development Bank (2010) explains: “Administrative data sources cannot be discussed in isolation from other major sources of data, such as sample surveys and censuses. Data produced through these sources should be viewed as complementary, designed to meet the needs for policy planning and evaluating government programs and policies.”

A challenge in combining multiple data sources arises when different data-collection processes provide different measurements of related target concepts. These inevitable differences result from factors such as including differences in data-collection protocols, the phrasing of questions and the incentives of respondents. The latter is particularly prevalent in administrative data-collection when the purpose is linked to a government or private objective.

Differences among data sources do not necessarily prevent the synthesis of different sources of information. Overcoming the challenges associated with conflicts across data sources requires understanding of the reasons for differences across sources. If the differences between sources are systematic, procedures can be identified to reconcile them. Section 7.6 provides examples of manual and model-based procedures for reconciling conflicting data sources.

Formation of Registers and Register Systems

In setting up register systems in developing countries, the experiences and observations from statistical offices in developed countries may provide valuable lessons. Wallgren and Wallgren (2007) emphasize the importance of maintaining the quality of register systems: “It is generally not a good idea to produce statistics directly from the received administrative registers because these are not adapted to statistical requirements. The object sets, object definitions, and variables need to be edited, and it will often be necessary to carry out some kind of processing so that the register fulfils the statistical requirements for objects and variables. The register-statistical processing, which aims to transform one or several administrative registers into one statistical register, should be based on generally accepted register-statistical methodology.”
An important variable in a register system is the identifying variable. Wallgren and Wallgren (2010) emphasize the importance of maintaining high-quality identifying variables to assist with probabilistic record linkage and to simplify the process of combining multiple data files.

**The Illusion of Cost Savings**

A major motivation for using administrative data is to reduce cost, in that data collection is not required as in surveys and censuses. But the costs of data collection may be transferred to statistical offices through requirements such as data management and processing. Human resources are required to maintain and verify data quality, computational infrastructure is needed to manage large volumes of data and methodological expertise is required to develop and implement statistical techniques to draw inferences from administrative data. Statistical offices are advised to evaluate cost requirements when seeking new sources of administrative data.

**The Importance of Quality Control and Assurance**

It is essential to adopt best practices to maintain data quality standards. Standardizing definitions and data-collection protocols and identifying variables are helpful when reconciling data sources, and can reduce the magnitude of the error in a final estimator based partly or completely on administrative data. Supervision of data collection and clear instructions for collectors have been used to improve the quality of the resulting data. Providing sufficient metadata will help users to interpret the data. Many of the principles that apply to quality control of survey data extend to administrative data.

**The Value of Surveys and Censuses**

Survey and census programmes are valuable in a statistical system that incorporates administrative data. Asian Development Bank, 2010 notes: “Improving the administrative data systems in a developing country is a long-term process, and sample surveys are needed to fill the data gaps for monitoring and attainment of the MDGs. Censuses provide benchmark data on the structure and other basic characteristics of the population under study”. In contexts such as frame maintenance and model-based small-area estimation, administrative data help to optimize survey and census programmes, which have the advantage that statistical agencies can participate in data collection and estimation. Statistical offices can help to design questionnaires, follow-up procedures for non-response, editing and imputation techniques and estimation procedures. In the context of the IACS database, Carfagna and Carfagna (2010) explain that: “… sample surveys must be performed or the statistical system will produce biased results.” Survey data can provide information about the biases of an
administrative source. Although administrative data are attractive for the apparent savings in data-collection costs, the benefits are often apparent when they are used in conjunction with information from surveys or censuses.

**Evaluating the Impacts of Selectivity**

As discussed in the Section 1, the existence of a selection process that defines a particular sub-group of a population distinguishes administrative data from information obtained through surveys or censuses. This defining characteristic underlies many of the benefits of administrative data such as reduced data-collection costs and spatial and temporal detail. It is also the source of challenges associated with the use of administrative data: a population of interest may differ from a population defined by administrative processes, for example, and variables of importance for administrative purposes may differ from definitions of related concepts relevant for statistical purposes. Buehns *et al.* (2014 and 2012) discuss the issue of selectivity in the context of “big” administrative data sources. In using administrative data sources, an evaluation of the impacts of the selection process on estimates is important.
References


Project on Integration of survey and Administrative Data. Report of WP2: *Recommendations on the use of methodologies on the integration of surveys and administrative data.*


