UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses

Prepared by the Conference of European Statisticians Task Force on Assessing the Quality of Administrative Sources for Use in Censuses

United Nations
Geneva, 2021

Preface

The main purpose of this publication is to provide the producers of population and housing censuses with guidance on how to assess the quality of administrative data for use in the census. The Guidelines cover the practical stages of assessment, from working with an administrative data supplier to understand a source, its strengths and limitations, all the way to the receipt and analysis of the actual data. The Guidelines cover key quality dimensions on which an assessment is made, using various tools and indicators. For completeness the Guidelines also include information about the processing and output stages of the census, with respect to the use of administrative sources.

The publication was prepared by a Task Force established by the Conference of European Statisticians (CES), composed of experts from national statistics offices, and coordinated by the United Nations Economic Commission for Europe (UNECE).

Acknowledgements

These Guidelines were prepared by the UNECE Task Force on Assessing the Quality of Administrative Sources for Use in Censuses, consisting of the following individuals:

Steven Dunstan (Chair), United Kingdom
Katrin Tschoner, Austria
Christoph Waldner, Austria
Josée Morel, Canada
Lionel Espinasse, France
Stefan Dittrich, Germany
Tobias Kahlenberg, Germany
Thomas Körner, Germany
Ingeborg Vorndran, Germany
Sheelagh Bonham, Ireland
Brendan Murphy, Ireland
Alaa Atrash, Israel
Yael Feinstein, Israel
Gerardo Gallo, Italy
Donatella Zindato, Italy
Snezana Remikovic, Montenegro
Eric Schulte Nordholt, Netherlands
Christine Bycroft, New Zealand
Abby Morgan, New Zealand
Janusz Dygaszewicz, Poland
Krzysztof Woźnica, Poland
João Farrajota, Portugal
Sandra Lagarto, Portugal
Paula Paulino, Portugal
Dmitrii Calincu, Republic of Moldova
Valentina Istrati, Republic of Moldova
Marina Pérez Julián, Spain
Alberto Salcedo, Spain
Şebnem Beşe-Canpolat, Turkey
Muharrem Gürleyen Gök, Turkey
Mehmet Şaban Ucari, Turkey
Louisa Blackwell, United Kingdom
Adriana Castaldo, United Kingdom
Sara Correia, United Kingdom
Sara Haylock, United Kingdom
Charlotte Hillyard, United Kingdom
Jack Sim, United Kingdom
Stephan Tietz, United Kingdom
Clare Watson, United Kingdom
Marina Wright, United Kingdom
Tom Mule, United States of America
Eduard Jongstra, UNFPA
Diana Beltadze, Eurostat
Sorina Vâju, Eurostat
Ian White, Independent expert
Fiona Willis-Núñez, UNECE

The Guidelines were developed and agreed upon by the entire Task Force. Each chapter was drafted by a team under the leadership of one or more individuals, as follows:

- Census methodologies and uses of administrative data for censuses: Sara Correia and Ian White
- Quality framework: Sara Correia and Sorina Vâju
- Source Stage: Sorina Vâju, Josée Morel, Diana Beltadze and Steven Dunstan
- Data Stage: Christoph Waldner, Tobias Kahlenberg and Sara Correia
- Process Stage: Sara Haylock, Abby Morgan, Adriana Castaldo and Steven Dunstan
- Output Stage: Sara Correia, Marina Pérez, Sandra Lagarto and Steven Dunstan.

The Task Force extends particular thanks to the United Kingdom's Office for National Statistics for the invaluable contributions of many of its staff to these final Guidelines, most notably Steven Dunstan and Sara Correia who ensured that the Task Force met its goals and who brought this product to its completion.

Contents

List of boxes

List of figures

List of tables

List of country case studies

Acronyms and abbreviations

ABPE	Administrative data-based population estimate
ABS	Australian Bureau of Statistics
AIDA	Automazione Integrata Dogane ed Accise (Integrated Automation for Customs and Excise, Italy)
CAPI	Computer-assisted personal interview
CATI	Computer-assisted telephone interview
CAWI	Computer-assisted web interview
CAxI	Computer-assisted multi-mode interview
CES	Conference of European Statisticians
CIS	Customer information system
CT	Census test (Portugal)
DES	Dual estimation system
DA	Demographic analysis
ESS	European Statistical System
ESSnet	European Statistics System Network
FPC	fichero precensal (pre-censal file, Spain)
GSBPM	Generic Statistical Business Process Model
HMRC	Her Majesty's Revenue and Customs
INSEE	Institut national de la statistique et des études économiques (National institute of statistics and economic studies, France)
ISSR	Integrated system of statistical registers (Italy)
ISO	International Organization for Standardization
LMS	Legal marital status
MOU	Memorandum of understanding
NHS	National Health Service (United Kingdom)
NIP	numer identyfikacji podatkowej (tax identification number, Poland)
NRFU	Non-response follow-up
NSO	National statistical office
NZ	New Zealand
ONS	Office for National Statistics (United Kingdom)
PBR	Population base register
PE	Population estimate
PESEL	Powszechny Elektroniczny System Ewidencji Ludności (Universal electronic system for registration of the population, Poland)
PIA	Privacy impact assessment
PII	Personally identifying information
PIN	Personal identification number
PPHC	Permanent population and housing census (Italy)
PR	Patient register
QA	Quality assessment
REGON	Rejestr Gospodarki Narodowej (Business identification number, Poland)
ROC	Receiver operating characteristic
SCD	Statistical census dataset
SDC	Statistical disclosure control

SE	Statistic Estonia
SOL	Signs of life
SP	Statistics Portugal
SPD	Statistical population dataset (Portugal)
Stats NZ	Statistics New Zealand
UK	United Kingdom
UKSA	United Kingdom Statistics Authority
UNECE	United Nations Economic Commission for Europe
UPRN	Unique property reference number
VOA	Valuation office agency
VQS	Variable quality system

Executive summary

The use of administrative data in censuses continues to increase across the countries of the UNECE region and beyond, whether it be to support a traditional census, or under a combined or register-based census methodology whereby the population is enumerated and/or the census variables populated via administrative data sources. It is therefore important that National Statistical Offices (NSOs) understand the strengths and limitations of administrative data for use in their censuses to ensure that the right decisions are made about the use of such data.

These Guidelines aim to provide census producers with a practical guide for assessing the quality of administrative data, through a series of Stages of assessment. The Guidelines draw on quality frameworks and best practices adopted by NSOs across the world, including the widely-used framework of Statistics Netherlands (Daas et al. 2012), the New Zealand Total Error Framework (Zhang 2012) and the deliverables from the Statistical Network Methodologies for an Integrated Use of Administrative Data in the Statistical Process project (SN-MAID, 2014).

The Guidelines are based on four Stages: Source, Data, Process and Output, with the first two Stages being the principal focus of the Guidelines, providing an assessment of input quality (i.e. the quality of administrative data sources, set against their use in a census). The Process and Output Stages are provided for completeness and give the reader information about the key processes and considerations for transforming administrative data for use in a census, and for assessing the quality of census outputs that are based on administrative data.

The Source Stage covers the assessment of the administrative source through working with the data supplier and reviewing relevant metadata. The Stage includes an assessment of whether the source can meet the needs of the census, under the quality dimensions of relevance, accuracy, timeliness, coherence and comparability. An assessment is also made of the accessibility and interpretability of the administrative source, covering any restrictions on access and use, and public acceptability. Finally, an assessment is made of whether the data supplier can meet the needs of the NSO, considering factors such as the strength of the relationship with the supplier and the status of the supplier.

The Data Stage covers an assessment, based on an analysis of the actual data (as transmitted by the data supplier) and through comparisons with other sources. The Stage includes the validation of data on arrival, an assessment of accuracy and reliability, including coverage and measurement errors; timeliness and punctuality; and an assessment of linkability. For the Source and Data Stages, the assessment is against key data quality dimensions, for which various tools and indicators are provided.

The experiences of several countries are included throughout the Guidelines, via basic illustrations or via more detailed case studies. The Guidelines also provide a set of key recommendations which are summarized in the concluding chapter, along with suggested areas for further work.

Chapter 1. Introduction

1.1 Background

1. In 2017, the UNECE Task Force on register-based and combined censuses prepared the Guidelines on the use of registers and administrative data for population and housing censuses[1]. The Guidelines included a section on "data sources and their quality" with a general discussion of this topic. Experts at the UNECE-Eurostat Expert Meeting on Population and Housing Censuses (Geneva, 4-6 October 2017) identified the quality of administrative sources as a topic of primary importance for many countries. As a consequence, the Expert Meeting called for the establishment of a new UNECE Task Force on measuring the quality of administrative sources for use in censuses, building on the work of the previous Task Force.

2. The Task Force was established in 2018, with its Terms of Reference[2] approved at the February 2018 meeting of the Bureau of the Conference of European Statisticians (CES) in Helsinki (14-15 February 2018). The Task Force reported to the UNECE Steering Group on population and housing censuses, which in turn reports to the CES and its Bureau.

3. The objective of the Task Force was to develop guidance on the measurement of the quality of administrative sources for use in censuses[3]. The terms of reference stipulated that guidance should be developed that is relevant to all UNECE countries, and that it should build on the work of Eurostat's ESS.VIP ADMIN project[4] on the use of administrative sources in the production of official statistics.

4. The Task Force met in person during the 2018 and 2019 UNECE-Eurostat Expert Meetings on Population and Housing Censuses, and held a further in-person meeting in Geneva, Switzerland on 5-6 March 2020.

5. The Task Force presented annual reports of its progress to the UNECE-Eurostat Expert Meetings on Population and Housing Censuses in 2018, 2019 and 2020. A full draft of these Guidelines was circulated in advance of the 2020 Expert Meeting (online, 30 September-1 October 2020) and feedback received from participants was used to refine the Guidelines.

6. These Guidelines serve as a practical toolkit for the assessment and measurement of the quality of administrative sources for population and housing censuses.

[1] Available from http://www.unece.org/index.php?id=50794

[2] Available from http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/bur/2018/February/06Add.1-TF_on_quality_of_admin_data_for_censuses__ToR__apr.pdf

[3] The Task Force subsequently decided to adjust its title and the corresponding objective to 'assessing' rather than 'measuring' the quality of administrative sources for use in censuses.

[4] More information on this project can be found at https://ec.europa.eu/Eurostat/cros/content/ess-vision-2020-admin-administrative-data-sources_en

1.2 Use of administrative data in censuses

7. Administrative data sources are data holdings that contain information collected primarily for administrative purposes. This includes data collected by government departments, public bodies and other organizations for purposes of registration, transaction and record-keeping, usually during the delivery of a service. They include administrative registers (with a unique identifier) such as a country's population, business, address, education, health, employment and tax registers, as well as other administrative sources (without a unique identifier). Administrative registers and/or other administrative sources are used to create statistical registers, which are specifically used for statistical purposes, including for the census. The administrative sources most commonly used in censuses are outlined in Chapter 2 of the Guidelines.

8. The use of administrative data sources in censuses varies across countries. Such sources may be used to enhance or to supplement a traditional census, to conduct a combined census, or in the construction of a fully register-based census. There has been a clear trend towards increased use of administrative data in censuses. This has been motivated by the benefits administrative data can bring, including reduced cost and respondent burden, improved timeliness and frequency of results, improvements to quality, and greater flexibility to respond to user needs (see, for example, section 4.1 of UNECE 2018). Furthermore, the conditions within many countries have changed to support and facilitate the use of administrative data throughout national statistical systems (c.f. section 4.2 of UNECE 2018). This includes through changes in legislation, public and stakeholder acceptability, and through developments in technology and statistical methodologies.

9. The importance of administrative data has been further highlighted by the challenges that National Statistical Offices (NSOs) are now facing when it comes to collecting data directly from the population, whether due to a reluctance of the public to engage with the census, or their ability to do so. This challenge was emphasized at the onset of the Covid-19 crisis in 2020, where both the public's ability to engage with NSOs and at the same time NSOs' ability to engage with the public were affected significantly. The various ways in which administrative data are used in censuses are covered in Chapter 2 of these Guidelines.

1.3 Key risks to quality

10. For all the benefits that administrative data can bring, there are a number of key quality considerations that must be assessed and evaluated before incorporating an administrative source into a census. First, the NSO will have only limited control over the way the data are collected and processed. There is therefore a significant dependency on the authorities holding the administrative data. For example, if the administrative authority is unable to meet the NSO's requirements with respect to providing the right data at the right time, this will impact the timeliness of the census results. Similarly if the administrative authority does not adequately engage with the NSO on any potential changes to the source, this could impact coherence and comparability.

11. Second, the use of administrative data by the NSO for purposes other than those for which the data were originally collected raises privacy, security and legal concerns. The NSO must therefore assess public acceptability, seeing that the required assurances are in place and that they are communicated to the public (and to the administrative authority). The use must also be lawful. Without acceptance or agreement both from the public and from the data supplier, or a credible legal basis for the use of the administrative source, there would be significant risk to the reputation of the NSO and its ability to deliver a high-quality census. This can arise if the public changes their behaviour in the way they interact with the administrative authority or the NSO, due to concerns over the way the NSO is using their data.

12. Third, administrative data have (in general) not been collected for statistical purposes. As a consequence, the data sources may have adopted different concepts and definitions from those required by the census; they may refer to different reference periods; and they may have limited coverage of the census population. In addition, the accuracy and completeness of the data will be highly dependent on the importance of the data to the administrative authority's function. The administrative sources may also be subject to changes over time and inconsistencies in the way the data are collected across units of the population.

13. Finally, the complexity of the administrative data and the availability and completeness of the associated metadata will impact the ability of an NSO to understand, access and use an administrative source. For example, administrative data can be held in large, complex data structures, posing significant technical challenges for the NSO to consider and overcome. The complexity of administrative data may also impact the accessibility and clarity output quality dimension from a user's perspective. That is, users of the census may find it difficult to understand the use of administrative data in the census and the impact this use has on the quality of the census outputs.

14. These key quality considerations will inform decisions about the use of administrative data in a census. The Guidelines address each of the considerations in detail.

1.4 Scope and structure of the Guidelines

15. The focus of the Guidelines is on the assessment of the quality of administrative data sources for use in the census (i.e. Input Quality). They do not cover other sources *per se* (e.g. Big Data, commercial data). Nevertheless, much of the material within the Guidelines is applicable beyond administrative data (guidance on the quality assessment of Big Data can be found in UNECE 2014).

16. The Guidelines begin by providing information about the different census methodologies and how administrative data can be used under each of these methodologies, including the types of data sources used. The aim is to provide information that may be useful for NSOs that wish to incorporate new administrative data sources into the design of their censuses (Chapter 2).

17. The next chapter (Chapter 3) outlines the overall quality framework on which the Guidelines are based, which is built around four Stages of assessment. The Stages broadly relate to the lifecycle of using administrative data in the census: from understanding, assessing and working to acquire a source (the Source Stage); receiving the actual data and assessing its quality (the Data Stage); processing the administrative data for use in the census (the Process Stage); to the quality assessment of the census outputs that use administrative data (the Output Stage). The chapter also outlines the dimensions of quality which are assessed within each Stage, covering the associated errors (e.g. representation and measurement errors). The chapter concludes with an outline of the importance of carrying out feasibility research on the use of administrative data, explaining how the Stages within the Guidelines can be used for this purpose.

18. Chapter 4 covers the first Stage of assessment (the Source Stage), where information is gathered about an administrative source through communication with the administrative authority and by reviewing existing metadata. At this Stage the focus is on assessing the relevance of the source against the needs of the census, covering accuracy, timeliness, coherence and comparability, and accessibility. An assessment is also made about the institutional environment, including whether the data supplier can meet the needs of the NSO, considering factors such as the strength of the relationship with the supplier and the status of the supplier.

19. Chapter 5 covers the Data Stage of the assessment, where data are received from the administrative authority and are assessed through analysis of the data and through comparisons with other data sources. During both the Source and Data Stages, the assessment and measurement of quality is set against a number of data quality dimensions, using various tools and indicators. The two Stages together provide an assessment of Input Quality.

20. The information and insight gained through the Source and Data Stages is useful not only to determine whether a particular source should be used in the census, but also to determine the necessary processing of the administrative data for use in a census. In general, administrative data cannot be used directly in a census, due to conceptual and definitional differences and due to limitations of coverage, completeness and accuracy. It is therefore necessary to transform the data from administrative sources (including registers) using the information gained at the Source and Data Stages. Some of the most important processes and the associated quality considerations are covered in Chapter 6 of the Guidelines, including record linkage, the construction of statistical registers and the 'signs-of-life' methodology, enumeration of the census population, making decisions between sources, and data editing and imputation.

21. The Source, Data and Process Stages relate directly to the quality of census outputs in accordance with the European Statistical System (ESS) output quality dimensions. Conversely, the assessment of the census outputs will provide valuable information about where there may be limitations or concerns about the administrative data, or the processing of these data, that were not identified initially at the Source, Data and Process Stages. There is therefore an iterative process of assessment, which can inform both ongoing improvements to the

administrative sources (working with the administrative authority to improve the source), and improvements to the processing of the administrative data by the NSO. The assessment of the quality of census outputs which use administrative data is covered briefly in Chapter 7 of the Guidelines, which details the Output Stage.

22. Various country-specific examples and case studies throughout the chapters of the Guidelines illustrate the application of the Stages of quality assessment in practice.

23. Finally, Chapter 8 provides a conclusion to the Guidelines, with a summary of the recommendations that are presented throughout the earlier chapters. The final chapter also offers proposals for further internationally-coordinated work in the area of quality assessment.

1.5 Summary of recommendations

24. The discussions throughout these Guidelines lead to the following recommendations that NSOs are advised to follow:

i. Identify administrative sources against specific use cases, so that assessment can be made with respect to the expected or required outcomes of using the source for the identified use case.

ii. Build and support relationships between NSOs and data suppliers, with a legal basis for supply and use of data and collaborative mechanisms for feedback.

iii. Harness supplier relationships to ensure comprehensive understanding of source metadata.

iv. Assess the coherence and compatibility of the administrative source, striving for a detailed understanding of any differences between the required populations, concepts, definitions and time-related dimensions, and those available in the administrative source.

v. Understand restrictions and challenges to acquiring an administrative source and integrating it into a census, and weigh up the value of overcoming these challenges against the effort and risk entailed in doing so.

vi. Assess and manage the risk implied by use of an administrative source.

vii. Be transparent in communication with data users and with the public about how and why administrative data are used for the census, emphasizing procedures for ensuring effective use and for data protection.

viii. Undertake feasibility research as a ‘proof of concept’ and test runs with real data prior to including administrative data sources in census production.

ix. Make use of expert review and conduct comparisons between sources and over time to identify quality concerns in a given source.

x. Record and publish results of quality assessment at all stages.

xi. Develop an NSO-specific quality assurance framework and strategy, supported by clear and comprehensive documentation and training procedures, with a focus on continuous assessment and communication between the NSO, users and data suppliers.

Chapter 2. Census methodologies and uses of administrative data for censuses

25. This chapter summarizes the range of census types and uses of administrative data in censuses, which are common throughout UNECE countries. This will help NSOs within the UNECE region and beyond when using administrative data in their censuses – whatever data collection methodology is adopted.

2.1 Census methodologies

26. As has been noted in previous UNECE publications (UNECE 2015; UNECE 2018) there are several different ways to undertake the data collection process in a population and housing census. This section provides an overview of census types and where these Guidelines may be useful to statistical producers.

27. For the sake of simplicity this chapter summarizes only the three main categories of census data collection methods:

- 'traditional' census
- 'register-based' census
- 'combined' census.

28. The UNECE Census Wiki[5], which compiles information on the 2020 round of censuses as reported by member countries, indicates that the trend away from the traditional census continues apace. Out of 52 UNECE countries for which information is available, fewer than half (23) are conducting a traditional census in the current round (with some 13 countries planning to conduct a register-based census and 16 planning a combined approach to data collection). Nevertheless, as discussed below there are still opportunities and benefits for NSOs that conduct a traditional census to exploit the use of administrative data.

29. The key features of the three census types identified are summarized below. A more detailed discussion of the various census methodologies, including the necessary prevailing conditions, advantages and challenges, is given in the CES Recommendations (UNECE 2015).

2.1.1 Traditional census

30. The term 'traditional census' refers, in the broadest sense, to a census based on a direct count of all individuals, households and housing units and the collection of information on their characteristics through the completion of census questionnaires, either in paper form or electronically. The information is collected in the field by means of a full enumeration across the whole country in a relatively short period of time.

31. The information can be collected by one or more of the following methods:

[5] Available at bit.ly/UNECECensusWiki2020

- directly from households (with delivery and collection of paper forms undertaken by enumerators, the postal service or other methods)
- online, using electronic questionnaires
- by enumerators during a face-to-face interview of the household using either paper or electronic questionnaires.

32. Since 2001 some countries have made significant changes to their data collection operations while still falling within the definition of a 'traditional' design. For example, in the United States, the Bureau of the Census focuses on collecting only the basic, mainly demographic, data on a 'short form' in a full enumeration in the decennial census year. A large-sample household survey then collects and tabulates the more detailed demographic, social, economic, and housing data every year throughout the decade, replacing the need for a census long form that had previously been sent out to a sample of the population.

33. In contrast, France's INSEE has adopted a different approach: a 'rolling census' is conducted by means of a cumulative continuous sample survey, covering the whole country over the decennial period rather than an enumeration carried out simultaneously, in all areas, on a specific reference date. Such an annual survey may be conducted over the course of a year, in a particular month, or a shorter time frame. With such an approach it is possible to build a sample framework in order to produce

- national results with a single annual survey
- regional results by cumulating several consecutive annual surveys
- small-area results by cumulating data from a more substantial number of years.

2.1.2 Register-based census

34. The register-based census is a totally different approach initially developed by the Nordic countries in the 1970s, among which Denmark was the first to conduct a fully register-based census in 1981. Under this approach there is no direct collection of data from the population, and the traditional enumeration is replaced by the use of administrative data held in various registers (such as a population register, building or address register, social security register, tax records, etc.) through a matching process, usually making use of personal identification numbers. Once a good quality system of statistical registers has been established, this approach permits the (often more frequent) production of census data at a greatly reduced cost and with far less human effort.

35. This methodological approach clearly demands the greatest use of administrative sources and is therefore heavily dependent on establishing and ensuring the highest levels of quality of data from such sources.

2.1.3 Combined census

36. Since the 1990s several other countries within the UNECE region and elsewhere have developed innovative methods to conduct their census, combining the use of administrative data with the collection of an often reduced set of data from a field enumeration of the

population. The field enumeration may still be the primary method for collecting census data. However, administrative sources are used where available in order to reduce response burden and add extra information not collected in the census (e.g. income-related questions). As such, the field enumeration aims to derive specific variables for which the relevant data are not readily available from any administrative source. Under this combined approach, the field data collection may cover the whole population or just a sample.

37. This methodological approach has been used recently by several NSOs in their transition from a traditional to a wholly register-based census. These Guidelines have been written primarily to help statistical producers through such a transition or when running a combined or register-based census. Nonetheless, it will also support the assessment of administrative data used in a primarily traditional census.

2.2 Uses of administrative data

38. The extent of the use of data from administrative sources for the purposes of carrying out a population and housing census clearly will depend on the type of methodology used in the data collection operation.

39. Across the different types of census methodology described above, administrative data can be used in a variety of ways. Among these, the following use cases emerge as key:

- In the construction and optimization of census sampling frames and field operations (as adopted by USA and Canada)
- To replace and add new census variables (as adopted by the UK)
- In the construction of population registers and direct use of administrative data-based enumerations for the census (as seen in Spain and New Zealand, respectively)
- To enable the quality assessment of census estimates by comparison with administrative sources and to inform adjustments through, for example, editing and imputation (as adopted in Estonia)
- In a full administrative data-based census (such as in the Netherlands).

2.2.1 Construction and optimization of census sampling frames and field operations

40. The first use case is the use of administrative data for the construction and optimization of census dwelling / address frames and field operations. This includes assessing the quality of a census sampling frame constructed from administrative data. It also includes the use of administrative data to determine whether an address is likely to be occupied and by whom, or whether a certain address is likely to be ‘hard to reach’, thereby optimizing census field operations.

41. For those countries where some element of a field enumeration is retained – either in a fully traditional census or where a combined approach is adopted – data from administrative sources can be used to support the field operation. Many such countries may, for example,

use information from address or building registers to construct consistently-sized enumeration areas that contain broadly the same numbers of households or dwellings.

42. Alternatively, such information can be used to select appropriate samples of households or housing units where a full dataset is not collected from the whole population.

43. The quality assessment of administrative data-based census frames will benefit from an assessment of data sources at the Source, Data and Process Stages proposed in these Guidelines. However, given the iterative nature of the field operation (i.e. the census frame improves throughout collection), such an assessment may emphasize aspects of coverage (linked to relevance) over the accuracy dimension.

2.2.2 Replacing and/or adding new census variables

44. The second use case is concerned with assessing the quality of administrative data used to replace and add new variables to the census.

45. Where countries decide to reduce the size (and concomitantly, the cost) of a full field enumeration by adopting a combined census approach, data from appropriate administrative sources can be used to replace the information collected from a household questionnaire. For example, reliably accurate information on marital and employment status or the year of immigration may be readily available from administrative registers, thus eliminating the need to collect such data directly from individual persons.

46. Alternatively, a valid case may be made by users for the NSO to collect information in the census either that has been shown to be publicly sensitive or that requires a level of detail which many individuals may be unable to report accurately on a traditional questionnaire. For example, information relating to infant deaths may be culturally sensitive in some countries, while data on household income may often require potentially confidential information to be shared among other household members. In such cases the equivalent data relating to the linked individual may be obtained from administrative sources (such as vital registration or tax records).

47. Quality assessing source data at the Source Stage can aid the decision on what administrative sources to use in such cases. In addition, assessing the particular chosen source(s) at the Data and Process Stages will ultimately ensure the quality of the outputs. As such, when adding new or replacing existing census variables, it is suggested that producers follow all of quality Stages proposed in this guide.

2.2.3 Construction of statistical registers and the direct use of administrative enumerations

48. The third use case relates to the quality of administrative sources used in the construction of population registers and/or the creation of new census units from administrative data (referred to as ‘administrative enumerations’). The creation of such enumerations may include modelling the characteristics of high-quality records through direct linkage of administrative data (e.g. population registers) to the census data. A distinction is drawn between situations where NSOs are able to rely on unique identifiers to integrate multiple sources into one register, or where such identifiers do not exist (and thus

where reliance on deterministic/ probabilistic methods for entity resolution and to link sources on variables such as name, date of birth, address, etc is required).

49. The United Nations (2014) has noted that population registers are now well established in several countries, especially those in the UNECE region, where they have been effectively used as a statistical data source for decades and they may be considered the logical product of the evolution of a vital statistics system. They have become an important source of information for various statistical surveys, including, in particular, the population census.

50. Basic characteristics that may be included in a population register are date and place of birth, sex, date and place of death, date of arrival/departure, citizenship(s) and marital status. Depending on the possibility of proper linking with other registers, much additional information (though not recorded on the population register itself) may be added to the single record, such as language(s), ethnicity, educational attainment, parity, activity status and occupation. Moreover, if complete, population registers can produce data on both internal and international migration through the recording of changes of residence as well as the recording of international arrivals and departures. Such registers can thus be used as the direct base for a so called 'administrative enumeration' to replace a traditional field operation.

51. As with the previous use case, quality assessing source data at the Source and Data Stages will be essential in designing a methodology for the construction of population registers. Ultimately, this will be an iterative design process, where quality assessments at the Output Stage may reveal issues to be addressed at earlier Stages. As such, it is suggested that when constructing registers, producers of statistics follow all of quality Stages proposed in this guide.

2.2.4 Quality assessment and adjustments

52. The fourth use case relates to the quality of the data source to be used for the enhancing of existing census variables. In this type of use case, administrative data is used for the editing and imputation of an existing census variables, as opposed to the direct/complete replacement of a traditional collection.

53. Even in those countries that continue to carry out a traditional censuses, data from administrative sources can be used to either quality assure the information collected from households or to adjust such data where it can be shown that there are errors or omissions to the data collected in the field. For example, in those instances where the NSO chooses not to carry out a post enumeration survey (or surveys) to assess coverage and the quality of responses, data from administrative registers can be used to set parameters within which responses to particular questions may be considered acceptable.

54. Moreover, where the reported data in a traditional census contains errors of substance or omission, the incorrect responses may be edited and/or the missing responses imputed using either the information recorded in the census itself from similar households or the data relating to the variable and individual in question as recorded in a corresponding register.

55. When using administrative data to quality assess census data (collected in the field), the Source, Data and Process Stages are key. In addition, while outside the scope of these Guidelines, it is important to consider issues of circularity with respect to the overall design of the census. For example, where an administrative data source has been used to impute missing values in the census data, or altogether replace a census variable, it should not also be used in its quality assessment.

2.2.5 Full register-based census

56. Finally, the last use case, concerns measuring the quality of sources where the entirety of the census is conducted based on an administrative-based population register, instead of a traditional census methodology.

57. Clearly, the most widespread use of data from administrative sources occurs, by definition, where NSOs undertake a wholly register-based census. As such, in the context of a full register-based census, assessing quality at each of the proposed Stages is vital.

58. Here, the quality of census outputs is particularly dependent on the continuous improvement of source, data and process quality. In countries where register-based censuses are conducted, the quality and stability of the underlying administrative sources at these earlier stages is such that register-based census results are considered the 'gold standard'. The collection of census data in this way does not, however, preclude the NSO from undertaking a field-based post-enumeration survey as a means of independently quality assuring the coverage or content of the counts in the resulting census database.

2.3 Types of administrative sources

59. As the Conference of European Statisticians (CES) has noted (UNECE, 2015) the development of a register-based population census system (whether within the context of a full register-based or combined approach) is a long process, which might take many years. Many countries will choose to continue to retain elements of a traditional data collection in some way even when they start to use administrative registers as an alternative source of data.

60. This section of the Guidelines discusses briefly some of the types of administrative sources from which data are more commonly used by NSO for the purpose of the census and the particular uses to which the data from each can be put. Where appropriate, these uses refer to the particular topics that are currently recommended by the CES to be included in the census (UNECE, 2015).

2.3.1 Use of administrative sources to support a traditional census

61. The extent of the use by NSOs of data from administrative sources increases progressively from census to census as the move from a traditional field enumeration, through a combined approach to a full-register based census, develops. But even those countries that continue to adopt a traditional field enumeration are likely to use administrative data increasingly to support the census operation in one way or another.

62. For example, use of **address registers** is now commonly made by NSO to create lists of dwellings and households from which enumeration area can be constructed and mapped so that they provide balanced workloads for enumerators, or provide stratified sampling designs for post-enumeration or other sample surveys. The creation of such a purpose-built address list by the NSO may involve the amalgamation of data from several separate and independent registers (that may have been constructed for different administrative purposes) in order to minimize under- or over-enumeration. For example, lists of registered electors used for national and local voting purposes or lists of dwellings used by local authorities for assessing rateable values may not include all postal addresses used by national or commercial mail carriers. Moreover, buildings identified by a national mapping agency for the purposes of producing accurate large-scale official maps may identify the location of addresses that are not used for residential purposes and which are often excluded from the census database.

63. Also, those NSOs undertaking a traditional census may use data from administrative sources in the process of assessing the quality of the data collected on the household questionnaire. Thus, data from a **national vital registration** system, for example, should provide very accurate information on the numbers of births and deaths during successive 12-month periods before the census with which the data on the ages of young children recorded in the census can be compared and benchmarked. Similarly, data on any changes of address that are required to be reported to local authorities for the purposes of maintaining **population registers** can be used to validate the information collected on migration since the previous census.

64. It should be noted however that where such data are used not only to *assess the quality* of the information recorded on the questionnaire but also to *supplement* the census data itself to account for any missing or incorrect responses, then the census can be considered to have progressed from a traditional to a combined approach methodology.

2.3.2 Use of administrative sources to derive populations or particular census characteristics

65. The most common use of administrative sources for the purposes of a census is in providing data from which the required output variables can be derived without having to collect the relevant information directly from the public. The type, structure and content of such administrative sources will, of course, vary from country to country depending on the administrative purposes for which the data is used by the data owners, but the most common generic types of registers used for this purpose are summarized here:

66. *Population registers* are registers (often held by the national government department (and/or appropriate local authorities) with responsibility for internal security matters) that provide a frame of persons usually resident in a given country. They are typically maintained to fulfil any legal requirement that both nationals and foreigners residing in the country should register with the local authorities. Aggregation of these local accounts results in a record of population and population movement at the national and local level. Additionally, they often record information on some characteristics of individuals from which data on several core census topics can be derived, such as date and place of birth, sex, date of arrival/departure, citizenship and marital status for each resident person by place usual residence (however that may be defined).

67. *Social security registers* are registers held by official bodies typically for the purposes of the administration of national contributory social insurance programmes and the allocation of benefits and allowances encompassing, for example, the unemployed, families, pensioners, and the disabled and long-term sick. The data from such registers may be used to derive census attributes for such topics as sex, age, marital status, unemployment status, income and disability/health status.

68. *Tax registers* are registers held by national and local tax authorities for the purposes of the administration and collection of income tax, purchase taxes, building rates and other national and locally-levied taxes. The data from such registers may be used primarily to derive census data on personal or household income that might otherwise be difficult, or too sensitive, to collect directly on a household questionnaire. Other information held on such registers may also include details of sex, age, marital status, employment status, occupation, place of work and place of usual residence.

69. *Employment registers* are the registers from which the country's official employment and unemployment figures are derived. The data recorded may enable the NSO to derive census figures relating to the key socio-economic topics of economic activity, employment status, occupation, hours of work and place of work – the latter two enabling information on travel-to-work patterns to be analyzed.

70. *Business registers* hold information to underly the provision of range of services that can vary from country to country but most principally their aim is to register, monitor and store corporate information, such as a company's legal status, its headquarters, capital and legal representatives. The NSO may be able to use this information to derive census data on economic topics, particularly industry.

71. *Education registers* are maintained both centrally and by individual educational and academic establishments for the purpose of registering admissions and the performances of students as well as the employment of teaching staff. The data held may be used by NSOs to create census statistics on attendance, literacy and highest level of educational attainment – though it should be recognized that such available data may often only refer to the current student population. Data on such topics with respect to persons no longer formally attending at an educational establishment must therefore be obtained from other sources.

72. *Health registers* are maintained by locally-based health authorities for the purposes of providing health-related services, whether these are within the context of national health service or provided by insurance-based private agencies. The raw information they contain are usually treated as confidential but can be anonymised to a sufficient extent to allow them to be used by the NSO to create data on health status, domain and level of disability, and parity.

73. *Building and dwelling registers* are registers held usually by land and property valuation agencies and by local authorities responsible for the development of housing policies and urban planning. They may include information relating to the ownership, size and physical construction of individual housing units, but may not necessarily relate these to the persons living in them. The data held may enable NSOs to obtain data to create census

statistics relevant to the needs of a housing census, such as type of dwelling, floor space, floor level, construction materials and period of construction, and may also distinguish between residential and non-residential buildings.

74. NSOs may also be able to access data from other administrative sources to provide topic-oriented census outputs. For example: *registers of motor vehicles* may allow the collection of data on car availability; *registers of foreign nationals* may provide information on migrants, year of entry into the country, citizenship and asylum seekers; *lists of military service personnel* may (if access by the NSO is permitted) indicate employment within the armed forces; *prison registers* can provide some basic information on members of a population group that is particularly difficult to enumerate in a traditional census operation; and *registers held by public facility service providers* (may offer information on the availability of household amenities such as piped water supply, electricity and/or piped gas, and sewage and waste disposal facilities.

Chapter 3. Quality framework

75. The quality of statistics depends on whether the statistical output satisfies its intended use. For example, the ESS definition of quality is derived from the ISO 9000 family of standards, "the degree to which a set of inherent characteristics of an object fulfils requirements" (ISO 2015). In official statistics, the object may include "a statistical product, service, process, system, methodology, organisation, resource, or [data] input" (Eurostat 2020, p.17). In a census context, the quality of administrative data used should therefore be considered in relation to the ways data are collected and processed by suppliers and NSOs, through to the final census outputs.

76. Throughout the above stages however, errors may occur which will compromise quality. Here, error is understood as the difference between a final estimate and the true population parameter it represents. This is highlighted in the Generic Statistical Business Process Model (GSBPM), which provides a standard structure to describe most statistical processes and includes "quality" as an aspect which cuts across all its stages (ESSnet 2014). In addition, Lothian et al. (2019) also argued for the need to understand the whole statistical production process when dealing with alternative data sources such as administrative data. As such, assessing the quality of administrative sources requires mapping the errors which may occur before and after the data is supplied to NSOs and determining how any such errors can be mitigated (e.g. through changes to collection, processing and/or integration with other sources). As such, these Guidelines identify four broad Stages of census production: Source, Data, Process and Output. They then set out how the quality of administrative data may be assessed, by identifying the key quality dimensions at each Stage and the respective tools and indicators for quality assessment.

77. As well as drawing on the GSBPM, this approach also draws on Daas et al. (2009), who identified cross-cutting areas which concern quality or "views" of quality which they call 'hyperdimensions', relating to the source, metadata and data (2009 p. 3). Each of these views comprises several data quality dimensions, each of which is assessed via quality indicators. In line with this approach, these Guidelines also identify quality dimensions, indicators and methods used in the assessment of administrative data, with a particular focus on censuses. At the same time, it was considered that focusing on census production stages would be more intuitive for producers of statistics, for whom these Guidelines were written. Focusing on production stages highlights that quality is an inherent part of statistical design and enables NSOs to focus on the part(s) of the Guidelines which are most relevant to their use-case and/or current production stage.

3.1 Quality and Error in Censuses

78. Where official statistics are produced using a sample survey methodology, survey questions are designed and tested to reduce measurement errors, thus ensuring maximum accuracy and reliability. Thus, the error of the estimates produced are assumed to be caused by deficient sampling and are typically measured and communicated using the Mean Squared

Error (MSE) framework and/or through confidence intervals. However, such measurements do not capture non-sampling errors and these are particularly important in the context of censuses, where the aim is to the capture the full population. As such, similarly to statistics produced with administrative data more generally, the key sources of error in the context of censuses are not sampling errors, but representation (coverage) and measurement errors (Zhang 2012). A common practice is thus to adjust census estimates based on the results of a post-enumeration survey – although this can lead to controversy as complex dual estimation methods (see Chapter 7) are ill-understood by the public.[6]

79. Where administrative data and other alternative data sources such as big data are used in censuses, the range of possible errors is greater than in a traditional census, because data collection processes are not controlled by NSOs. Zhang (2012), drawing on Groves et al. 2004, distinguishes between two broad types of error in statistics produced using administrative data: measurement and representation errors. The first relates to errors in the measurement of characteristics (e.g. age, gender etc), while the second to errors in the representation of population units or objects (e.g. individuals or households in a census).[7] Zhang also distinguishes between the quality of single sources as provided by data suppliers and the quality of transformed and/or integrated sources, after processing by the NSO. This approach is mirrored in the Guidelines which assess the quality of single administrative sources (see Source and Data Stages below) and integrated sources (see Process and Output Stages), with a particular emphasis on identifying measurement and representation errors.

80. Furthermore, the total survey error (TSE) framework has also been adapted to assess the quality of administrative data. In contrast to MSE, TSE identifies a wider range of errors including validity, frame/coverage, nonresponse, measurement, processing and model errors. As such, TSE frameworks have sought to capture how a variety of errors accumulate throughout the statistical design and methodology, resulting in the final error of any given estimate. This approach has been adapted to report the quality of statistics which integrate administrative data (e.g. Reid, Zabala and Holmberg 2017, Rogers and Blackwell 2020). At the same time, the quality of statistics cannot be reduced to assessing error alone. When considering the integration of data from an administrative source into the census design, the impact of such integration on quality should be assessed in terms of the extent to which it adds error or uncertainty to the outputs, vis-à-vis the advantages of integration e.g. reducing response burden, increasing timeliness, reducing costs. As such, these Guidelines identify

[6] For example, the adjustments based on the post-enumeration survey carried out after the 1990 U.S. Census were subject to court proceedings and rejected by the U.S. Supreme Court and the 2000 adjustments were also subject to litigation (United States Census Bureau 2009).

[7] Based on Zhang (2012), In relation to input data, measurement errors relate to differences between supplied and target characteristics (e.g. gender, sex, age, ethnicity, occupation etc.) and include several types of error within variables including relevance (definition misalignment), mapping (errors in the re-classified measures due to poor equivalence between supplied and target classifications which may therefore require adjustments, e.g. through imputation) and comparability errors (errors between the re-classified and adjusted measures). Representation errors relate to the difference between the units supplied and the target units. They include errors relating to over and under-coverage (lack of alignment with target population), identification (errors in classifying a unit based on inconsistencies across multiple sources) and unit errors (errors in the statistical creation of statistical units of interest where they do not exist in any available data source).

additional dimensions which can affect the overall quality of census outputs including the institutional environment and the need to balance quality dimensions in order to meet user needs.

81. Following these Guidelines will help ensure that census estimates are based on the most appropriate sources and methods and are not misleading. At the same time, consideration should also be given to the way administrative sources are intended to be used in the census design (see Chapter 2). Given the variety of possible uses, this framework should be used flexibly and adapted to the level of quality required by different uses of administrative data by the NSO and different statistical requirements from the users of census statistics including the generally public, organizations, local and national governments. Inevitably therefore, quality assessment relies on skilled professional judgement throughout the entire statistical production process, from collection to publication, in order to meet the needs of users.

3.2 Measuring Quality

82. The quality of census estimates produced using administrative sources is particularly difficult to assess and/or measure due to the complexity and multi-dimensionality of the data used. As noted above, many factors affecting quality are not quantitatively measurable. Moreover, what constitutes 'fitness for purpose' and high-quality statistics will necessarily vary from one user to another e.g. some users may prioritize timeliness over accuracy. As such, it is important to assess/measure administrative data quality across the key dimensions which will be of interest to statistics producers and users. As such, what is meant by assessment and measurement is needs further clarification.

83. These Guidelines distinguish between *assessing quality*, meaning a qualitative evaluation, and *measuring quality* – meaning attaching a quantitative metric to this evaluation of quality. Where it is not possible to produce indicators for quantitative measurement, or where they have not yet been developed, these Guidelines recommend a qualitative assessment of their impact on quality. In addition to these, there are several additional principles which guide the production of official statistics (UNECE 1992) and which are applicable throughout the full statistical process and the wider NSO environment (e.g. commitment to quality, independence, data protection, statistical confidentiality etc.). These themes are relevant for all statistical processes and are not fully covered within the scope of the present Guidelines. However, it must be acknowledged that a census that uses administrative sources relies on data that were produced outside of the statistical system, in a different organization over which the NSO usually has no control.[8] For this reason, the impact of using these outside sources on these principles, must be considered carefully.

[8] In some cases, the NSO has some control over the register. In Switzerland for example, the Federal Register of Buildings and Dwellings or the Enterprise Register are sections within the Federal Statistical Office. Therefore, it might be feasible in a long-term perspective to integrate certain suitable registers within NSOs. Implications/advantages of this are briefly discussed in section 4.2.5.

3.3 Stages of Quality Assessment

84. To ensure these Guidelines are easy to follow, the quality assessment of administrative sources is considered across four broad stages of the census lifecycle. These are applicable regardless of census type (see Chapter 2). While statistical design is never entirely linear, thinking of how to carry out quality assessment in this way should enable statistical producers to quickly identify the key quality considerations which are most relevant to their own circumstances. The Stages are:

- **Source Stage**: A metadata-based quality assessment of new or re-supplied administrative sources to be used in the census. This Stage does not require NSOs to be in possession of the actual data, but it is crucial for the Stages that follow.
- **Data Stage:** The quality assessment of the raw administrative data supplied to NSOs by administrative authorities. This will require NSOs to validate the data supplied against the learning from the Source Stage. As well as basic validation, this Stage includes any processing required to establish the quality of the data supplied vis-à-vis what was expected, as well as comparisons with alternative sources.
- **Process Stage:** The processes often carried out on administrative data sources, in the context of censuses to transform the data for use in the census and/or to improve quality. The processes identified include data linkage; constructing statistical registers and the 'signs-of-life' methodology; enumeration using administrative data; methods for comparing the quality of variables across sources; and editing and imputation.
- **Output Stage:** The overall quality assessment of the census outputs produced using administrative data. While this is not conceptually that different from the assessment of the outputs of a traditional census, these Guidelines attempt to identify where this may differ.

85. These Guidelines are focused primarily on so-called 'input quality' of administrative sources and thus the Source and Data Stages. However, Process and Output quality are included for completeness and because ultimately, the question of whether or not the administrative data are good enough for census purposes can only be answered with respect to the planned use to which they will be put, or the census output they integrate. As such, the four Stages cannot meaningfully be separated. For the first two Stages, the Guidelines identify in detail the key data quality dimensions for assessment, the key tools used in completing their assessment and where possible, set out the criteria against which the assessment may be carried out. In addition, key issues in the assessment of process and output quality when census estimates are produced using administrative data are briefly reviewed. Across each of these Stages, areas for future guidance are identified where applicable.

3.4 Quality Dimensions

86. As previously noted, the quality of statistics and of administrative data is understood to encompass multiple dimensions which are not reducible to representation or

measurement errors e.g. statistics which are accurate but out of date, are of limited use. The quality dimensions identified by ESS include 1) relevance, 2) accuracy and reliability, 3) timeliness and punctuality, 4) accessibility and clarity, and 5) coherence and comparability.[9] However, for assessment of administrative data these "standard quality dimensions are not always applicable" (Daas et al 2008, p.2). On the other hand, they do capture all of the relevant aspects of administrative data quality. The following tables set out the dimensions for assessment of administrative sources used in these Guidelines, for each of the above-mentioned Stages.

[9] Alternative dimensions are used by various NSOs (e.g. Statistics Canada 2017, Australian Bureau of Statistics 2009). On the whole, these alternative frameworks cover approximately the same content albeit using different terminology or classifications.

Table 1: Quality dimensions at Source Stage

	Quality Dimension	Definition
SOURCE STAGE	Relevance and Accuracy	The degree to which the administrative data source meets the needs of the census. Covering the overlap between the census target population, concepts and definitions (relevance). The degree to which the data correctly describe the phenomenon they were designed to measure (accuracy).
	Timeliness	The lapse between the end of the reference period to which the information pertains and the date on which the information becomes available to the NSO.
	Coherence and Comparability	The degree to which the administrative source can be successfully combined with other sources used in the census, including linkability.
	Accessibility	The ease in which the NSO can obtain the administrative data, covering the impact of any restrictions, public acceptability of the use, the ease of data transfer and receipt, and the availability of metadata.
	The Institutional Environment	Organizational factors affecting the data supplier's capacity to supply data to the quality expected. Covering the strength of the relationship, previous experience, existence of formal agreements, risks associated with the status of the supplier and the supplier's quality standards.

Table 2: Quality dimensions at Data Stage

	QUALITY DIMENSION	DEFINITION
DATA STAGE	Validation and Harmonization	The data files provided to the NSO are in a readable format. Further data validation and harmonization arrangements are in place upon data transfer to NSO, to confirm that the expected variables/units/reference period have been supplied and ensure data processing by the NSO is consistent across census use cases.
	Accuracy and Reliability	The accuracy, completeness (for variables and population coverage) and coherence of the data supplied matches the requirements of the specific census use-case for which it will be used. Comparisons with alternative sources reveal acceptable levels measurement or representative errors.
	Timeliness and Punctuality	The timeliness and punctuality of the data supplied matches the requirements of the specific census use-case for which it will be used.
	Linkability	Adequate linkage variables are available (i.e. either common unique identifiers or a combination of variables which enable identification) and these are of sufficient quality to enable data linkage.

Table 3: Quality dimensions at Process Stage

	Quality Dimension	Definition
PROCESS STAGE	Accuracy of record linkage	Where multiple sources are linked (to each other or census responses), the linkage is accurate and unbiased, thereby improving the overall quality of the census methodology and/or dataset.
	Coverage and coherence of statistical registers and admin-based enumerations	Where census (sub-)population registers are constructed, or when admin data are used to supplement census collection, they adequately cover the target population/variables, thereby improving the overall quality of the census methodology and/or dataset.
	Accuracy of conflict resolution	Where different sources are linked and the same attributes are available in them, methods for deciding between sources are improving the overall quality of the census methodology and/or dataset.
	Accuracy of editing and imputation	Where census variables/units are derived/constructed through imputation or modelling techniques, this derivation is accurate and unbiased, thereby improving the overall quality of the census methodology and/or dataset.

Table 4: Quality dimensions at Output Stage

	Quality Dimension	Definition
OUTPUT STAGE	Relevance	The degree to which statistical outputs meet current and potential user needs
	Accuracy & Reliability	The closeness between an estimated result and the unknown true value – and how reliable these are over time and geographies.
	Timeliness & Punctuality	The lapse of time between publication and the period to which the data refer, and the time lag between actual and planned publication dates.
	Accessibility & Clarity	Accessibility and clarity refer to the actions taken in order to help the user find and understand the data he or she is interested in
	Coherence & Comparability	The degree to which data can be compared over time and domain. The degree to which data that are derived from different sources or methods, but which refer to the same phenomenon, are similar.

Source: Eurostat 2013 and 2018

3.5 Feasibility Research

87. It is unlikely that new administrative data sources will be integrated into census production without prior feasibility research by NSOs. The quality of a data source may be established by acquiring 'test data' and assessing its quality at the various stages suggested in these Guidelines. This will aid design thinking, i.e. designing a census methodology that makes the most of the available administrative data and considers the impact of its use on the quality of the census overall.

88. Firstly, feasibility research involves developing a detailed understanding of the administrative authority's data collection processes, the population covered, and variables included within the source as well as how accessible this data is (the Source Stage, Chapter 4). Secondly, supply, acquisition and ingestion of test data should be rehearsed, and test data examined in detail to identify quality issues and define cleaning and harmonization, along with validation checks (the Data Stage, Chapter 5). When data from multiple registers are combined, they can be used for verifying data quality on the one hand, and for selecting the most reliable variables and values, in accordance with the developed methodological rules, on the other hand (Chapter 6). Finally, estimates produced using test data can be compared with previous census estimates or another such 'gold standard', contributing to an assessment of the overall quality of the output (Chapter 7).

89. Generally, census characteristics cannot be acquired directly from administrative data sources, because they have been designed for other, non-statistical purposes and thus most of the definitions and classifications used by administrative authorities are different from standard statistical definitions. As such, data from multiple registers may be used in order to construct or derive certain census characteristics, while other characteristics may be covered by duplicate information in several registers. This makes feasibility research a key stage for developing methods for the derivation of census characteristics.

90. Census methodologists should address the following main challenges when deriving census characteristics:

1. Ascertain the international census standard (definition, classification, etc.) applicable to the target census characteristic;
2. Compare and contrast census definitions and classifications with the definitions and classifications used in the administrative source;
3. Test the accuracy of the administrative data recorded against alternative sources and work collaboratively with data suppliers to eliminate/mitigate any shortcomings;
4. Determine which and how many sources are required to derive and quality assure each target census characteristic;
5. Establish optimal rules for deriving each census characteristic and develop the necessary data processing software, optimised for the quality of outputs sought;
6. Where characteristics are not covered by any administrative sources, take steps to ensure creation of the necessary register or register part (e.g. suggest amendments in register procedures, the legal environment, etc.).

Box 1: Feasibility research in Estonia

In 2016, a pilot Population and Housing Census (PHC) was conducted in Estonia. Data for the mandatory census variable "Year of arrival in the country", was available in the country's administrative population register. Following an analysis of distributions however, the variable in the register was not directly used within the census, as in the first half of the 1990s (when the register was first established), the years of 1994 or 1995 were recorded as the year of arrival in the country for many persons. Comparing the distributions of year of arrival in the register to alternative migration data sources, immigration in Estonia in the 1990s should not be that large. To address this issue, data of PHC2011 and different population register variables (e.g. entry creation date and country of birth) were used, so that the created census variable could correspond as closely as possible to the definition in the UN Principles and Recommendations for Population and Housing Censuses (2008, Revision 2).

Box 2: Feasibility research in Israel

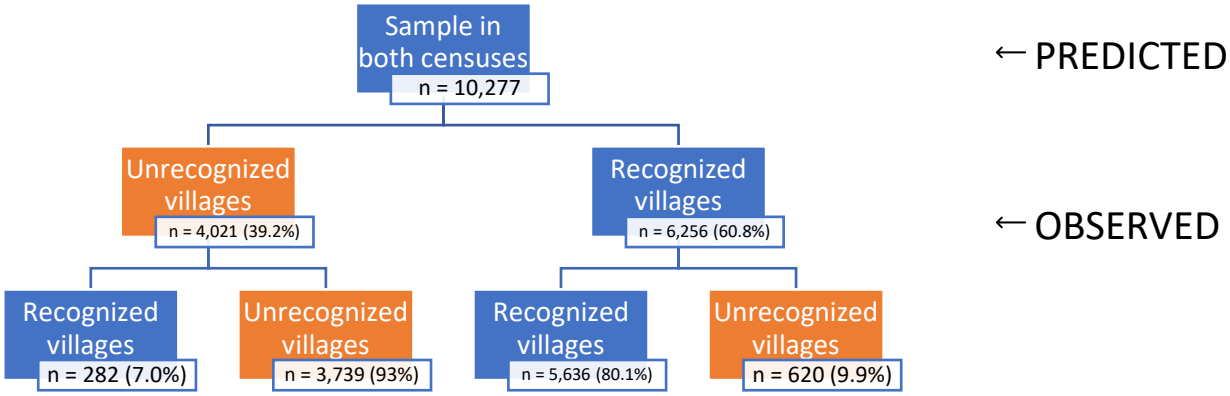
In Israel, feasibility research has been undertaken to develop methods for choosing the best address for the hard-to-reach Negev Bedouin population, by comparing estimates produced using administrative files, to those produced with the last traditional census in 1995. The Negev Bedouin is an ethnic group that includes approximately 283 thousand Arab Muslims, living in the Negev desert. They are a unique population as they traditionally live as nomadic tribes with a unique culture (e.g. 16 per cent of men are polygamous). In the traditional 1995 census, Bedouin households were interviewed, and their places of residence marked on maps. However, this population is considered hard-to-reach as about one third of this population lives in unrecognized villages, which are not connected to public infrastructure like electricity, water or paved roads. In addition, they have low levels of engagement with government agencies.

As such, research was conducted to explore the potential of administrative data in determining the geographic location for this population, based on the CPR (Central population register). Within it, every person has a unique PIN, which is linked to individuals' day-to-day interactions with government agencies and services. Additionally, each CPR record contains a link to records of the person's father, mother and spouse and demographic variables.

However, by comparing the CPR to previous census data, it is known that this register carries inherent errors and incompatibilities with census definitions, including omission of residents (foreigners), inclusion of non-residents (emigrants) and purposely incorrect address registration: 20 per cent of the population doesn't report their most recent address. Furthermore, there are limitations particular to the Negev Bedouin population as although Bedouins in the Negev desert are expected to be registered in the CPR, their address registration does not allow for accurate location. This is especially true for individuals living in unrecognized villages, registered under tribe names rather than the geographic area they live in, because the tribes could be scattered over all the geographic area of the Negev desert. Moreover, Bedouins of the unrecognized villages intentionally register themselves in the CPR as if they live in recognized villages in order to services such as educating their children in schools at the recognized villages. Finally, even Bedouins, who have a "real address" in one of the recognized villages may not be recorded with sufficient accuracy.

In this research, the first stage (initial location) was to use current CPR address and their 1995 CPR address, in order to locate individuals at the reference day. For example, if their address has not been changed in the CPR between the years 1995-2019, it means that they still live in the same geographic area reported in the 1995 census with their offspring. The second stage was to use a 'signs of life' methodology (see Chapter 6) based on other administrative sources (e.g. marriage records, address changes, local tax, water service, students at school and electric service), in order to improve the accuracy of location data. The results were then compared to those obtained via the traditional 1995 census. This method was then tested and re-evaluated by replicating the methodology with 2008 census data (Figure 1) and it was found that approximately 90 per cent of the sample found to be living at the same geographic area that was predicted via this method. This result was further confirmed through a small field test (n = 110).

Figure 1: Results predicted through administrative method (level 1) versus observed (level 2) on 2008 census in Israel



91. The above-mentioned challenges are best addressed through feasibility research as in the examples from Estonia (Box 1) and Israel (Box 2). The second example in particular highlights both the challenges and opportunities administrative data can present in the production of statistics about hard-to-reach populations. However, reaching an adequate harmonization of register and census concepts can be a complex and time-consuming activity which should not be underestimated. As such, it is recommended that the inclusion of administrative data sources into census production should be preceded by adequately resourced feasibility research which provides a 'proof of concept' for the planned integration of administrative data into the census production. In addition, making the four Stages of quality assessment an integral part of feasibility research will enabled census producers to directly apply the learning from feasibility research to the census production context and to better inform users on the quality of data sources.

92. Drawing on a review of the literature and the experience of census producers, the remainder of these Guidelines focus on tools and indicators for assessing the quality of administrative data sources, against each of these dimensions. In the chapters that follow, in addition to the work of Daas and colleagues, these Guidelines also draw on comprehensive suites of quality indicators for administrative data developed by others (e.g. Iwig et. al. 2013, Eurostat ESSnet MIAD 2014; ESSnet KOMUSO 2016, 2019).

Chapter 4. Source Stage

93. This chapter provides a guide to the key quality dimensions, the process of assessment, and associated tools and indicators for evaluating the quality of administrative data sources to be used in the census production – both at first acquisition and when they are regularly re-supplied to the NSO. Normally, no data are accessible during this stage. However, the search for information about the administrative data sources begins, most likely through communications and exploratory meetings between the NSO and the administrative data source supplier.

94. The evaluation in this stage should lead to a recommendation on whether to proceed with the acquisition initiative (or continue the re-supplying of the data source). If the decision is to go ahead, the administrative data supplied will undergo a more detailed evaluation at the Data Stage.

95. It is necessary to assess source quality both at the first acquisition of an administrative data source and in each instance when it is re-supplied to the NSO. This is because the characteristics of any data set that has previously been supplied might have suffered changes in terms of concepts, classification, collection methods etc.

4.1 Source quality dimensions

96. The data quality dimensions to be considered at this stage of quality assessment are Relevance; Timeliness; Coherence and Comparability; Accessibility and Interpretability; and the Institutional Environment. The dimensions are described below, with the processes, tools and indicators for assessment provided in the sections that follow. It should be noted that failure to reach minimum acceptable quality against any of the dimensions cannot be compensated by success in the other dimensions.

4.1.1 Relevance and Accuracy

97. Relevance reflects the degree to which an administrative data source meets the needs of the NSO with respect to the intended use. To be deemed relevant, the administrative data source must, for example, fulfil the reasons for its acquisition. This could be with respect to reduced costs or respondent burden; improvements to the quality of census outputs; or through the delivery of enhanced or new census outputs. To achieve this, the administrative source should be representative of the population of interest for the census (the target population) and the measurements from the population should align with the needs of the census. A key part of the assessment of relevance is understanding the context in which the administrative data have been collected.

98. As part of the assessment of relevance, the accuracy of the administrative data is also considered. Accuracy refers to the degree to which the data correctly describe the phenomenon they were designed to measure. It is important to understand how the collection, processing and quality assurance carried out by the administrative organization might affect the accuracy of the resulting data and their usefulness.

4.1.2 Timeliness

99. Timeliness refers to the period between the date to which the information pertains and the date on which the information becomes available to the NSO. The timeliness of the information will affect its relevance.

4.1.3 Coherence and Comparability

100. Coherence reflects the degree to which the administrative data can be successfully combined with data from other sources used by the NSO, i.e. census data, within a broad analytical framework, over time. The use of standard concepts, classifications and target populations promotes coherence within and between censuses. Therefore, a clear understanding of the operational definitions used by the administrative supplier, the purpose of data collection and the impact on comparability of changes in an administrative source over time and across domains factors in assessing coherence.

101. It is often a requirement to link an administrative source at the level of the census statistical unit to integrate the data into the census design. The comparability of identifiers across the different data sources to be linked is therefore a consideration under coherence.

4.1.4 Accessibility and Interpretability

102. Accessibility refers to the ease with which the NSO can obtain (and understand) the relevant administrative data items in their entirety. This includes an understanding of any restrictions (legal and those imposed by the supplier); public acceptability; the ease of data transfer and receipt (suitability of the form or medium for transferring data and costs); and the availability and clarity of documentation and metadata. It is crucial that the use of the administrative data source is based on a legal framework that gives the NSO the unequivocal right to access and use the data and the metadata for statistical purposes.

4.1.5 The Institutional Environment

103. The Institutional Environment refers to the organizational or institutional factors that may have an impact on the data supplier's capacity to supply data to the quality expected and to the agreed timetable (punctuality). This includes the strength of the relationship with the data supplier, including the effectiveness of communication channels and how responsive the supplier is to the NSO's requests. It includes the existence (or potential for) formal agreements and risks associated with the status and complexity of the supplier organization. It also includes the quality standards and procedures adopted by the administrative organization(s).

4.2 Tools and indicators

104. The quality of an administrative source should be assessed against the quality dimensions outlined in the section above. The following provides guidance on the process of assessment, including tools and indicators for evaluating an administrative source for use in the census.

4.2.1 Relevance and Accuracy

105. An understanding of the differences between the administrative population and the required census population, and between the measures/variables in the administrative source and the required census characteristics is important to assess relevance and accuracy. The error arising from these differences is referred to as representation and measurement error respectively (Zhang, 2012). At the Source Stage of assessment, it is possible to gain some understanding of these errors and their impact on relevance (as outlined in the subsections below) based on metadata about the supplier's purpose and methods of data collection. The impact of representation and measurement errors on accuracy and reliability are also considered at the Data and Process Stages (Chapter 5 and Chapter 6).

4.2.1.1 The Census target population (representation)

106. To assess relevance, the NSO must determine whether the set of objects in an administrative data source align with the population units of interest for the census (the target population). An object is the basic element of the population on which information is collected, this could be a person, household, dwelling, event or transactions, etc. The following indicators are proposed for establishing relevance, with respect to representation. Against each indicator is a series of questions to help guide the assessment:

- Alignment (of the objects) with the census target units.
    i. How comparable are the administrative objects with the census target units?
    ii. What definitions, methods and processes are used to identify and include an object in the source?
    iii. Are there any laws or regulations that define the objects?
    iv. Are any checks carried out by the data holder to ensure the definitions hold?
    v. In the case of misalignment with the census units, is a transformation possible that could meet the census needs?
- Coverage (of the set of objects) against the census target population.
    i. Does the coverage of the objects meet the needs of the census?
    ii. Is there evidence of under-coverage (objects that are missing from the source, but are part of the census target population) and over-coverage (objects that are in the source, but are not part of the census target population) that would impact the usefulness of the source?
    iii. Are there any differences across geographical areas due to differences in practices by the data holder or due to legislation that need to be considered?
    iv. Are there any rules, legislative or regulatory requirements, including penalties for non-compliance that may impact on the inclusion and exclusion of objects on the source?

v. What methods and processes are adopted by the data holder to include new objects that meet the required inclusion criteria / definitions (e.g. registration procedures) and to remove objects that no longer align with the target population for the administrative source (e.g. deregistration procedures)?

vi. In the case of coverage errors, are there other data sources that could be used in combination with the source to overcome, for example, under or over-coverage in the source?

4.2.1.2 The census variables/Characteristics (measurement)

107. To assess relevance, the NSO must also determine whether the information collected from the objects on an administrative data source meets the needs of the census, with respect to the target concepts (e.g. ethnicity, employment status, household size, tenure status, etc.). The following indicators are proposed for establishing relevance, with respect to measurement:

- Availability of the target variables/characteristics.
    i. Does the administrative source include the variables needed against the intended use for the census?
    ii. Do the variables / characteristics broadly cover the relevant census reference period?
- Alignment of variable concepts, definitions and classification with the census needs.
    i. Are the administrative concepts, definitions and classifications comparable with the census needs?
    ii. Is there a difference between the data holder's ideal target concepts and what is actually achieved through their operational target measure used in the collection?
    iii. In the case of misalignment with the census concepts, definitions and classifications, is a transformation possible to meet the census needs?
- Alignment/measurement error against the census reference period.
    i. What is the frequency of collection for a variable / characteristic?
    ii. Are there known delays between an event or phenomenon occurring and being captured in the administrative source (e.g. parents may not have to register a birth for several weeks on a country's birth register)?
    iii. Are there time stamps recorded on the data source to indicate what period a data item relates to?
    iv. Are there any incentives or disincentives for a data subject to update their information as and when their circumstances / information changes on the administrative source (e.g. benefits or penalties for not doing so / or doing so)?

- Quality of collection and potential for measurement error against the census concepts.
    i. What is the data holder's purpose for collecting the data and how might this influence the quality of the data?
    ii. Are there any legal obligations, targets or incentives (or lack of incentives) that could influence the quality of the data collection?
    iii. Does the data holder's collection process raise any concerns about the quality of the variables, including the potential for any biases? This could include whether data are recorded by proxy and therefore not reported directly by the data subject (increasing the potential for misreporting).
    iv. What procedures are in place to validate and check data on entry by the data holder?
    v. Are there any incentives or disincentives on the data subjects to provide complete and accurate information to the data holder?
- Quality of data processing and potential for processing error by the data holder.
    i. Does the processing carried out by the data holder suggest the quality of the resulting data will meet census needs?
    ii. What checks are carried out by the data holder to assure quality?
    iii. Are data edited or imputed? If so, when and how, and is there an indicator on the data source to identify when an edit and imputation has taken place?
    iv. Are there any rules, regulations or incentives on the data holder that may impact on the way data are processed?

108. At the Source Stage of assessment, the evaluation against the indicators is usually based on a qualitative assessment (e.g. indicating whether the need is fully met, partially met or not met against each indicator, with an explanation of why, based on the answers to the question set). A quantitative assessment of representation and measurement error is carried out in the Data Stage (based on analysis of the data) under the dimension of accuracy and reliability (Chapter 5).

109. The assessment against the indicators should inform a decision (often based on experience and expert judgement) on the use (or continued use) of a source in the census. The decision should take account of whether or not the data source can meet the needs of the census (e.g. reductions in costs and respondent burden, improvement and enhancements to the census outputs), set against any costs or risks (referenced under the Institutional Environment and Accessibility dimensions below).

110. There are various quality frameworks described in the literature that provide similar indicators as in this chapter against the different dimensions of quality, along with question sets and scoring systems for informing the assessment (e.g. Daas et. al. 2009; SN-MAID, 2014; Iwig et. al. 2013; Statistics Canada's Administrative Data Evaluation Guide (Lavigne & Nadeau

2014); Statistics Austria's Quality Assessment of Administrative Data, Documentation of Methods Framework (Statistics Austria 2019)). The New Zealand case study (section 4.4.1) provides a practical example of a framework being used to assess administrative sources for use in the census.

4.2.2 Timeliness

111. An administrative source may cover the relevant time period for the census, but to be useful it will also need to be available in time, against the schedule for the census. The following indicator is proposed to assess timeliness:

- Timeliness and frequency of supply against census needs.
    i. What is the lag between an event or phenomenon occurring and being captured in the administrative source?
    ii. What is the lag between the end of the reference period for the administrative data and the date the data can be made available to the NSO?
    iii. How frequently can the data be supplied to the NSO, set against the needs of the census?
    iv. Are there any requirements, in terms of the delivery method and required formats and data structures the NSO uses that could impact on the data supplier's timeline?
    v. Can the NSO structure and process the data in time for the needs of the Census, from when the data becomes available?

112. In instances where the data are unlikely to be available in time, the NSO may wish to establish whether a provisional version of the dataset can be made available ahead of schedule. In such cases, the dataset may be incomplete and subject to higher levels of error. There may therefore be a trade-off to consider between the timeliness of the data and accuracy.

113. As referenced against the dimension of Institutional Environment below, it is important to include the delivery dates against the data reference periods within formal agreements with the data supplier. Although the data may be available on time to meet the requirements of the data owner, they may not necessarily be delivered to the NSO in time, while the latter carries formal responsibility for timely delivery of the census.

4.2.3 Coherence and Comparability

114. It is important to assess the degree to which an administrative source can be successfully combined with other data sources for use in the census. The information gathered under the indicators provided to assess relevance can be used to assess coherence. This includes information about the differences between the underlying concepts, definitions, classifications and methods between the administrative data source and the other data sources for combined use within the census.

115. For a full register-based census, it is important to analyze the census characteristics and administrative data source; mapping and ascertaining the extent to which the information in the administrative data source facilitates the derivation of the relevant census characteristics. In particular, the NSO should establish whether or not the data recorded in the registers conform to the definition of the census characteristics. In the case of partial or no conformity, the NSO should examine the causes of non-conformity between the census characteristics and the information available in the administrative data source.

4.2.3.1 Comparability

116. Administrative data are subject to changes and difference over time and across domains due to changes in legislation, regulation and procedures, which can affect the concepts, definitions, classifications and coverage of a source. More generally, the changes can impact on all of the indicators under representation and measurement, as outlined under the dimension of relevance. This is of particular importance for the census, where stability over time can be a key concern. The following indicator is proposed to assess comparability:

- Comparability over time and domains

  Are there any changes across time or differences across domains (e.g. geographical areas) affecting the:

i. the definition of an object and coverage of the objects on the administrative source relevant to the census?

ii. the concepts, definitions and classifications associated with the variables on the administrative source of relevance to census?

iii. the data collection, processing and quality assurance procedures that could impact on the quality of source for census purposes?

4.2.3.2 Linkability

117. A consideration under coherence and comparability is the ease with which an administrative data source can be linked with other relevant datasets for the census. The Estonian case study (section 4.4.3) provides an example of how it is possible to link multiple different administrative data sources with few different unique IDs. The following indicators are proposed to access the linkability of a source:

- Presence of a unique key for linkage

  i. Does the source include a unique identifier that is common with the unique keys required for the census linkage?

  ii. Is the identifier available for all of the relevant objects on the source?

- Presence of a unique combination of variables for linkage

  i. Does the source include a unique combination of variables (e.g. name, age and address), which could be used for the census linkage?

  ii. Are the unique combination of variables present for every object on the source?

118. The quality of linkage variables is also assessed at the Data Stage (Chapter 5) and the quality of the linkage process is covered as part of the Process Stage (Chapter 6).

4.2.4 Accessibility

119. The following indicators are proposed for the assessment against the accessibility dimension:

- Restrictions on data access and use
- Public acceptability
- Easy of data transfer and receipts
- Interpretability of the source – clear and comprehensive metadata.

120. The sections below provide details of the relevant information for assessment against each of the indicators.

4.2.4.1 Restrictions on data access and use

121. It is important to identify any restrictions that may impact on the NSO's ability to access and use an administrative source. For example, existing data protection restrictions embedded in legal acts can impose certain limitations on the data acquisition and processing (especially when data are protected with extra security measures or laws at state level). Such legal acts may be specific to particular data sources (see Estonian case study in section 4.4.3, for example) or may be more generic allowing the NSO access to such data sources as and when required, subject to the agreement of the data owner. The data owner may also impose further restrictions on the supply of data and the permitted use. These can include:

- suppression of records or variables;
- disclosure treatments (pre-delivery), such as encryption of identifiers, perturbation, banding or top-coding of the supplied data;
- restrictions on how the data can be used;
- restrictions on the retention of data and rules on deletion / disposal;
- rules on disclosure methods that must be applied by the NSO, affecting the census outputs.

122. The NSO should establish and describe any restrictions that apply, on which an assessment can be made as to the impact (and risks) of the restrictions on the use of an administrative source in the census. As part of the assessment, the NSO should also consider whether it has the capability to abide by the restrictions. This could include the technical and procedural safeguards the NSO must adopt. The safeguards would generally form part of a Memorandum of Understanding (MoU) or Data Security Agreement with the data owner / supplier. In particular, the MOU may describe how Personally Identifiable Information (PII) will be protected.

4.2.4.2 Public acceptability

123. Whether an NSO can access a data source for use in the census may also depend on public acceptance. The public must understand and be supportive of, or at least not hostile to new approaches and uses of their information. If the public are opposed to the use of an administrative data source, there is a risk to quality. For example, this could change the way the public interact with the census or an administrative source used in the census. The NSO should therefore be transparent about the use of administrative sources in the census, highlighting the benefits to the public, whilst providing assurance against privacy and security.

124. To assess public acceptability, the following tools or processes can be used:

- Public consultation or engagement
- a Privacy Impact Assessment (PIA)
- a Data Ethics Assessment

125. A ***Public Consultation or Engagement*** exercise may be carried out by the NSO on the use of administrative data in the census (or for other statistical research or outputs). This can take various forms, including formal consultations, questionnaires (through surveys or via the NSO inviting feedback via its website), qualitative research into public attitudes, or the use of Citizens' Panels. Citizens' Panels aim to bring together members of the public (to be representative of the population, or to reflect different population groups of interest) to assess their views and opinions.

126. A ***Privacy Impact Assessment (PIA)*** is a formal process resulting in a document that describes the process, findings and results that helps the NSO to consider the effects of a new programme or service (or proposed policies and plans) on the privacy of individuals. As a risk management tool, used in the planning phase of a programme or service initiative, PIAs assist organizations to more fully consider the privacy implications of a given proposal. PIAs are also used to ensure data controllers can meet their obligations under the General Data Protection Regulation (under European Law). A PIA can be applied to the various usages an NSO may wish to make of a data source in the design of the census. The New Zealand case study in section 4.4.2 describes the privacy risks involved and the mitigation measures used by the NSO to eliminate or reduce each of the risks.

127. A ***Data Ethics Assessment*** is carried out to establish whether the access, use and sharing of public data, for research and statistical purposes, is ethical and for the public good. NSOs may use an ethics self-assessment tool (e.g. UKSA 2020), but may also use a formal body to provide expert advice or endorsement, such as a Data Ethics Advisory Committee (e.g. the UK National Statistician's Data Ethics Advisory Committee[10]).

[10] For more information see https://uksa.statisticsauthority.gov.uk/about-the-authority/committees/national-statisticians-data-ethics-advisory-committee/#:~:text=The%20National%20Statistician%E2%80%99s%20Data%20Ethics%20Advisory%20Committee%20%28NSDEC%29,advise%20the%20National%20Statistician%20on%20the%20ethical%20

128. The findings from public consultation and engagement, PIA and ethics assessments, can help an NSO assess public acceptability of the use of administrative sources in the census (and for other NSO statistics).

Box 3: Statistics Canada's Trust Centre

Statistics Canada has a Trust Centre that outlines how information is protected, placing privacy as a top concern. This includes how societal needs for new data insights and the protection of privacy are balanced, using a modern "necessity and proportionality" framework. The Trust Centre provides clear and comprehensive information to reassure the public on the use of their data, including through the use of infographics and short videos accessible via the website. One such video ('Joe Anonymous') explains how the public's data are used, including the importance of bringing together multiple sources. There is an emphasis on the work and culture within Statistics Canada to protect data, including a promise to protect the identify of people, their families and their businesses.

Openness and transparency are at the heart of the Trust Centre and information about administrative sources to be acquired and used by Statistics Canada is published (and updated regularly) on the website. [Available at https://www.statcan.gc.ca/eng/trust]

4.2.4.3 Ease of data transfer and receipt

129. The Data Supplier might use very different data models, formats, schemas, software and hardware to what the NSO is familiar with. This includes how data are held and transmitted, (including the security arrangements for transmission). The data structures could also be complex and file sizes extremely large (particularly for transaction data). It is important that the NSO understands such differences and complexities, in order to assess whether it is feasible to receive and ingest the datasets into the NSO's systems. This process can also include negotiations with the supplier on the development of processes and systems to facilitate the transmission of datasets in a format that meets the needs of the NSO. However, this can be a time consuming and costly process.

130. More generally, cost is a key factor to be considered when assessing ease of access. This can include costs imposed by the data supplier, or costs incurred by the NSO in developing its capability to receive an administrative dataset (for example, if new software or hardware needs to be purchased). It is important to assess any costs against the expected value a new administrative source will bring.

131. In practice, details of the arrangements for the transmission of data to the NSO, including the files structures and format (e.g. flat files, a relational database; SAS, Excel or text formats, etc), the variables, the frequency of supply and dates for delivery, data standards and agreed costs, would be included in Data Sharing or Delivery Agreements between the NSO and the supplier. Such agreements would be signed by authorized managers at each of the organizations.

4.2.4.4 Interpretability of the source – clear and comprehensive metadata

132. An assessment of interpretability relates to the existence and availability of comprehensive and clear metadata and documentation about the administrative source. Without this, it is not possible to understand and assess the administrative source against the intended use. The metadata should include details about the:

- administrative organization and the purpose of the collection
- concepts, definitions, classifications and protocols used
- collection, processing, validation and quality assurance methods and procedures
- reporting units and variables; including data dictionaries, file structures, formats and relationships within the data.

133. This information is important for the assessment against the other quality dimensions outlined in this chapter. It will often be the case that clear and complete metadata will not exist for all aspect of an administrative source in the initial phase of exploring the source for use by the NSO. It is therefore necessary to work with the data supplier to build the relevant metadata. This relies on good communication with the supplier and a willingness of the supplier to work with the NSO (see Institutional Environment below). Depending on the complexity of an administrative source, an NSO may decide to set up secondments for staff to work within the administrative organization to develop an in-depth understanding of a source.

Box 4: Metadata templates for assessing administrative sources

New Zealand

Statistics New Zealand has a Guide to Reporting on Administrative Data Quality (Stats NZ 2016), with an associated Metadata Information template for Admin Data (available at Stats NZ 2020). The template is a useful resource for capturing metadata about an administrative source, covering general information about the administrative organization, the data collection, population objects and variables, changes over time and aspects of accessibility.

The Netherlands

Statistics Netherlands Checklist for the Quality Evaluation of Administrative Sources (Daas et al 2009), provides a useful template (the Annex to the paper) for recording information and metadata about a source. The ordering of the dimensions and indicators within the template directs the user through the recording and assessment against the metadata efficiently – ensuring problems are revealed early on before moving on to later stages.

The Statistical Network on the Methodologies for an Integrated use of Administrative Data (SN-MAID) Project

Deliverable B2.3 (Source) and B2.4 (Metadata) (SN MAID, 2014) provide checklists including quality indicators and fields for recording metadata about an administrative source, which is used to assess the quality of the source for use in statistics. The checklists draw on the work of Daas et al. (2009).

4.2.5 The institutional environment

134. The NSO is completely reliant on the data supplier to collect, process, and deliver the administrative data to the quality expected and to the agreed timetable. The NSO is also reliant on the quality of the information the supplier provides about the data (see Interpretability, section 4.2.4.4 above) and about any foreseen changes to the data. It is therefore important to assess confidence in the data supplier's ability to meet these needs. The following indicators are proposed for the assessment of the Institutional Environment:

- the strength of the relationship with the supplier
- previous experiences with the supplier
- the existence of formal agreements
- the risk posed by the status of the supplier
- the supplier's quality standards.

135. **The strength of the relationship.** There should be processes in place for managing the relationship with the data supplier; ensuring there is a continuous dialogue. These should include mechanisms for:

- the communication of the NSO's requirements to the supplier
- the timely communication (by the supplier) of any changes that might affect the data source (e.g. changes to the legal basis for the data, to concepts and classifications and to the processes and procedures for data collection, management and supply);
- raising any questions with the supplier about the data source
- feeding back findings to the supplier that could result in improvements to the source.

Box 5: A Quality Assurance Toolkit: Communication with data supply partners

The UK Statistics Authority's Administrative Data Quality Assurance Toolkit (UKSA 2015b) describes "practice areas" associated with data quality, including an area for Communication with data supply partners. The area covers the importance of collaborative relationships with data collectors, suppliers, IT specialists, policy and operational officials. Highlighting the importance of formal agreements detailing arrangements (see below), as well as regular engagement with all involved parties. There are three levels of assurance proposed, depending on importance: Basic, Enhanced and Comprehensive.

136. **Previous experiences.** This includes how responsive a supplier has been to the NSO's requests, whether any issues have arisen with previous supplies of data (e.g. late delivery,

unexpected errors), whether the supplier has provided accurate information in the past about a data source (this might have been established through checks at a later stage by the NSO).

137. **Formal agreements**. This includes whether written agreements (legal or otherwise) exist or can be developed, covering:

- roles and responsibilities of the NSO and supplier. This could include whether the NSO has a role in the approval of any changes to an administrative source used (or to be used) in the census
- the legal basis for the supply of data and any security/confidentiality requirements
- the specification of requirements, as per the Data Sharing/Delivery Agreement referenced under the Ease of data transfer and receipt section above.

Box 6: Statistics Netherlands System of Base Registers

In the Netherlands a system of administrative base registers is adopted, comprising 13 registers on population (residents and non-residents), addresses and buildings, enterprises, real estate (boundaries, ownership, value, etc.), topography (maps: land, water, roads), motor cars (model, colour, ownership, etc.), taxable income, labour (wages, employers, social benefits, etc.) and subsoil (sewerage, cables, etc.). The system of base registers is based on legislation and supports the production of statistics (including census) by Statistics Netherlands.

The use of data from base registers is compulsory for governmental agencies. The objective is that all users of the system contribute to the quality of the data. Therefore, users are obliged to notify the holders of base registers if they have alternative data that are considered to be of better quality (with the exception of the NSO, due to legal considerations). Users of base registers can rely on their validity. Statistics based on base registers demand only a limited amount of data editing. The registers adopt standardized approaches and identification numbers for linkage, so the statistical data are generally coherent.

Each base register has its own project board. All groups of stakeholders are represented in these project boards. Project boards operate within the legal framework and see to it that the register data fulfil the legal requirements (quality, completeness, etc.) and that the data are correctly applied. Project boards act as advisory boards to the responsible cabinet ministers and meet a few times per year.

138. **The status of the supplier**. The risk associated with the status of a supplier should be assessed by the NSO, considering whether the supplier is an established, stable and reputable organization. This should consider whether there is any legal or regulatory basis to the administrative function the supplier carries out that would give confidence in the sustainability and quality of the source. Risks associated with the complexity of the supplier

organization(s) involved in the collection, processing and delivery of the source should also be considered. For example, there could be multiple bodies or organizations involved, each impacting on the quality of the final data supply.

139. **Supplier's quality standards**. An assessment of whether the supplier can meet the quality expectations of the NSO should be made. This should consider the Information on the principles, standards and guidelines adopted by the supplier for assuring quality, including the procedure in place covering collection, processing and the supply of data to the NSO. Evidence of how the supplier checks whether the standards are being met is valuable, this could be through internal or external audits by regulators or professional bodies. The supplier may also produce quality reports, which should be reviewed by the NSO. A more detailed assessment based on key aspects of the administrative source is included under the relevance quality dimension above.

140. Following an assessment of the data supplier against the criteria outlined above, the NSO can evaluate the risks associated with the supplier delivering the administrative data on time, against the required quality criteria.

4.3 Recommendations

1. Identify relevant and promising administrative sources for use in the census (see Chapter 2).
2. Set out clearly the required target population, variables and concepts, along with the anticipated outcomes for using an administrative source in the census on which to base the assessment.
3. Understand the restrictions and challenges to acquiring and integrated administrative sources into the census, including where changes may be needed to the NSO's methods, processes and computing systems.
4. Build and maintain clear and comprehensive metadata capturing all relevant quality information about a source (this will provide a valuable resource for the NSO). Structured metadata using an appropriate, agreed-upon metadata standard format is important (Cornell University Research Data Management Service Group 2020).
5. Develop a good understanding of the data supplier, the context and purpose of the data collection and the quality standards they uphold.
6. Build strong relationships with the data supplier, to ensure effective sharing of information – building a common understanding of each other's needs.
7. Put in place formal agreements, which outline clearly the NSO and data supplier requirements, roles and responsibilities.
8. Carefully assess the value of acquiring and using an administrative source, against any risks and costs. This can be with respect to the stability of a source over time and the risk of a supplier failing to deliver data on time or to the quality expected.

9. Ensure there is a sound legal basis to the receipt and use of an administrative source, with effective safeguards in place to protect the privacy of the data subjects.
10. Be clear and transparent about the use of administrative data, showing evidence that the benefits outweigh any privacy concerns.
11. Accept that objects, definitions, concepts and time reference periods within an administrative source may not align with the census targets. It will, therefore, be necessary to transform data and make judgements on what levels of misalignment are acceptable.
12. Assess quality on a continuous basis (using the process and tools outlined) – responding to any anticipated or known changes to a source.
13. Document and publish the strengths and weaknesses associated with administrative sources used for the census, so that users have confidence in the use and can take account of any limitations.
14. Be prepared that it will take time to understand and acquire administrative data sources for use in the census. Particularly, where a programme of work is required to develop registers for use in the census (as per Estonia case study).

4.4 Case studies

4.4.1 New Zealand: Source assessment

141. In March 2012, the New Zealand Government agreed to a Census Transformation strategy. Part of the first phase of that programme was to complete a first broad look at the potential for administrative data to produce the long-form (social and economic) information currently provided by the census (O'Byrne, E, Bycroft, C & Gibb, S, 2014). This process identified administrative data sources related to the census topics and used quality measures to assess how likely those sources were to satisfy the information needs previously met by the traditional census. The investigation did not include population counts and demographic breakdowns which were investigated elsewhere.

142. The purpose of this work was to provide an early indication of the likely ability of existing administrative data sources to produce census long-form information and to guide decisions about where to direct more in-depth analysis.

143. The steps in the process included:

1. **Identification of data sources** – achieved through tapping into existing Stats NZ use of administrative sources, web searches, and contact with government agencies.
2. **Understanding the nature and content of potential administrative data sources** – achieved through review of publicly available information, discussions with experts from Stats NZ and the source agencies.
3. **Quality assessment** – using five critical quality dimensions.

4. **Assigning a quality rating** - the likelihood that administrative data could satisfy a census topic.

144. The quality measures used in the assessment were adapted from existing quality dimension frameworks (such as the Stats NZ quality model, Eurostat, 2009 and 2011). The five measures identified as relevant for this assessment were: relevance, accuracy of coverage, accuracy of linkage, timeliness, and accessibility. These quality measures were chosen because they are strongly discriminatory, in the sense that they are essential for the use of administrative data for census information and are also measures for which reasonable judgements can be made from metadata.

145. This assessment was done by jointly assessing as many administrative data sources as may be needed to satisfy that census variable. For each variable, each quality dimension was rated as Excellent, Good or Poor, which determined an overall rating of 'Likely', 'Possible' or 'Unlikely' to be satisfied by administrative data sources.

146. The key questions considered for each dimension are outlined in Table 5.

Table 5: Key questions for each dimension

QUALITY MEASURES	MAIN QUESTIONS FOR ASSESSMENT
Relevance	How close are the administrative data to the statistical concept? (the census topic is used as a proxy for the statistical concept) Who/what should be included in these data? (target population) Who/what is actually included in these data? (observed population)
Accuracy of coverage	Are there missing people or responses? (undercount) Are there duplicate records or other people who should not be included? (overcount)
Accuracy of linkage	Is it possible to link the data to the census population or dwelling lists?
Timeliness	How frequently are the data updated? How long after the reference date are the data available to Statistics NZ?
Accessibility	Are there privacy or legal issues around using these data? Are there any other barriers to access?

Source: Stats NZ

147. The study showed which administrative sources would be most important in providing census-type information, and detailed analysis of most of the variables identified as being 'possible' or 'likely' has now been completed. One of the most important findings was that a majority of the current census variables were unlikely to be obtained from administrative sources, and a survey component would still be needed.

148. The quality ratings used are shown in Table 6.

Table 6: Quality ratings

QUALITY MEASURES	DEFINITION OF QUALITY RATING		
	EXCELLENT	GOOD	POOR
Relevance	The data collected in the administrative sources are very close to the statistical concept.	The data collected in the administrative sources are not exactly the same as the statistical concept, but are close, or related to a similar statistical concept that might be acceptable.	The data collected in the administrative sources are not at all relevant to the statistical concept we are interested in.
Accuracy of coverage	The coverage (net, under and over) is similar to the census.	Most of the population is covered and those who are missing are 'missing at random'.	Coverage (net, under and over) is very low, or there is bias in the distribution of missing values.
Accuracy of linkage	Data have excellent individual identifiers that can link the units in one dataset to other datasets.	Data have good individual identifiers.	Data have no individual identifiers. Data linkage is not possible.
Timeliness	Data are updated at least every year and available to Statistics NZ within two years.	Data are updated at least every two years and available to Statistics NZ soon after.	Data are updated sporadically, or with delays of more than two years.
Accessibility	No privacy or legal concerns exist. Statistics NZ understands the data and has a good relationship with the administrative data owner.	Some privacy or legal concerns exist with one or more key datasets.	Serious privacy or legal concerns exist. No relationship with administrative owner or no history of using the data.

Source: Stats NZ

4.4.2 New Zealand: Privacy impact assessment

149. Privacy impact assessments (PIAs) are a useful tool when considering the accessibility dimension of quality, specifically the legal implications of administrative data use and for building public trust. In New Zealand, the Office of the Privacy commissioner provides guidelines and templates to support organizations completing PIAs. This guidance outlines 12

privacy principles (these principles are drawn from the Privacy Act, 1993 and range from collection of data to use of unique identifiers) to be considered as part of a PIA. It also includes guidance on the key questions to ask during the process, some of the common risks to be aware of, as well as possible mitigation strategies to consider. Prior to the New Zealand 2018 Census, Stats NZ engaged an external organization to complete an independent PIA on the planned use of administrative data in the census. Stats NZ later completed and published an additional PIA covering the intention to extend the use of administrative data to mitigate the lower than expected response rate. The overarching goal of a PIA in this context is to bring together information about what, why, and how a NSO wants to use specified administrative data, and to assess the potential value gained against a range of privacy considerations.

150. Key topics covered in the second edition of the 2018 Census PIA include:

- Information about the benefits of using administrative data in the census and detail about how security is managed through the process of constructing the final census dataset.
- A summary of relevant legislation
- A summary of the privacy assessment for each of the 12 privacy principles
- Recommendations and action plan to minimize harm
- A risk and mitigation table containing risk ratings (consequences and likelihood) for each of the 12 privacy principles along with some additional principles to reflect obligations under the Statistics Act 1975

151. The PIA concluded that using administrative data in the census is lawful, safe, and beneficial to New Zealanders.

4.4.3 Estonia: Improving data through legislation

152. Statistics Estonia (SE) carried out work during 2010-2013 in cooperation with data source owners and scientific communities. The goal was a quality assessment of administrative sources to be used in the census production. The requirements for those census characteristics laid down in the regulations of the Council of Europe and the European Parliament, as well as the regulations of the European Commission were analysed. The coverage of each census characteristic was mapped, and suggestions were made for the formation of census characteristics in future and for quality analysis (EU 763, 2008).

153. On the basis of this analysis it was concluded that as many as 20 different administrative sources (held by nine different authorities or ministries) would be necessary to provide data of sufficient scope and quality. SE was given a mandate to determine the minimum universal criteria for all those registers that were required to provide the data to meet the needs of users.

154. SE was made aware of the limitations in use of registers, the main cause of which was the lack of sufficient metadata information provided by register holders. The metadata that did exist had been compiled merely to satisfy the administrative purposes for which the data were collected and were often not relevant for the statistical use of the data. In particular

there were often conceptual inconsistencies between the definitions and classification adopted in the register and those necessary for use in the census. Nor was coverage of the base population or the availability of topic variables in the registers always compatible with national census requirements – particularly where variables related to self-defined statuses.

155. The target for 2014 was to work out a package of legal and organizational measures to improve the quality, timeliness and coverage of the dataset for the register-based census based on the bottlenecks pointed out in the methodological report.

156. From 2014 Statistics Estonia actively started to participate in formal deliberations with the relevant authorities with the aim of making the necessary revisions to the legal acts governing the specific data sources required for census purposes. State authorities were requested to state in their legislative proposals whether a new administrative data source was going to be established or an existing one modified. Any data sharing mode was also to be prescribed. Provisions for the scope to start or improve the data collection process were also covered in the legislation.

157. SE were charged with the responsibility for improvement of data quality in the registers. Accordingly, it devised a roadmap based on suggestions given by experts, and prepared an improved business model in order to facilitate better cooperation between administrative registers. SE worked out an action plan up to 2020, which comprised different tasks for data source owners. The most urgent of which was to create a legislative environment for adding any necessary new characteristics to the registers (such as occupation, industry and place of work) and for updating these characteristics in the registers (including the Tax Board registers, planned working register, business register, etc).

158. The next critical task was to improve the accuracy of residence registration to gain better coverage for households and institutional populations and tenants. For that SE initiated a State-level project, launched by the Ministry of Interior, for adding archival data on families and relationships between family members to the Population Register. This would improve several census characteristics (such as legal marital status and relationship within household).

159. Amendments to the legislation relating to foreigners has helped to improve data collection on the foreign-born population. This has allowed improvements to registration procedures in order to obtain more complete information on new arrivals (including characteristics on education, marital status, relationships between family members).

160. Altogether, about 20 different suggestions were made to data source holders to improve data source quality using the legislative framework.

161. In order to create linkable data some basic rules, prescribed by special governmental regulations, were adopted by 16 register holders from 2016:

- All data in registers for persons, enterprises and dwellings must be identified (using unique codes);
- Address data should be used in all registers according to the established standard; and
- Metadata should be available and updated.

162. Another important aspect related to the quality of the data source used concerned data transfer. It is necessary to have a fool-proof and reliable environment for transferring data from different registers to the NSO. In Estonia such an environment – named X-Road – facilitates the transfer of large quantities of data between institutions or the provision of individual persons with their personal data. Data capture for census purposes was allowed, according to the government regulation, through X-Road. Previously, data owners used to use e-mail or file transfer protocol (FTP) as encrypted .csv or .xls files.

163. The quality standard was prepared for assessing data sources. In the quality standard the numerical values were fixed for accepted biases in census variables and hypercubes, where the following quality dimensions of data were taken into consideration:

- Relevance (coverage, conceptual differences, etc)
- Timeliness & Periodicity (last data of record update, lags in supply, etc)
- Accuracy: especially of linkage variables to assess linkability of source.

164. By 2020 ES had reached the position whereby 38 different variables relating population and dwellings required by the current EU census programme had been derived from 26 different administrative sources. (Pilot Census Report, 2019).

Chapter 5. Data Stage

165. This chapter provides a guide to the key data quality dimensions, tools and processes for the assessment of administrative data at the Data Stage of production. This refers to the quality assessment (QA) of raw administrative data as supplied to the NSO, with reference to the expectations and requirements established through the metadata-based assessment at the Source Stage. The Source and Data Stages together provide an overall assessment of Input Quality, with respect to an administrative data source (see UNECE (2018), Chapter 6).

166. The quality of administrative data at the Data Stage is assessed against several dimensions including readability and validity, accuracy and reliability, timeliness and punctuality and the linkability of the source data. In what follows, these dimensions are explored in greater detail (section 5.1, along with the tools and indicators for their assessment or measurement (section 5.2).

167. At the Data Stage it is possible to establish a baseline for the quality of the individual administrative datasets supplied, on which edit and validation rules can be developed and resupplies of data can be assessed against. At this Stage a level of processing of the data and linkage to other sources may be necessary in order to make the data usable for QA and to establish their quality vis-à-vis other sources.

168. The results of the QA at the Data Stage inform any corrections that are necessary (including through the resupply of data by the supplier). They also inform the necessary processing of the data for use in the census design, through an understanding of the error that must be adjusted or accounted for (Chapter 6). Furthermore, they provide information needed to understand the implications of any errors in the sources on the final census outputs (Chapter 7), which would need to be communicated to users.

5.1 Data quality dimensions

5.1.1 Harmonization and Validation

169. A general assessment of the accessibility of the data is part of the assessment of quality at the Source Stage (see Chapter 4). However, it is crucial for the NSO to ensure that the data files actually transmitted are in the required 'readable' format i.e. the databases are structured in a way which can be ingested and read by the NSO's systems. Where this is not the case the NSO may be unable to process the transferred data files.

170. In addition, further data validation and harmonization arrangements should be in place upon data transfer to the NSO, ensuring consistent use across census use cases. Here, the Data Stage provides the opportunity to validate the dataset supplied against metadata collected at the Source Stage, the reference period and other data requirements (e.g. for particular variables). In addition, validation checks and harmonization arrangements may be developed based on previous experience of working with test data (see section 3.5 on feasibility research).

5.1.2 Accuracy and reliability

171. An assessment of the accuracy of the input data should be conducted with a view to identifying 'measurement' and the 'representation' errors within the administrative dataset (see Chapter 3), as described in Zhang's (2012) two-phase life-cycle model and adopted in quality assessment literature (e.g. Stats NZ 2016 and KOMUSO 2019 WP1).

5.1.2.1 Representation Errors

172. Representation errors (errors relating to the target units, see Chapter 3) might occur if data are not reported correctly to the administrative authority resulting, for example, from non or delayed self-registration on an administrative register (e.g. birth, death or full population register). In addition, some data records may not be transmitted to the NSO because of technical problems or be transmitted with errors, if units are not maintained properly by the authority (e.g., resulting in duplicates). It should be noted that representation errors may cause measurement errors where the unit of statistical measurement changes. For example, a person missing in the administrative population register may lead to an understated value for the variable 'size of household'. Furthermore, for an assessment of the overall coverage of a dataset, an examination of both over and under-coverage is needed. Under-coverage may be of particular importance with respect to so-called 'hard-to-reach' populations (see Chapter 3).

5.1.2.2 Measurement Errors

173. Implausible or missing values are indicative of measurement errors, i.e. errors within variables and thus reduce the accuracy of the raw data (see Chapter 3). To assess whether a value is implausible or missing, it is important to examine both specific records, but also variable distributions for all records. Reasons for a lack of accuracy might be technical, for example, errors in the process of data transfer; or systematic resulting, for example, from inadequate submission or maintenance at the supplier end, particularly if the variable is not, itself, of administrative importance for the data supplier, and is therefore not systematically recorded (such as a person's occupation in the Austrian tax register) (Eurostat ESSnet KOMUSO 2019 WP1, 2.2.2).

5.1.2.3 Re-supplied data

174. In general, the regular maintenance, and updating, of the data source conducted by the data supplier/owner will improve the quality of the data. However, many registers are subject to changes in structure and/or content resulting, for example, of internal administrative requirements and processes. These may in turn lead to a loss of quality, particularly regarding comparability. Therefore, where data is being supplied periodically there is a need for additional, longitudinal, quality assessments. Nonetheless, re-supplied data offers the opportunity to assess the reliability of specific variables, i.e. the closeness of initially supplied values to the subsequently re-supplied values within a dataset. In general, it is assumed that later (i.e. more up to date) values are more accurate.

5.1.3 Timeliness and punctuality

175. It is important that the difference between the reference date to which the data refer and the date on which they are supplied to the NSO is kept to a minimum, since the longer the delay, the less relevant those data become, even though they may still be accurate (cf.

UNECE 2018, 4.3.4). This gap between reference date and acquisition by the NSO is referred to as timeliness.

176. Moreover, punctuality – the difference between the expected date of delivery and the actual date of delivery – is also important as the NSO will usually have a responsibility for producing census outputs to an agreed schedule and would not want any delay in the supply of the census data to affect this.

5.1.4 Linkability

177. Often determining the quality of a dataset will require its linkage to another for comparison. In addition, if the NSO relies on more than one source of administrative data for its census dataset, it is necessary to be able to link data from the different sources at the unit/record level (see Chapter 6). The degree of success of such linkage will affect both the accuracy and the relevance of the input data.

178. A common unique identifier reduces the effort required to link the data by easing the process of evaluating completeness and accuracy of matching. In the absence of such an identifier it is more difficult to link data reliably. In this case record linkage using multiple variables that are common to the units in each data source (typically, name, date of birth, sex, and address) may be possible (see Chapter 6). In this case, the NSO needs to be assured that such 'matching' variables are of sufficient quality in all sources, otherwise the quality of record linkage, and thus the reliability of the data, will suffer. The quality of the linkage variables will ultimately impact on the risk of false matches and false nonmatches in later production stages (cf. Eurostat 2014, section 3.5.2) (see also Chapter 6).

5.2 Tools and indicators

179. The following tools and indicators are useful to the NSO in assessing the quality of raw data against the dimensions discussed in 5.1 above. This application of the tools and indicators supports a consistent assessment across different sources, to decide whether administrative data are fit for purpose.

5.2.1 Harmonization and validity

180. To ensure the readability and validity of the transmitted data files, it is crucial to implement technical checks to validate the data files against the expected data format. If this validation fails, the NSO may require the data files to be re-submitted in the correct format. Before such checks can be meaningfully carried out however, data must often undergo a basic cleaning and/or harmonization process, so that they are comparable to other sources and are optimised for use with the NSO's statistical software.

181. Examples of harmonisation processes include, consistent coding of missing values, formatting of date variable types and removal of dealing with duplicate records from the dataset. Data harmonisation rules should be agreed within the NSO and applied consistently to data, regardless of the different census use cases for which they are intended. In addition, data harmonization and validation results should be documented.

182. Previous literature has identified specific indicators which can be used to assess the validity (e.g. Daas et al. 2009; Eurostat ESSnet MIAD 2014, Cerroni, Di Bella and Galiè 2014). These include:

- The variables supplied are correctly named and formatted (e.g. numerical, categorical, data variables etc);
- The correct reference period has been supplied;
- The variables match the expected pre-defined content, established through the metadata collected at the Source Stage (and through feasibility research, where possible);
- No unexpected differences between current and previous supplies of the data source are found with respect to number of records and variables (further examined below);

183. Harmonization rules and validity checks need to be developed based on production needs and specific planned uses of administrative data within the census design. As such, this guidance is not prescriptive. In addition, linking records from the supplied data to another reliable data source at the unit level provides a tool for determining whether or not the correct reference date is supplied (Asamer 2016). It is also possible to check variables with date specification to determine whether or not they are compatible with the census reference date. A correct reference date is important, especially for changeable variables such as current activity status for seasonal workers. Where possible any such inconsistencies should be corrected at the Process Stage (Chapter 6).

5.2.2 Accuracy and reliability

5.2.2.1 Representation Errors

184. A variety of indicators can be used to measure the accuracy of the supplied units/objects, providing an assessment of representation errors in the data (see Daas et al. 2009; Eurostat ESSnet MIAD 2014, Cerroni, Di Bella and Galiè 2014)[11]. Basic indicators include:

- Total number of objects/units (for comparison against expected count);
- Percentage of duplicate units/objects (if possible).

185. Additional or **enhanced indicators** (see Cerroni, Di Bella and Galiè 2014, p. 128) include:

- percentage of objects involved in non-logical relations with other (aggregates of) objects, described elsewhere as ‘inconsistent objects’;

[11] As noted in the glossary, some of the literature (e.g. Zhang 2012), the term ‘object’ is used to refer to the units within an administrative dataset. The term is used to distinguish between units in the administrative data and the statistical units after this data has been transformed in some way. This is particularly relevant in cases where the unit (or ‘object’) in the administrative register differs from the target statistical unit. For example, where a tax register, where the units of a yearly tax returns (i.e. the same person may make several returns in one or multiple years), is converted into individual ‘people’.

- per cent of objects involved in implausible but not necessarily incorrect relations with other (aggregates of) objects, described as 'dubious objects'.

186. A broad assessment of over and under-coverage of the data can be made by computing and comparing the total number of units, as well as cross-tabulations of frequency/percentages across characteristics (e.g. sex, age, geography) on an **aggregate level**, between the administrative source and other/alternative sources taken as reference or a comparative 'gold standard'. The indicators suggested by Cerroni, Di Bella and Galiè (2014, p. 129) include:

- Under-coverage:
    - per cent of objects of the reference source missing in the supplied source.
- Over-coverage:
    - per cent of objects in the source not included in the reference population;
    - per cent of objects in the source not belonging to the target population of the NSO.

187. The above is subject to two key assumptions. Firstly, a complete base register (including the target population of the input data) must be available to compute over-coverage.[12] For instance, deceased persons may still be reported by a country's central population register but may be identified as such in a central social security register. Secondly, it should be clear which objects of the complete base register should be included in the input data to compute under-coverage. For instance, school-aged children in the statistical population register should be largely covered by the register of enrolled pupils and students.

188. Finally, comparisons can be made at the unit/object level to determine the percentage of units/objects which are consistent within and across sources, described elsewhere as a measure of 'authenticity' of the objects (Cerroni, Di Bella and Galiè 2014, p. 128). An example of inconsistent objects might be where each unit or row within an administrative dataset represents an event of registration (e.g. with a doctor) which includes, name, address code, together with date of registration as well as (maybe) date of deregistration. Two objects relating to a single person are inconsistent if the period of registration of the objects at different addresses overlaps. The percentage of inconsistent objects provides an indicator of error or conversely, accuracy. However, as noted by Zhang, object-level analysis has its limitations as sources may differ at the micro level but result in similar statistical measures such as means, medians etc. Unit-level analysis "may fail to reveal such statistical equivalence" (2012, p.45). In addition, where unit-level comparisons are made between multiple sources, it is important to note the possible impact of selectivity bias within the linkage process on any resulting differences.

[12] In the literature, base or core administrative registers are often distinguished from additional registers (e.g. Daas et al. 2009). Base or core registers are those assumed to have the most exhaustive coverage of the target resident population.

5.2.2.2 Measurement Errors

189. As above, previous literature contains **basic indicators** to measure the completeness of the variables supplied within administrative datasets at the **aggregate level** (e.g. characteristic variables such as age, sex, ethnicity etc) (see Daas et al. 2009; Eurostat ESSnet MIAD 2014, Cerroni, Di Bella and Galiè 2014). They include:

- Number and percentage of missing values within key variables;
- Number and percentage of out-of-range values within key variables (for example a recorded age of 120 years)
- Number and percentage of implausible values (based on, for example, cross-tabulations of different variables);
- Prevalence of unexpected frequencies, patterns or outliers, based on frequency/distributional analysis of key variables (**aggregate** comparisons with external sources, as well as expert knowledge can be used to identify data oddities also);

190. The degree of consistency of the supplied data at the aggregate level, namely that relationships between related variables are consistent and not implausible, provide a measure of the accuracy of variables. However, to assess consistency at the micro level, **enhanced validation checks** for related variables within a supplied data set should be carried out. Based on previous literature, key indicators include:

- per cent of objects whose combinations of values for variables are involved in non-logical relations;
- per cent of objects with dubious variable values or objects whose combinations of values for variables are involved in implausible but not necessarily incorrect relations (i.e. outliers);
- prevalence of objects with missing values for key variables that have different characteristics to complete objects;
- prevalence of objects with values imputed by the data provider for the main variables of interest; and
- prevalence of rounding for the main variables of interest (can be detected by analysing the distributions)

191. **Statistical techniques and metrics**, such as frequency distributions, can reveal unexpected patterns and outliers. For example, a cross tabulation of age and marital status may lead to the identification of implausible cases, such as a 5-year old child that is married. Other examples include the comparison of date of birth with that of other events in the German case study in section 5.4.1, and the cohesion analysis of address data in the Polish case study in section 5.4.2. Observed patterns might indicate systematic measurement errors, as illustrated for example in the case study from Germany (section 5.4.1). Note that if

inconsistencies are identified and the data supplier cannot fix such problems, then certain edit steps (as part of the Process Stage, Chapter 6) may be necessary.

192. Similarly to the assessment of representation error, an efficient way to assess variable accuracy, especially in the preliminary analysis of data and the very first time the data are used, is the **comparison of data**; that is, where the input data are checked by means of comparison with other independent sources that contain the same variable. Suitable independent sources for comparison could include a national survey (such as a labour force survey) or a previous census (see, for example, Asamer 2016).[13]

193. More complex methods for assessing the accuracy of administrative data, where administrative data are linked to a comparative source (which includes the variable / concept of interest) are described in the literature. Bakker (2012) uses structural equation models to estimate the validity of administrative variables, using survey data. The model is applied to data on age, gender, educational attainment and wages. Scholtus and Bakker (2013) also used a simulation study to test the robustness of the model to additional components of measurement error, to misspecification of the measurement model and to small sample size.

194. Oberski et. el (2018) apply a generalised multitrait-multimethod (GMTMM) model, under a general framework for evaluating the quality of administrative and survey data simultaneously. The framework allows both survey and administrative data to contain random and systematic errors (therefore does not assume the survey has no error, like other methods (Yucel and Zaslavsky, 2015)). Their approach accommodates common features of administrative data such as discreteness, non-linearity and nonnormality and so may improve on other models used (such as structural equation models).

5.2.2.3 Re-supplied data

195. Administrative data may be re-supplied to ensure NSOs have access to the most recent and relevant data for use in the census. As with data supplied for the first time, the first step to assess the quality of re-supplied data is to perform a macro-level comparison of the main key metrics, such as total number of records, number of missing values, etc against what was expected to be received. For resupplied data, a comparison with previous supplies will identify any unexpected differences across the datasets that may indicate a quality concern. Furthermore, longitudinal comparison between the data supplied in the active and previous period is important for revealing possible quality changes, especially in terms of coverage, completeness, and linkability.

196. For key variables that are expected to be stable over time , it is possible to compare values for the same unit (e.g. a person) over time to check for unexpected changes. These

[13] It should be noted that consistent values and cross-tabulations generate through different sources and methodologies (e.g. administrative data and survey data) suggest that both sources are likely to be correct. Inconsistent values leave an open question as to which result is most accurate, i.e. closest to the true population value. This depends how survey questions are answered, and how the administrative source is collected, which again highlights the importance of the Source Stage. It is not always true that the administrative data source will be less accurate (e.g. see literature on receipt of state benefits). A more sophisticated analysis is needed to determine the accuracy of both the administrative and external source to determine the cause of inconsistencies found.

checks are easier for 'invariant' variables, such as date of birth or place of birth, and for data where a unique key is available and stable over time. Even for changeable variables such as legal marital status or highest level of education however, such checks may still be possible in a restricted form. Such longitudinal comparisons can serve as an internal measure of the **reliability** of the data, by providing indicators such as the means or medians of differences or relative differences between the newest and previous data supplies.

197. If there is no key variable that is stable over time however, then at least the distribution of the variables can be used to compare the periods.

5.2.3 Timeliness and punctuality

198. Measures of timeless and punctuality as defined above can easily be determined by comparing the reference date, the specified delivery date, and the actual delivery date of the data. The following indicators are suggested by Cerroni, Di Bella and Galiè (2014, p. 130):

- Timeliness
    - Time difference (days) = (Date of capturing the change in the source by the data source holder) – (Date the change occurred in the population)
    - Time difference (days) = (Date of receipt by user) – (Date of the end of the reference period over which the data source reports)
- Punctuality
    - Time difference (days) = (Date of receipt by NSO) – (Date agreed upon; as laid down in the contract).
- Overall time lag
    - Total time difference (days) = (Predicted date at which the NSO declares that the source can be used) – (Date of the end of the reference period over which the data source reports).
- Delay
    - Time difference (days) = (Date of receipt by NSO) – (Date of the end of the reference period over which the data source reports)

5.2.4 Linkability

199. Often, administrative data sources will be linked to other sources be it the census enumeration itself or other administrative sources. A quality assessment of the variables in the single sources to be used in the linkage provides general information to inform the design of a successful linkage process as described in Chapter 6.

200. Regardless of whether a unique key or identifier variable is available or whether several variables will be used in combination to identify matches in the linkage process, these indicators should inform the choice and evaluation of the quality of linkage variables supplied, including:

- Number of or percentage of duplicate linkage keys, which can be calculated for a single identification variable (e.g. a personal identification number), or a combination of variables which provide the linkage key (e.g. age, date of birth, address);
- Accuracy indicators as described in previous sections (including, missing values, implausible values, etc.) the extent to which measurement errors are not random – e.g. evidence of any systematic biases (e.g. where certain population groups may be subject to greater error against the linkage key variable(s)).

201. Finally, if and where the linkage variables have been provided to the NSO in an encrypted or 'hashed' form (i.e. masked via a one-way algorithm to protect the privacy of the data subjects), it must be verified that the hashing performed by the supplier matches the hashing algorithm used at the NSO. Otherwise, it will not be possible to link the data supplied to other data sources, undermining the relevance of the data. Chapter 6 provides further details about the linkage of encrypted keys.

5.3 Recommendations

1. As noted in Chapter 3, before using an administrative data source within census production, at least one test run with real data is advisable, if not essential. Such a test should be carried out early enough to allow a readjustment of the technical infrastructure and processes to guarantee the readability of the data.
2. Check that the data supplied matches the metadata collected at Source Stage and that the correct reference date has been supplied.
3. Compute and monitor basic indicators of the supplied data to gage possible representation and measurement errors.
4. Verify consistency of objects resp. variables within a supplied dataset through enhanced validation checks.
5. Use statistical metrics to reveal unexpected patterns and outliers.
6. Total number of records and cross-tabulations are compared with independent comparable sources, to assess accuracy.
7. Ensure NSOs have the ability to clarify data queries with supplier. Where queries regarding the data arise post-supply, there should be adequate mechanisms in place to ensure these can be resolved.
8. To improve input quality and ensure consistency, provide feedback to the data supplier about any anomalies (such as inconsistencies within the dataset) found, at least on an aggregated level, providing that the relevant laws on data protection allow this.
9. Where data is being supplied periodically there is a need for additional, longitudinal, quality assessments.

10. Determine the timeliness and degree of punctuality of data supplies.

11. Determine the quality of linkage variables to guarantee the best possible linkage results (see Chapter 6).

5.4 Case studies

5.4.1 Germany: The quality of the data provided from the local population registers for the 2021 census

5.4.1.1 Introduction

202. The German National Census 2022[14] is a combined census using data from multiple sources. Data from all local population registers of the approximately 11,000 municipalities – administrated by around 5,100 local registration offices – is the fundamental source of data, but other information (not specifically relevant to this case study) is collected from a variety of other official sources such as the Federal Mapping Agency, the Federal Ministry of Defence, the Federal Foreign Office and the Federal Ministry of the Interior, Building and Community. In total, six deliveries of data from local population registers are scheduled for the Census 2022. Since giving a subsequent notice of departure/arrival is allowed in Germany there will be two deliveries the calculation of the population figure is based on – one with reference date equivalent to the census reference date and one with reference date three months after the census reference date.

203. This case study focuses only on the quality of German population registers data and problems that occurred during the delivery of that data in January 2019. The 2019 data simulated the largest dataset from the population registers that is to be delivered in the context of the census 2022. The data delivery in January 2019 was a test run to assess the quality of the raw data, to test data transmission, to optimize existing techniques of data processing and to test the transmission of historical data records. Note that some critics considered a test with anonymized data or a random sample to be sufficient.

204. The case study focuses only on the examination of input quality. Note that for statistical usage there is no unique identifier for a person available in Germany.

205. In general, the data set contains every person who was registered with first or second residence at the reference date on January 13 2019. Furthermore, the data includes historical records on recent changes in the registers close to the reference date.

206. Since the previous national census in 2011, measures have been taken to improve the quality of German population registers. Firstly, when a habitant moves from one municipality to another the registration offices in the two municipalities communicate automatically. Moreover, local registration offices communicate any change in its population register to the Federal Central Tax Office; and since every person has unique tax ID it is highly likely that the

[14] The census was originally scheduled for May 2021 but was postponed to May 2022 due to the Covid-19 pandemic.

number of first residence duplicate records in the data of 2019 has shrunk since 2011. However, this trend is still under examination.

5.4.1.2 Readability

207. In Germany, a standardised, universal format has been determined for the transmission and delivery of data from a local population register. The recipient (which, in the case of data to be used for the purposes of the census, is the Federal Statistical Office) only accepts the data if it is delivered in this format. This helps to improve the input quality of the data.

208. At least four municipalities tried to transfer some variables in a format that violated general formatting rules, but could not deliver the affected data records which consequently led to an incomplete data delivery. For subsequent data deliveries the format of these variables was broadened so that this problem should not occur again.

5.4.1.3 Accuracy

209. Recalling that population registers are administrated locally, it is no surprise that the accuracy of the data varies across municipalities. The following two examples illustrate the variety in the degree of accuracy of the data:

210. In the first example, for more than 40 municipalities in one or more of the three variables 'date of moving to an address', 'date of moving to the municipality' or 'date of registration', some 75 per cent or more of all first residence records contain the same date. It can be assumed that this is an error made during a data merge that became necessary because of a consolidation of two or more municipalities. Such peculiarities can be critical for handling first residence duplicate records.

211. In the second example, in some 120,000 data records, one or other of these three, and some other, 'date' variables recorded dates that were earlier than the date of birth. In particular, one State had 60 per cent of these erroneous records in its registers.

212. In order to improve input quality, the municipalities received feedback about anomalies found in their data on an aggregated level, the plausibility checks on the data need to be expanded and an exchange with the producers of software for the population register subsequently took place.

5.4.1.4 Completeness

213. During the 2019 data delivery, several technical problems arose which also had a negative impact on the completeness of the delivered data.

214. Due to an error in the software the municipalities used to retrieve the data, approximately 1,200 municipalities had transmitted files with missing data records. This error was detected only by accident: For some of those municipalities the software provider as well as the municipality itself initiated a data delivery, due to miscommunication. (In some cases the software provider holds a mirrored register of all the data from the relevant municipalities). A comparison of these two deliveries showed that the software provider failed to transmit some data records. The software provider had to schedule a second delivery replacing the former one. The data delivered by the municipalities themselves was deleted.

Therefore, the technical infrastructure needs to block the integration of doubled deliveries or of a delivery that consists of data combined from different senders.

215. Generally, it is hard to identify, if some records are missing, since the recipient may have no information on the exact number of records that have to be delivered. The recipient can only compare the number of the transmitted data records on first residences in a municipality with its own rolled-forward population estimates. However, it is not uncommon for these two figures to differ by up to several percentage points.

216. Some municipalities did transmit, for every data record, missing values for some variables. This showed up an incomplete data retrieval from the local databases. For instance, the variables 'most current date of moving to Germany' and 'country of origin' (which should be empty if it is Germany) were blank for all data records in approximately 1,200 municipalities. Prior to integrating the data into the database it is thus important to check whether or not variables are missing throughout the entire municipality due to technical problems.

5.4.1.5 Time-related dimension

217. Some municipalities were not able to compile their data until several days after the reference date. Thus, a person who, for example, reports a (subsequent) notice of departure in such a municipality during the intervening period is not covered. To reduce the possible damage of such a mistake during future data deliveries, it is crucial that municipalities create the capability to retrieve an historic state of their registers.

5.4.1.6 Conclusion

218. Technical problems during and before the delivery lowered significantly the quality of the data received from local municipalities' registration offices. Hence, the test run for the Census 2022 in January 2019 was important to assess procedural and technical flaws. A test run with anonymized data or a random sample would not have detected most of the described flaws. Its timing more than a year prior to the next Census 2022 data delivery provides enough time to analyse and eradicate errors and to optimize data processing capabilities on the central as well as local level. Furthermore, municipalities were informed about anomalies found in their data on an aggregated level since it is legally forbidden for the Federal Statistical Office to return individual data records. This will, hopefully, help to improve the input quality of the data delivered from the population registers.

5.4.2 Poland: The Polish variable quality system

5.4.2.1 Introduction

219. For the purposes of censuses in Poland, data are collected from multiple sources, including administrative ones. Registries and database systems are characterized by a wide variety of content and complexity of structure, resulting from the fact that they are created for different purposes and are managed by different data owners. Accordingly, the standards of storage, accuracy and recording methods adopted in each case also vary. The lack of uniformity exists not only between the registers but also with the data within any one particular register.

220. The quality of data from administrative sources used has an effect on the quality of the census results. Therefore, adequate input quality is a prerequisite (although not the only one) to obtain correct census results. Thus, when using administrative sources (not only in the context of census production), the key elements are to identify and understand problems and errors encountering in the data, and then unify and correct the data. For the assessment of the input quality, the first point is especially important.

221. Having assessed the viability of using particular administrative sources, the process of managing the quality of data collected from administrative sources in Poland is divided into three parts: input, process and output quality. This process of managing quality is monitored constantly. At Statistics Poland the variable quality system (VQS) is used for this purpose. The VQS is a system for viewing, analyzing and reporting data from administrative sources.

222. At first, the VQS validates the data received. The process involves applying a set of rules assessing the dataset for completeness, consistency, and correct format for consumption into the system. A key consideration is the completeness and accuracy of the unique identifiers provided in the data supply – this is critical to ensure high quality data integration. Missing or erroneous values in the unique identifier field prevent records from being integrated effectively across multiple data sources. Data that do not pass the validation assumptions are set for correction – a harmonization process to align the data to the expected standard.

223. Following both the validation and correction steps, a data quality improvement report is produced to inform decisions on whether to approach the data provider in efforts to improve the data quality at the point of supply, or to complete any additional data processing. It enables Statistics Poland monitor closely all the changes that are taking place in administrative data sources used in official statistics, and permits the automation of the calculation of quality indicators for both input and output data. This case study focuses on the assessment of the input quality.

5.4.2.2 Accuracy and Reliability

224. The VQS contains the results of the polish data profiling of the raw data. Data profiling is a procedure with which the user obtains, among other things, information on the accuracy of the raw data. It gives access to a series of statistical metrics:

- ordinal position
- data type
- count (number of records)
- non-null count
- data length.

225. For numeric variables:

- minimum value
- maximum value

- mean
- median.

226. For character variables:

- pattern count
- unique count
- minimum length
- maximum length
- frequency distribution
- pattern frequency distribution.

227. Within the VQS a cohesion analysis of address data is conducted to check their accuracy and consistency. The address consists of the following hierarchical levels of the territorial division:

- voivodship (or province, of which there are 16 in Poland)
- powiat
- gmina
- locality
- street.

228. Considered separately, the individual address field values may comply with the standard, but as a whole do not form a consistent address string appearing in the National Official Register of the Territorial Division of the Country (TERYT). To consider the address to be valid, the correct parts are not, in themselves, sufficient. The logical structure must also be kept: that is, the street must be located in the locality, the locality in the gmina, the gmina in the powiat, the powiat in the voivodship. Only addresses following this structure are considered as consistent. Cohesion to the street is considered as full cohesion, cohesion to the level of the city (compatible sequence of the voivodship, the powiat, the gmina, the locality) or the gmina (compatible sequence of the voivodship, the powiat, the gmina) needs to be improved or supplemented by other available data. With respect to the cohesion analysis, the VQS generates the following quality indicators:

- TERYT dictionary comparability (number)
- change of TERYT dictionary comparability (per cent)
- conversion dictionary comparability (number)
- change of conversion dictionary comparability resulting from various stages (input, output) (per cent)
- level of cohesion of address variables (flag).

229. In order to check the completeness of a variable, the VQS generates the following quality indicators for every variable:

- completeness (number)

- change of completeness (per cent).

5.4.2.3 Timeliness and Punctuality

230. Long-term, effective and transparent cooperation with administrative data owners is crucial. In Poland, the acquisition of data for census purposes is supported by a legal framework including both a Statistics Act and a Census Act. The VQS records information on the reference date of the data and the date of receipt of the data by Statistics Poland.

231. Data are usually collected at the end of the year or according to the date of the relevant survey. Data for the needs of the decennial census are collected as soon as possible during its implementation, permitting the necessary time required to process the data. In order to maximize the relevance of the data, the collection should either be as close as possible to the reference date of the census or, if the receipt of data is a continuous process, as close as possible to the reference date of the data.

5.4.2.4 Linkability

232. Completeness and accuracy are crucial for unique identifiers such as:

- the PESEL number: a permanent numeric symbol, widely used by administrative registers relating to persons, which uniquely identifies a natural person and through its uniqueness allows to distinction between many people bearing the same name and surname
- REGON: business identification number
- NIP: tax identification number.

233. Identifiers should be characterized by the required number of characters and the compliance of the check digit. The high quality of identifiers is of utmost importance in the course of data integration. Missing or erroneous values do not allow the same entities to be identified in different sources. The VQS generates the following quality indicators for identifiers:

- number of correct identifiers (number)
- change of number of correct identifiers (per cent).

5.4.2.5 Conclusion

234. Within the methodological framework for improving the input, process and output quality, the VQS is an important tool for controlling data quality, make quality comparable among different suppliers, and monitoring quality changes over time.

Chapter 6. Process Stage

235. Once administrative data are received and quality assessed by the NSO, they will require some processing to make them usable for the census. For example, they will need to be integrated into the census design and any quality issues will need to be addressed (e.g. conceptual misalignment with the census definitions and concepts, coverage and measurement errors). The Process Stage of these guidelines provides a basic overview of some of the key processes used to integrate administrative data into the census and the related quality concerns (also see KUMUSO, Quality Framework for Multisource Statistics, 2019 WP1 for quality indicators, measures and methods for assessing process (and output) quality).

236. The required processing of the administrative data is informed by the information that is gained from the Source and Data Stages. For example, the assessment of the linkability of an administrative source informs how data are linked. An understanding of coverage error informs the processes for integration of data to achieve the coverage needed for the census. The assessment of accuracy of the administrative data will inform the editing and imputation stages; and will provide the necessary insight to support decisions about how sources should be used together to construct the census variables. However, processing can introduce additional error that can be systematic or random, thus introducing bias or variance in final estimates. For this reason, it is important that processes are appropriately tested and evaluated, and that error is managed along the whole statistical production chain.

237. We focus on some of the most common processes that are required when using administrative data for the census. These are: record linkage; processes for assessing coverage error in statistical registers, or when administrative data are used in the enumeration of population units; conflict resolution, when inconsistencies are found in the values of data items from different sources; editing and imputation. Each of these processes is described in more detail in the following sections, along with the challenges associated with them and ways to assess quality of the processed data, based on the available literature and the experiences of different countries.

6.1 Record linkage

238. Almost every administrative data source requires some form of record linkage to other data sources (including administrative and survey data), whether it be for validation purposes or for ensuring adequate coverage of the census population units and variables. For example, two or more data sources may need to be combined to achieve better coverage of the target population, including to adjust for potential over-coverage (see section 6.2). Likewise, linkage of multiple sources may be necessary to provide complete and accurate data for the census variables (see section 6.4).

239. Many countries integrate administrative data from multiple sources to create so called administrative-based statistical registers, including address, population and business registers

(see UNECE 2018, Chapter 8 and section 6.2 below). And even countries without statistical registers, are moving towards maximizing the use of administrative data in the production of their core population, social and business statistics, which traditionally were primarily based on a census collection or a survey.

240. This makes record linkage one of the most important processes for using administrative data in the census. In addition, linkage is used to assess the quality of the separate administrative sources with respect to measurement and representation error (as described in the Data Stage) and also to assess the quality of integrated data and statistical registers (see sections 6.2 to 6.4 below). It is therefore important to assess the quality of the linkage process, through an assessment of the linkage variables or keys (as described in the Source and Data Stages) and through an assessment of the process itself (as outlined in the sections that follow).

241. The impact of linkage error on the overall accuracy of population and census estimates should also be considered, including both representation and measurement error (see Daan Zult et al 2019). For instance, missed and false links can lead to over- and under-coverage of the census population and can cause the wrong value to be assigned to a census variable (e.g. employment status, household composition) for a given unit (e.g. a person, household). Address data often need particular attention, as they can be used for both linking data for an individual (e.g. as a linkage key in combination with name/date of birth) and linking individuals together to form households. However, people do not always alert an Administrative Authority when they move address, or a registered address might not be the primary address of residence, thus the accuracy of address data can be poor in administrative sources. In addition, linkage error can introduce bias in dual-system estimation (Abbott 2009).

242. Methods for linking data typically rely on the existence of common unique keys (or identifiers) across the sources to be linked. For example, Poland has developed a population list from integrated administrative sources called the Statistical Census Dataset (SCD). The SCD is constructed through a set of integrated 'core' administrative sources with high coverage, completeness and timeliness. The datasets are integrated using a unique identifier, the PESEL number, a 11-digit permanent numeric symbol that uniquely identifies every person registered in the PESEL database (Polish abbreviation for the Universal Electronic System for Registration of the Population).

243. In the absence of common unique identifiers, other common identifying variables, such as address, name, sex and date of birth, may be used to link records from multiple sources. Although this is more challenging and subject to much higher levels of error, as outlined below.

244. In some cases, the NSO may only have access to anonymized or 'hashed' identifiers in the administrative data. Hashing is a practice that is often used in computer science to protect confidentiality of individuals or other entities in data. It involves applying an algorithm to every piece of information in the original data (e.g., a name) to create a string of characters that uniquely identifies that information and mask the original data. Hashing has some

important quality implications for data linkage (see Shipsey and Plachta (2020) for a description of methods for linking with anonymized data, the challenges and limitations).

245. Linkage methods are of two main types: deterministic, when matches are made based on a set of common identifiers, and probabilistic when matches are made based on model-based linkage weights (Harron, Goldstein and Dibben 2015). Probabilistic matching does not require record values to be identical between two records but relies on similarity between records. One additional linkage method that can be applied to unlinked records after deterministic and probabilistic methods are applied, is clerical linkage, which involves a visual inspection of the unlinked records. Clerical linkage is not possible to do when the data are hashed.

246. Linkage error can occur through unlinked records that should have been linked (also known as 'false negatives') and linked records that shouldn't have been linked (also known as 'false positives').

247. Two very common methods for assessing linkage quality are:

- Estimation of false positive and false negative rates, through for example a clerical review of a sample of linked records. Although clerical review can only be done when the data are not hashed. If the data are hashed, the NSO should try and obtain access to a sample of the linked records with all the original information on the data to assess the linkage.
- Comparison of the distributions of characteristics of linked and unlinked records, e.g., age, sex, ethnicity. Differences in characteristics suggest bias is introduced by linkage error, which means that certain types of records (e.g., individuals) will not be linked because they are more difficult to link.

248. The assessment of linkage error using the methods described above is presented in the United Kingdom and New Zealand case studies, sections 6.7.1 and 6.7.3.

Box 7: Methods for data linkage and the assessment of linkage quality: a UK cross-government review

The importance of linking administrative data for the public good (including for the census) is widely recognised and resulted in a cross-government review within the United Kingdom to develop guidance on data linkage methods, covering the quality assessment of linkage. The review drew on the work of experts across government, academia, the private sector and internationally. The outcome was a series of articles covering: the future of data linking methods; quality assessment in data linkage; longitudinal linkage (design principles and the total error framework); preserving privacy; linking with anonymized data; and procedures for improving efficiency (see ONS 2020).

6.2 Statistical registers and the 'signs of life' methodology

249. As mentioned in section 6.1, integrating data from different sources for use in a census is becoming increasingly common, and record linkage plays an important role in this process. Two key quality dimensions related to the integration of data from various sources are coverage and coherence, as integrating data is done to assess and possibly reduce coverage error, and it also enables and requires assessment of coherence of information across sources and over time. There are also coverage/coherence issues that are specific to administrative data, such as, people not always alerting the relevant Administrative Authorities when there is a change in their residence or other personal/household characteristics.

250. One example of data integration for use in a population and housing census are administrative-based population or address registers. By linking information from the available sources at the record level it is possible to determine individuals or households that are resident in a country and their characteristics. The integrated data from these sources become a statistical register, namely, a database that can be used for further processing and analysis to produce census-type outputs (see UNECE 2018, Chapter 8).

251. Some of the key processes involved in the construction of a statistical register are:

- Identifying the data sources to be used
- Linking the sources
- Developing and applying a set of rules to make decisions about which records should be included in the final estimates
- Resolution of conflicting information (e.g., date of birth, or address) between the linked sources
- Editing and imputation.

252. The quality considerations and indicators suggested in Chapter 4 and Chapter 5 will help identify the data sources to be used in a statistical register. Linkage error has been discussed in section 6.1. Quality assessment of conflict resolution, and editing and imputation, are covered in sections 6.4 and 6.5. This section focuses on the application of decision rules and some quality considerations related to this process. The section also discusses other methods of coverage assessment that can be used in statistical registers along with decision rules.

253. Decision rules, or 'activity' rules, are criteria for inclusion that are often applied when constructing population registers to ensure that only individuals who meet some pre-defined usual residence criteria are included in the final estimates. This process is sometimes known as the 'signs of life' (SOL) method and is a widely used tool to minimise over-coverage in statistical registers (e.g., records that are not part of the usually resident population). Spain uses 'signs of presence' on four income administrative sources, including data from the tax agency files and social security files, and movements detected within the statistical register *Padrón*. These administrative sources are compiled at the individual and household level, and

individuals who have reached the threshold level of signs of presence are considered 'active' and included within the population count, whereas all the others, called 'inactive', are not (see Vega Valle et al 2020 and the case study from Spain, section 6.7.2 for more details).

254. The UK Office for National Statistics uses a similar approach to decide which records from selected administrative sources should be included in their administrative data-based population estimates (ABPEs) (see ONS 2019b). For example, in an earlier version of the ABPEs a record was included in the population estimates if it was present on two of the selected administrative sources, while in a subsequent version of the ABPEs, strict criteria for inclusion were applied to each source separately (only include records who had a sign of activity within the last 12 months) and the rule of including records only if present on two sources was removed (with data linkage only used to deduplicate records that appeared on multiple sources). The sub-sequent version of the ABPEs aimed at further reducing the over-coverage found in the previous version, at the expense of increasing under-coverage (records that are missed from the population estimates), as under-coverage was expected to be addressed using a coverage survey combined with a dual-system estimation type of method.

255. The success of a SOL method relies on the availability of good indicators of signs of activity in the individual or combined administrative data. The application of the method typically involves making some assumptions, which determine who is considered as active and who is not, and NSOs should be clear about those assumptions and provide relevant supporting evidence. In particular, the choice of signs of activity indicators (or decision rules) should be informed by an assessment of quality at the Source and Data Stages (see Chapter 4 and Chapter 5), including consultation with data suppliers, cross-validation between sources and over-time, and expert opinion.

256. As already mentioned in the case of the UK, the application of SOL methods may be combined with other methods to assess and account for coverage error in statistical registers. One of these is to conduct a survey that is independent from the statistical register and use the combined information from the survey and the register to estimate the number of records that are missed in the register (or in the survey) and improve the final estimates. In the context of population estimates, this is like conducting a post-enumeration survey after the traditional census and applying dual-system estimation (also known as capture-recapture) methods to assess the level of under-coverage in the census (Abbott et al 2020).

257. Over-coverage in statistical registers can also be assessed through linking the register to a survey through an approach called 'dependent interviewing', which aims at verifying administrative records in the field. This approach has been used in Italy and in some other countries (e.g., Israel) that have successfully transitioned to primarily administrative data-based censuses (Brown et al 2020). However, not all countries can carry out dependent interviewing, due to ethical and privacy concerns (see Chapter 4).

258. In Italy dependent interviewing (with a sample of households drawn from the Population Base Register (PBR)) and a SOL methodology (using other administrative sources) is used in combination to estimate and adjust for over-coverage error in the PBR. In addition, a sample survey of addresses drawn from the Statistical Base Register of Addresses (SBRA) is

used to adjust for under-coverage. As a result of this process, the population estimates are obtained by applying correction coefficients for both under- and over-coverage errors to individual data on the PBR. The Italian case study, section 6.7.4, provides details of the complete methodology.

6.3 Enumeration of population units: administrative data-based models

259. Related to the construction of statistical registers, administrative data can be used to enumerate population units (e.g. individuals, households, occupied addresses), to support or supplement a census field collection. This approach was used in both New Zealand to address under-coverage in their 2018 Population and Housing Census and in the United States of America (USA) to improve the efficiency of their field Non-Response Follow-Up (NRFU) operation.

260. The approach involves linking integrated administrative data sources to a "gold standard" dataset (in this case the traditional census) to build models to assess the quality of the administrative data and to determine under which conditions the administrative data are used for the census. The approach allows for partial usage of administrative record information where they are believed strongest.

Box 8: Determining occupancy at an address (the United States Census field operation)

For the United States Census the aim was to use administrative data to determine vacant and non-existent addresses and to enumerate occupied addresses as part of the Non-Response Follow-Up (NRFU) Operation. For example, where the administrative data predicted (based on defined cut-offs) that an address was un-occupied, the field contacts could be reduced, thus reducing costs and improving efficiency. Predictive models were developed based on the relationships observed in 2010 between census outcomes (as a "gold standard") and from government administrative records and third-party data. The performance of the models was then tested as part of the 2015 and 2016 census tests, and via a retrospective evaluation using the 2010 Census. Multiple administrative sources (government and commercial) were used, including tax, social security, health, housing and Postal Service data.

The performance of the models was used to determine cut-offs to guard against under-coverage (where addresses are incorrectly classified as vacant by the administrative-based model), while aiming to minimise non-response follow-up workloads. Specific attention was paid to the performance of the model by different geographic areas, with different concentrations of population groups (e.g. Hispanic and Non-Hispanic Black populations). This resulted in further development of the strategy to protect against misclassification of addresses as unoccupied (section 3 of Administrative Records Modeling Update for the US Census Scientific Advisory Committee, 2017 provide details of the quality assessment that was carried out).

Box 9: Direct Enumeration (the New Zealand 2018 Census)

The New Zealand 2018 Census used administrative data to enumerate people that had been missed from the field collection. Census data (previous and current) were linked to administrative records to build models that were used to assess the quality of the administrative data and to determine how and when they would be used to include people, families and households in the census.

The primary aim of the administrative enumeration was to target under-coverage in the census. The linkage method was therefore designed to minimise false positives (i.e., to minimise the number of administrative records incorrectly excluded from the census dataset because they were wrongly linked). Furthermore, an adjustment was made as part of the final enumeration process to reduce false negatives (i.e., administrative records that were incorrectly not linked, and thus added to the census dataset in error, causing over-coverage).

The administrative records that were selected for inclusion following the linkage process, were divided into those to be included into dwellings (with families and households created), and those included at a small geography only (with no relationship to dwelling and no family or household created). This decision was driven by statistical models that were specifically developed to predict the reliability of administrative data for representing households. The models (which used census data) were assessed using receiver operating characteristic (ROC) curve analysis.

To assess the performance of the approach, an indication of the coverage patterns for the census after the administrative enumerations were included was carried out. A newly developed Dual System Estimation (DSE) benchmark population provided the most suitable estimate of the true census usual resident population available at that stage. Population distributions by age, sex, ethnicity and geography were produced and showed that the 2018 Census dataset was largely consistent with the benchmark and in most cases, the inclusion of administrative records in the file greatly reduced (but not resolved all) under-coverage (Stats NZ, 2019a). These indicative findings provided confidence in the new methods when the census data were released. Case study 6.7.3 from New Zealand provides more details of the approach, with a focus on how the quality of the linkage and statistical modelling processes were assessed.

6.4 Conflict resolution/decision between sources

261. As mentioned in section 6.2, when combining administrative data to create statistical registers there may be inconsistencies in the values of key variables across different sources. For example, once a decision has been made on which administrative records to include in the usually resident population, if their address on two or more sources are different (e.g., due to delays in individuals communicating a change of address, administrative processing delays, second/multiple homes), then the NSO may need to decide at which address the records should be included. Conflicting (or multiple) address information and any related decision may cause under-coverage in some areas and over-coverage in others. At an aggregate (e.g., national) level this may not be an issue because whichever the address the person may only be counted once, but at a small-area level this may matter, if the two addresses are located in two different areas, as it will cause over-coverage in one area and under-coverage in the other.

262. Abbott et al (2020) describe three approaches for deciding between sources in the context of address conflict: i) remove the record from the population; ii) split the record between the different locations according to weights (e.g., half if two locations); iii) choose which source to believe is the most likely to be up to date based on the characteristics of the individual or the administrative variables (this approach could also include using additional data sources where the same individual appears). The first approach increases under-coverage in the population estimates. The other two may produce acceptable population estimates at an aggregate level but may introduce significant biases due to coverage and linkage error in estimates at lower level of disaggregation, such as, age and sex. The last two approaches have been tested in the UK as part of their development of administrative data-based population estimates, and further research is ongoing in this area (ONS 2016, section 6).

263. Similar approaches to use quality information/indicators on individual sources to measure the quality of attributes in statistical registers, when the same attribute is available in different sources, have been used in Austria and in Spain. In the Austrian full register-based census a combined quality indicator is calculated using the Dempster-Shafer theory, also known as the theory of belief functions and a generalization of the Bayesian theory of subjective probability (Dempster-Shafer Theory: see Shafer 1992). A comparison with an external source is carried out to assess the associated statistical rules (Statistics Austria 2019).

264. The Spanish population register lacks information on the legal marital status (LMS) of individuals. To estimate LMS, therefore, several registers are used to obtain complete information (Argüeso 2019), including data from the Tax Agency, the Civil Register, the Social Security database and the Central Register of Foreign Nationals. Since an individual may appear in multiple data sources with conflicting information, decisions rules are applied to determine the most plausible value. The decision rules are applied for each person after which a value for LMS may be given. If cases remain unassigned, a value is imputed depending on age, information in past censuses, and number of household members. The results generated by this method are promising and further research is on-going.

265. To summarize, methods for deciding between sources when the same attributes are available in different sources typically rely on decision rules approaches, like in the SOL methods (see section 6.2). Different approaches should be considered and tested by NSOs, according to the census specific needs, and based on quality information gained at the Source and Data Stages (see Chapter 4 and Chapter 5), including from metadata, consultation with data suppliers, expert opinion, pre-processing checks and comparison with external sources.

6.5 Editing and Imputation

266. The quality assessment at the Source and Data Stages (Chapter 4 and Chapter 5) will inform whether the administrative data has undergone or require editing (to deal with incorrect/implausible values) and/or imputation (to deal with missing values) and what type. Editing and imputation may be required both on the single source and the integrated data.

267. In the Austrian register-based census seven 'base registers' are used to provide basic information on the respective census topics (e.g., the Central Population Register determines the number of people with their main residence in Austria, and their basic demographic characteristics). These base registers are supplemented by eight 'comparison registers', which are mainly used for validation purposes. That is, one base register is selected to provide the information for a certain census variable, and the comparison registers are used to confirm these values (see Schnetzer et al 2015). However, in some cases the comparison registers also provide data that are either fully or partly missing in the base registers. The combined dataset from the base and comparison registers, called the Central Database (CDB), is enhanced with imputations for item non-response and implausible values, which creates the Final Data Pool (FDP). Quality is assessed throughout, from metadata and contact with suppliers (e.g., to understand the reliability of the data for the intended purpose and how they dealt with missing or implausible values), to checks on the raw data (e.g., for item non-response), and checks on the register-based output through comparison to an independent external source (Statistics Austria 2019).

268. Three imputation methods have been applied in the Austrian register-based census: deterministic editing, statistical estimation (including hot-deck and logistic regression) and statistical matching. For example, hot-deck imputation has been used to impute legal marital status and involves aggregating individuals into groups ('decks') by attributes which are strongly correlated with the target variable. The marginal distribution of the target variable within a deck (with existing values) is used to impute the target variable in the corresponding deck (with missing values). In the final assessment of data quality in the FDP, a quality indicator for the imputation is computed.

269. Schnetzer et al (2015) suggest the use of classification rates to evaluate different imputation models. This involves applying the imputation method to already existing data and compare the results of the imputation process with the true values for each unit. The classification rate is derived as the ratio between the values that match and the numbers of all compared units. The classification rate is like a hit ratio and can be used for categorial and numerical values.

270. Chambers (2001, cited in Schnetzer et al 2015) describes five quality-related properties that imputations should fulfil:

- **predictive accuracy:** the imputed values should be as 'close' as possible to the true values
- **ranking accuracy:** the imputation process should preserve the order of imputed values (for attributes which are at least ordinal)
- **distributional accuracy:** the imputation procedure should preserve the distribution of the true data values
- **estimation accuracy:** the lower-order moments of the distribution of the true values should be reproduced by the imputation process (for scalar attributes)
- **imputation plausibility:** the imputation procedure should result in imputed values that are plausible.

6.6 Recommendations

- As mentioned in Chapter 5, the accuracy and completeness of linkage variables should be assessed prior to linking data from different sources.
- Overall linkage rates (number of links over total number of records) and false positive/negative rates should be assessed and reported against. Thresholds for linkage error should be pre-determined and the trade-off between minimising false positive or false negative links should be considered.
- Coverage error in the statistical population register should be assessed and accounted for. This can be achieved using comparisons with other sources, including via the 'signs of life' methodology and using surveys (which can be specifically designed to adjust for over- and under-coverage).
- The choice of signs of activity indicators (or decision rules) when constructing statistical registers should be informed by an assessment of quality at the Source and Data Stage, and different methods (and underlying assumptions) should be tested.
- Models can be used to both assess the quality of administrative data for the purposes of enumerating population units (against a dataset that is taken as the 'truth') and to determine when and how to use the administrative data for this purpose.
- When deciding between sources when the same attributes are available in them, different approaches should be considered and tested, according to the census specific needs, and based on quality information gained at the Source and Data Stages.
- The quality of editing and imputation should be assessed both on the individual sources and on the integrated data, and different imputation models should be assessed.

6.7 Case studies

6.7.1 United Kingdom: measuring linkage quality when replacing a census variable with administrative data

271. The decennial Census of England and Wales is conducted by the Office for National Statistics (ONS) to enumerate the population, and record population and household characteristics. ONS are looking to replace a census question on "number of rooms" for the 2021 Census using administrative data. Some elements of this work remain to be completed; however, process quality has been tested through using 2011 Census data.

272. The 2011 Census asked the two questions "how many rooms are available for use only by this household?" and "how many of these are bedrooms"? The responses are used to derive occupancy rates by comparing the rooms/bedrooms that are available to the "rooms/bedrooms required". A negative occupancy rating implies there are fewer rooms/bedrooms available than required by the household (overcrowding). The information allows central and local government to develop appropriate housing policies and plan future housing provision. The quality of response to the 2011 Census number of rooms question, as measured by Census Quality Survey, was considerably lower (67 per cent) than that of the number of bedrooms question. This and the motivation to reduce respondent burden led to ONS considering administrative data as an alternative way to meet the information need. Also, following a consultation for the 2021 Census, overall users indicated using Valuation Office Agency (VOA) administrative data would have a positive impact on their work. The Valuation Office Agency (VOA) is an executive agency of Her Majesty's Revenue and Customs (HMRC). It has been responsible for banding properties for Council Tax since the tax was first introduced in the early 1990s.

6.7.1.1 Measuring process quality

273. The variable unique property reference number (UPRN), a unique alphanumeric identifier for every spatial address in the UK, was used to link VOA and census data. To ensure high quality linkage, the uniqueness of this variable was measured in both VOA and census data. In census data, responses with a non-unique UPRN were treated as if they have missing number of rooms as these cases cannot be linked to the administrative data with certainty. Duplicate UPRNs in census data occurred if two or more different census addresses were assigned the same UPRN. An example of this might be where a ground floor flat and a first-floor flat are assigned the same UPRN but have different census address identifiers. This is likely to be due to matching error when address records in census are linked to the address frame, as the method includes an element of probabilistic matching.

274. In VOA data, records with a non-unique linkage variable - which accounted for 1 per cent - were excluded. This is like duplication in the 2011 Census data. Other VOA records are 'cleaned' prior to data linkage which account for approximately 3 per cent. This included removing records that hadn't been assigned a UPRN by Geoplace (0.2 per cent) and records with duplicate UPRNs (0.3 per cent).

275. The linkage rate of 2011 Census responses with administrative data by UPRN is also measured. This is an important method in quality assessment as unlinked records are the main reason for the missing variable 'number of rooms'. Excluding wholly imputed households (non-responses) and non-unique records, 96 per cent of 2011 Census households linked to VOA property data.

276. Prior to edit and imputation, the distributions of unlinked and linked census records are compared on key household variables, such as accommodation type and number of usual residents. A similar comparison was carried out for missing number of rooms in linked and unlinked datasets. Although some differences in distributions were observed, crucially the number of available "donor" records where number of rooms was non-missing was sufficient: when broken down by a single household variable and by local area the number of donors always exceeds those with missing values. The edit and imputation processes have been tested for ten local authorities with the highest percentage of missing number of rooms.

277. Further research is required to establish if administrative data can provide comparable information with the 2011 Census category for dwellings "above or within commercial building" and to understand how to address the small proportion of records where VOA property type and census accommodation type appear contradictory.

6.7.2 Spain: Use of administrative data in the construction of a census data base for the 2021 Spanish Census: the 'signs of life method'.

278. The 2021 Population Census in Spain is viewed as a microdata database with approximately 47 million records, one for each resident.

279. For census enumeration, administrative records contain a vast amount of relevant information, despite being collected by authorities for purposes unrelated to population counts. Administrative sources are linked together to create a population register to identify who is residing in the country, and therefore produce population estimates.

280. The basic structure for the population count is based on *Padrón*, the Spanish population register where all residents in each municipality of Spain are recorded. Individuals are required to register in the municipality they live in and, as there are many advantages, residents normally do register.

6.7.2.1 Process Quality

281. When using the *Padrón* for census purposes, an adequate statistical register must be constructed. After receiving the original *Padrón* database referenced at January 1st of each year, a statistical treatment is carried out. Some assumptions are made around the presence of foreign nationals in Spain whose registrations have expired or about to expire. Moreover, population figures are statistically corrected to ensure they meet the 'usual resident' definition, for instance, applying the twelve-month residence concept. In short, population figures are obtained from *Padrón* but they are not exactly the result of counting the registered population as some individuals are eliminated while others are added.

282. From the whole population register, approximately 1.7 per cent of the individuals are erased (excluding them from the population counts) while approximately 0.15 per cent are added and included in the population counts.

6.7.2.2 'Signs of life' method

283. To identify which individuals are usually resident, the 'signs of life' (SOL) method is applied. All individuals are analyzed within the available administrative data sources and the movements are detected in *Padrón* for the months following the reference date. The four key administrative sources used to assess SOL are as follows:

1. Tax Agency and local tax files.
2. Social Security Insurance Database: Includes individuals with insurance and beneficiaries (employees and pensioners).
3. Labour market-related sources including:
    a. Unemployment National Service Database that provides a job seekers file to include individuals unemployed.
    b. Social Security Affiliation Registers that provides affiliation information of the employed population.
    c. Public Aids Database that provides information about benefits recipients.
4. Central Registry for Foreign Nationals Database that provides supplementary information about foreign nationals living in Spain such as date of application for residence permit, licence or rejection of residence permit, expiration dates residence checks etc.

284. Through using the SOL method, individuals who reach the threshold of presence signals within administrative data will be identified as 'active' and will be included in the population counts. In addition, individuals not meeting the threshold will appear 'inactive' and will not be included. These SOL from administrative data can also be compiled at individual and household level, therefore information is available about how many household members are 'active'.

285. Furthermore, for both Spanish and foreign nationals, the movements in *Padrón* are taken into account in the following months after the reference date. There are certain movements that require the direct intervention of the person, or a residence check made by a municipality, which generates a high probability of the person to be residing in Spain at the reference date. Also, other movements are good indicators of a person not residing in Spain at the reference date. These movements can be used to identify individuals that are 'usually resident'.

286. For minors, it is considered to be a sign of presence that an adult in the same household themselves shows signs of presence. Minors who do not meet this requirement are excluded from the population. The possibility of using information about enrolled students in official studies is currently being analysed.

6.7.3 New Zealand: Process quality assessment when including administrative enumeration in the New Zealand 2018 Census

287. For the first time, the New Zealand 2018 Census dataset includes administrative (admin) records for direct enumeration of people who were missed by the census field collection, replacing the use of 'substitute' imputed records in previous censuses. These administrative enumerations are drawn from a New Zealand resident population derived from administrative data which has already been assessed for input quality, and quality limitations are known (Gibb et al, 2016; Stats NZ, 2017). The administrative enumerations are then only added to the census dataset if individuals were in New Zealand on census night and were census non-responders (Stats NZ, 2019a). This case study focuses on how we measure and assess the accuracy of our linking and statistical modelling processes.

288. The administrative enumeration methodology is designed to achieve a final census dataset with the highest possible coverage of the census target population. We are most concerned with eliminating potential over-coverage due to the use of administrative records, both nationally, and for local areas, and expect that this will result in some remaining under-coverage. Linkage processes are designed to ensure that administrative records are added only for people who have not already responded to the census. Statistical models have been developed to manage the known quality limitations of the administrative resident population.

289. At the highest level, the process of including administrative records in the 2018 Census dataset involves linking the census responses with the administrative data, selecting administrative records to be included into dwellings (with families and households created), and included at a small geography only (with no relationship to dwelling and no family or household created). At each stage of the process we assess the quality of the process and decide if the methodology is acceptable.

290. The link between the census responses and administrative population is achieved using a fully automated probabilistic linkage process designed to minimise false positive linkages (Stats NZ, 2019b). The quality of the linkage process is assessed through estimating the false positive and false negative link error rates. The false positive estimate is derived from manually checking a small sample of linked records, and the false negative estimate is based on an approach developed by Choi, 2019 in which we estimate the missed matches from a subset of the census forms that met the criteria for inclusion in the administrative data with a high level of certainty (so we should be able to match). The overall link rate achieved is high (97.7 percent) with false positive links estimated at 0.6 +/-0.3 per cent and false negative matches estimated as 1.21 per cent (Stats NZ, 2019c). The high link rate coupled with low error rates give us confidence that the linkage is of acceptable quality. We are mostly concerned with false negative matches and the potential for them to impact on accuracy by contributing overcoverage to the 2018 Census dataset, so we include an adjustment for these false negatives later in the administrative enumeration process.

291. The methodology used for allocating administrative records to dwellings, and the subsequent step into small geographic areas is designed to balance the quality limitations of the administrative data against the quality requirements of the 2018 Census dataset (Stats NZ, 2019a). Again, the driving dimension of quality is accuracy. To assess the quality of the

administrative data for inclusion in census, we developed statistical models (using current and previous census data) to predict reliability of administrative data for representing an entire household (Gath & Bycroft, 2018; Stats NZ, 2019d). We use census data for training and assessing the models (assuming census responding households represent the truth A model score is generated for each administrative household, representing how reliable the administrative data is for the entire household in a given dwelling. A model score cut-off determines which of the non-responding administrative households will be added to the census dataset. The model is assessed using receiver operating characteristic (ROC) curve analysis and by analyzing performance metrics such as sensitivity, specificity, and precision (Stats NZ, 2019d) across a range of model score cut-offs. We see medium to high scores on the sensitivity measure (the proportion of correct administrative households that we include) across the full range of cut-off scores giving us confidence we are able to correctly identify most of the high-quality administrative households. In contrast, we see greater variability in the specificity measure (the proportion of incorrect administrative households that we exclude) indicating we are also likely to include some administrative households without the correct membership.

292. With the remainder of the available administrative population, we undertake two adjustments prior to selecting records for inclusion in the census dataset to ensure we are not introducing people to the census file who should not be included. We first adjust for potential overcoverage in the administrative population (using a strict 'signs of life' approach) and then adjust for duplication caused by missed links between the received census forms and the administrative data. A model similar to that used for inclusion of households is applied, predicting the likelihood that the administrative meshblock reflects a person's true usual residence meshblock, and people with scores greater than a cut-off are included.

293. Much of the quality assessment process involves determining where to set the model cut-off scores – considering relevance, accuracy, coherence, and interpretability of the methodology and data produced. The cut-off for inclusion of administrative records in dwellings has been set as a balance between strict criteria of obtaining exactly the same people in the household as we observe in the census, and including administrative households that reflect similar adult–child patterns as the census, even if we cannot guarantee that all household members are the same. The cut-off for inclusion of administrative records in small geographic areas once again represents a trade-off; between maximising the use of administrative data to improve national demographic counts and minimising the number of individuals enumerated in the wrong area.

294. The quality assessments outlined have several limitations due to subjectivity in judgements, statistical assumptions, and challenges with the underlying administrative data. The linkage error assessment of false positive links is dependent on the quality of judgements made by clerical reviewers, while the false negative link assessment relies on the assumption that the records used in estimation are representative of those not eligible. The modelling assessments are also limited by the subjectivity in setting an appropriate model cut-off score, robustness of underlying assumptions such as census response data representing the truth (which extends into assuming no within household non-response), and the lack of information

available for determining when administrative data is incorrect. Future work on process quality assessment will include further methodological development, testing of assumptions, and exploration of alternative quality assessment tools for these processes.

6.7.4 Italy: The combined use of survey and register data for the Italian Permanent Population Census count

6.7.4.1 From door-to-door enumeration to the Permanent Population Census

295. The Permanent Population and Housing Census (PPHC) has been designed based on Istat (Italian National Institute of Statistics) modernization program, which places the integrated system of statistical registers at the core of statistical production. The role of field surveys in such system is to feed registers, in the broad sense of assessing their quality and integrate information that is missing, incomplete or of insufficient quality.

296. If the 2011 Census, though being register-assisted, was still a conventional census, comprising an exhaustive field-collection [15] , the PPHC is based on an upside-down relationship between field enumeration and registers, where register data are supplemented by field data collection.

297. At the core of the PPHC is the Population Base Register (PBR, in Italian Registro Base degli Individui), whose main administrative source are the Local Population Registers of Italian municipalities. Together with the Statistical Base Register of Addresses (SBRA) and with the thematic registers on education and employment, it provides the basis for the production of population census data, while ad hoc surveys are used to measure coverage errors of the PBR and to collect data for variables non-replaceable (or only partially replaceable) through the registers.

298. More precisely, two separate sample surveys (Areal survey and List survey) are conducted annually in self-representatives municipalities (i.e. with > 17,800 inhabitants), and once in 4 years, according to a rotation scheme, in non-self-representatives ones, for a yearly total of about 1,500,000 households (of which 450,000 for the Areal survey and 950,000 for the List survey).

299. In the Areal survey, a sample of addresses and/or enumeration areas (depending on the quality of addresses in a given municipality) drawn from the SBRA is canvassed “blindly” (as in conventional censuses) in order to enumerate every household.

300. The List survey, based on a sample of households drawn from the PBR, is conducted with a mixed mode technique (CAWI, CAPI, CATI), with a first phase of so-called “spontaneous response”, and a second phase of field follow-up on non-respondents by enumerators. For each non respondent household a pre-coded “outcome” is registered in the survey monitoring system at the end of the field-work.

301. The same questionnaire is used in both surveys (except for the List of household members, which in the List survey is pre-filled with PBR data) and includes not only partially

[15] Municipal Population Registers were used to guide field-enumeration i.e. as enumeration lists to mail out questionnaires, while other administrative sources integrated into the Additional List of Auxiliary Sources were used to correct the list under-coverage i.e. to enumerate people usually resident but not yet registered.

or non-replaceable variables, but all the hypercubes variables with the purpose of using the information collected to test the quality and the coverage of data already available in registers.

6.7.4.2 The combined use of register and survey data for assessing and correcting coverage errors of the PBR

302. With the aim of producing the population count, survey data are used to correct PBR data within a Dual System Estimation model aimed at estimating coverage errors of the register. If in the traditional census a Post Enumeration Survey (PES) is often used to measure the census undercount (with the PES being the second 'capture' while the census itself is the first 'capture'), in the PPHC the PBR represents the first 'capture' whilst the annual sample surveys and the 'administrative signs of life' represent the second 'capture'. Furthermore, differently from a typical PES, aimed at measuring under-coverage, in the PPHC design the second 'capture' aims at measuring and correcting both under-coverage and over-coverage of the PBR.

303. On the field, the second 'capture' is two-folded, with the Areal survey used for measuring the under-coverage error of the PBR, and the List survey used, together with information on "administrative signs of life" derived by the Register of Usually Residents according to administrative sources (AIDA), for measuring the over-coverage error of PBR. As a result of this process, the population count is finally obtained by applying correction coefficients for under-coverage and over-coverage errors to individual data in the PBR.

304. More precisely, through the linkage with the PBR, the Areal survey allows to estimate the number of individuals usually resident in the municipality who are not included in the PBR.

305. On the other hand, again through the linkage with the PBR, the List survey allows to estimate the number of individuals included in the register who are no more usually resident in the municipality. To this aim, non-respondent households are classified according to their "coverage status" based on the outcome registered in the survey monitoring system.

306. However, since the survey itself might be affected by under-coverage errors, failing to reach all individuals actually usually resident, a further step is undertaken before calculating the over-coverage rate. Within the subset of 'potential over-coverage' individuals (individuals still present in the municipality according to the PBR and not found on the field), a distinction is made based on 'signs of life' in the municipality tracked down in AIDA. Non respondents to the List survey are thus 'recovered' if they show strong (i.e. of at least 8- months) "signs of life" in the same municipality where they are recorded in the PBR; while individuals lacking such 'signs of life' in the municipality are confirmed as the actual register over-coverage. The 'signs of life' considered for this purpose are the following: being public servant, private employee or self-employed; receiving a retirement pension; attending school (including pre-primary) or university; receiving any unemployment benefit or the basic income; being a fiscally dependent family member of an individual with strong signs of life.

307. The correction coefficients to be applied to individuals in the PBR are obtained through the following steps:

a) calculation of the raw non-weighted rate of under-coverage per each profile[16] as the ratio between the newly enumerated individuals (i.e. individuals not expected according to RBI) and the total number of individuals enumerated

$$p_{ij,under} = \frac{Newly\ Enumerated_{ij}}{Total\ Enumerated_{ij}}$$

b) calculation of the raw non-weighted rate of under-coverage per each profile as the ratio between individuals expected according to the PBR and not found at the survey (or not 'recovered' according to AIDA) and, at the denominator, the same individuals plus individuals expected according to the PBR and enumerated at the survey (or 'recovered' according to AIDA)

$$p_{ij,over} = \frac{Expected\ and\ not\ found_{ij}}{Expected\ and\ not\ found_{ij} + Expected\ and\ Enumerated_{ij}}$$

c) calculation of the raw coverage corrector

$$corr_{ij} = \frac{1 - p_{ij,over-coverage}}{1 - p_{ij,under-coverage}}$$

d) calculation of direct and indirect estimates. Direct estimates calibrated for over and under-coverage for each profile are first calculated for sampled municipalities. The calibration process binds the survey sample weights to the known population totals, derived from PBR, for each profile. Indirect estimated are then calculated in order to reduce direct estimates' variability for sampled municipalities and to calculate estimates for non-sampled municipalities

e) calculation of the average corrector 2018-2019[17]. For each estimation domain and separately for over and under coverage, the average of 2018 and 2019 correctors is calculated, weighted with the respective demographic sizes. The estimate of the 2018-2019 average corrector is therefore obtained as the ratio between the weighted averages of the estimates of the over-coverage corrector and the under-coverage corrector.

6.7.4.3 Population count as a result of PBR correction

308.At the end of the process, a 'weight' is attached to each individual in PBR, which 'corrects' for the coverage errors of the register estimated for that municipality. The weight applied to residents in the register will be equal to 1 if PBR, for a given municipality, is affected by neither over-coverage nor under-coverage errors (or if the two errors compensate each other).

[16] All individuals who have the same profile in the municipality i.e. the same citizenship ('Italian' or 'foreign') get the same corrector value.

[17] Due to insufficient stability of the estimates between 2018 and 2019, an average population corrector of the 2018 and 2019 data has been adopted for each estimation domain.

309. If the estimated under-coverage of PBR is greater than the estimated over-coverage, the corrector applied to each individual of PBR will be higher than 1 and the total population will result higher than that of PBR.

310. Conversely, if PBR's estimated under-coverage is lower than the estimated over-coverage, the corrector applied to each PBR record will be lower than 1 and the total population will be lower than that of PBR.

311. Following the validation of the population count, the data collected both in the Areal and the List survey is used in conjunction with PBR data and data from the thematic registers on employment and education, using predictive statistical models, to produce data on education, foreign country of citizenship and labour force status.

Chapter 7. Output Stage

312. This chapter provides a guide to the quality considerations, tools and processes for the measurement of census output quality where estimates are produced through the integration of administrative data sources into the census design (also see UNECE 2018, Chapter 9). Section 7.1 covers the output quality dimensions on which an assessment should be made and section 7.2 provides details of addition tools and processes that can be used to assess quality against the dimensions.

313. While measuring output quality moves beyond the quality of the sources *per se*, producing high quality estimates using administrative data is the ultimate goal. As such, these Guidelines would not be complete without considering output quality. At the same time, it must be emphasized that all the preceding quality stages contribute towards the quality of the outputs. As such, in the case of a combined or full administrative data-based census methodology, a census design which is informed by the rigorous assessment of quality at the source, input and process quality stages will ultimately result in high quality outputs (also see KUMUSO, Quality Framework for Multisource Statistics, 2019 WP1 for quality indicators, measures and methods for assessing output quality).

314. Measuring output quality cannot be reduced to the estimation of overall uncertainty of the estimate (the accuracy dimension); rather, it should include an assessment across all other quality output dimensions. The introduction of administrative data will likely lead to gains in some dimensions and losses in others. Achieving the right balance across the quality dimensions is therefore the key to best meeting user needs. Whilst not comprehensive, this chapter aims to cover the output quality dimensions and some of the key quality tools and processes used to assess them.

7.1 Output quality dimensions

7.1.1 Relevance

315. Relevance refers to the degree to which the census outputs meet the needs of users in terms of both coverage and content. Data are relevant when they relate to the issues users care about most. This dimension may require NSOs to adjust the direction of their programmes over time, as needed. However, assessing relevance is subjective because it often depends on varying user needs. The challenge, therefore, for a census programme is to balance any conflicting user requirements and to go as far as possible towards meeting the most important needs within resource and other constraints (UNECE 2015). Section 7.1.6 provides details on meeting used needs and balancing quality dimensions.

316. Various tools and approaches can be used to assess relevance, including the use of user needs surveys and consultations; user satisfaction surveys; by building user feedback mechanisms into the census process; and by analysing the usages of census data (see UNECE, 2018 p28).

7.1.2 Accuracy and reliability

317. The accuracy of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure. More simply put, accuracy is the proximity between an estimate and the unknown true value. It is usually characterized in terms of error in statistical estimates and is traditionally broken down into bias and variance. In a census context, variance applies in situations where a portion of the questionnaire is used for a sample of persons or households, or where only a sample of records is processed; or can be introduced during the processing stages (e.g. probabilistic imputation and linkage – see chapter 6). Accuracy can also be described in terms of measurement and representation error (as described throughout the guidelines).

318. Reliability is the degree of closeness of initial estimates to subsequent estimated values (the concept is listed by the ESS together with accuracy; however, it is also related to comparability - see below). Administrative data, by nature, can be subject to improvements in accuracy over time (e.g. coverage can improve, as lagged registrations and de-registrations become available; and the quality of measurements can improve also). Therefore, an NSO can make use of "new" data to improve the census statistics, revising previous estimates. However, this needs to be balanced against user needs with respect to revisions. Methods for assessing the accuracy of census outputs are provided in the sections 7.2.1 and 7.2.2 below.

7.1.3 Timeliness

319. Timeliness refers to the lapse of time between the period to which the census data refer (e.g. Census Day) and the date of publication of the data. A combined or register-based census often allows for census estimates to be produced in a more timely and frequent manner than a traditional decennial census – indeed, this is one of the greatly-hailed advantages of census transformation. In light of this, the timeliness of estimates that can actually be produced should be a key quality consideration, and thus improvements should be made to this aspect wherever possible. The timeliness of the data themselves is an important determinant of the timeliness of the output, thus linking back to the quality Stages discussed in the preceding chapters. There is often a trade-off between timeliness and accuracy. It may be the case that different users will have different views on the balance between the two, and as such they may not have the same view on the effect of improved timeliness vis-à-vis accuracy (see section 7.1.6).

320. Several straightforward timeliness metrics can be found within the literature. Quantitative indicators can be applied to measure the time lag for the final results e.g. between data collection, data acquisition, data linkage and publication of statistics. For example, the overall timeliness may be calculated as the time from the end of reference period to receiving administrative data supply, divided by the time from the end of reference period to publication date, multiplied by 100 per cent (Eurostat ESSnet KOMUSO 2016; Eurostat 2014; Eurostat 2013; UNECE 2018).

7.1.4 Coherence and comparability

321. The ESS Quality Framework defines coherence and comparability as the adequacy of statistics to be reliably combined in different ways and for various uses and the extent to which differences between statistics can be attributed to differences between the true values

of the statistical characteristics. The ESS Quality framework and the UNECE 2020 Recommendations expands the definition to include 'the degree to which the census information can be successfully brought together with other statistical information within a broad analytical framework. Comparability can be seen as a special case of coherence, where coherence is the degree to which data that are derived from different sources or methods, but refer to the same topic, are similar, while comparability is the degree to which data can be compared over countries, regions, subpopulations and time.

322. Measuring the extent to which estimates produced using administrative data are internally and externally coherent and comparable is a centrally important aspect of output quality for all census types, including those which make use of administrative data. Such estimates should be coherent with the known characteristics of the population, longitudinally, across geographies and population characteristics (see section 7.2.2). In addition, it is important to assess the extent to which census integrated statistics are internationally comparable and to communicate this to users.

7.1.5 Accessibility and clarity

323. Accessibility is defined generally as the ease with which users are able to access the statistical information. Within the context of output production this includes ease with which the existence of the data and metadata can be ascertained by the user as well as the suitability of the format and/or medium through which this information can be obtained. Clarity relates to the availability of any supplementary information or metadata that may be necessary of help the user to interpret and understand the accompanying published data. The concept of a 'clarity' is essentially the same as 'Interpretability'. Section 7.2.6 provides details on quality reports and metadata that should be accessibly and understood by users.

7.1.6 Meeting user needs and balancing quality dimensions

324. Whether or not administrative data are used in statistical production, assessing the overall quality of estimates produced should take into account each of the quality dimensions. This includes not only the accuracy dimension – the aspect which is most often reported in relation to survey methodologies – but also the remaining quality dimensions. In a census context, the overall quality of estimates is thus about establishing the balance across the quality dimensions which best meets the needs of census users. To achieve this, it is necessary to consult users throughout the census design process, but also to give them access to the general information and specific metadata they need to appraise quality decisions and feedback on quality assessments undertaken by statistics producers. As such, quality reporting and quality metadata are essential (see section 7.2.6). In addition, the continuous improvement of input and process quality will ensure that output quality also improves. The former will be aided by the implementation of the necessary supplier feedback mechanisms (section 7.2.3) and the latter through independent expert review of methods (sections 7.2.4 and 7.2.5).

7.2 Further tools and processes

7.2.1 Assessing the accuracy of population estimates (coverage error)

325. In several countries the overall accuracy of census population estimates has traditionally been carried out based on a Dual-System Estimation framework (DSE) which involves conducting the traditional census (i.e. taking a census 'stock' at one point in time); following this up with a large post-census coverage survey (also at one point in time); and then relying on the DES which uses capture-recapture methods to estimate under- and over-coverage (O'Hare 2019). These estimates could then be adjusted based on administrative data on deaths, births and migration flows, for each year between decennial censuses. Alongside this, in some cases (e.g. the UK 2011 census) have carried out small post-census surveys, where data are collected on all census questions and then matched to census responses, in order to measure respondent error.

326. For some of the census types and use cases described in Chapter 2, the traditional methods for determining overall coverage and quality are still applicable. However, new or revised methods are necessary in the case of population estimates produced primarily from administrative records, as is the case with a combined or full administrative data-based census. These methods, including the use of dependent interviewing and the "signs of life" methodology, as were described in chapter 6. This continues to be an area of significant interest across NSOs, with ongoing developments across a number of countries (see The Survey Statistician, 2020, Vol.82, 27-39 for a summary of new and emerging methods).

7.2.2 Demographic analysis: comparison with alternative sources

327. Demographic analysis (DA)[18] can be applied to assess the accuracy and to understand the coherence and comparability of census outputs. DA involves systematic comparisons, establishing thresholds of acceptability and understanding any significant discrepancies. As such, it cannot be carried out without the conceptual research at the Source Stage or the validation and harmonization work at the Data Stage. It may also require multiple sources to be combined in order to meet the target population at the Process Stage.

328. The census estimates which integrate administrative data are validated against alternative sources – e.g. survey data, previous census data or alternative sources. When using DA, it is important to keep in mind that estimates in two sources can be different across different sex-age or other breakdowns. These differences could be caused by different target populations, different reference dates or population changes (when comparing to historical census data), by conceptual differences and variations in classification between the variables being compared across sources, and/or by differences in sampling, collection methods and approaches to data processing. As such, any such comparisons must be made in light of the results of the assessment at the Source and Data stages.

[18] See O'Hare (2019) for an introduction to the method and its limitations.

Box 10: Demographic analysis in Spain

In Spain, the precensal file (the FPC) is constructed based on the Spanish population register (Padrón) by applying a 'signs of life' method in order to enumerate the census population. The population figures obtained in the FPC are then compared at the minimum geographical level with the official population counts, with the main objective of detecting and correcting possible under- and over-coverage problems.

To ensure the quality of the FPC population figures, population is disaggregated by the most relevant demographic variables and compared for each level of the variables: sex, age (year by year), nationality type (Spanish/foreign) and nationality (disaggregated by countries). These micro comparisons help to establish the consistency of common variables.

Analysis of specific sub-populations is carried out to check for possible over-coverage problems. The most significant differences between the precensal file and the official population counts are due to the administrative nature of Padrón, as it is not a statistical register but an administrative one and, as such, requires processing to add and remove units as necessary (for instance adding births or removing deaths).

On the other hand, to avoid possible under-coverage in the precensal file, all people listed in each of the available administrative sources (e.g. tax files, social security files, unemployment files etc.) that have not been found in Padrón on the reference date, are checked. If there is strong evidence that a person is actually residing in Spain (given their presence in several administrative sources) but is not registered in Padrón, this person is incorporated into the population of the FPC.

A common example of this situation is people who have been removed from Padrón some months before the census reference date, 1 January, and who then appear again shortly afterwards, e.g. in February. This is the case, for example, with foreigners whose registration has expired and for which the renewal takes a few months to be completed.

Box 11: Demographic analysis in Canada

For the first time in 2016, the Canadian census programme gathered income information solely from administrative data sources. The estimates produced with these data were compared, to the extent possible, with other data sources. Comparison analysis focused on various topics including individual income by source, coverage issues, conceptual and processing differences, and regional differences. Given the sensitivity of most income indicators to such methodological differences, however, users should use caution when comparing 2016 census income estimates to those produced using other household income surveys, administrative data or earlier census data.

329. Having considered the assessment of individual data sources at the Source and Data stages and of sources combined into statistical registers at the Process stage, it is possible to make professional judgements about whether or not differences found through DA are within an acceptable margin. This will vary from country to country and thus it is recommended that such standards are locally developed.

7.2.3 Supplier feedback mechanisms and data quality incentives

330. The continued improvement of census estimates which integrate administrative data relies on the continued improvement of the administrative data collected by the administrative authority who supplies it (including the various organisations that might supply data for an administrative register, such as authorities of the municipalities). Achieving this requires adequate feedback mechanisms between the supplier and the statistical producers and the existence of the right kind of incentives for both the administrative authorities collecting the data and the individuals whose data they collect.

331. It is often the case that a supplier of the data, is also a user and will thus have an interest in the quality of the census results, which can support the relationship between the NSO and the supplier. Communication between the NSO and the various stakeholders was discussed in detail in Chapter 4. Good communication mechanisms will contribute towards closing the gap between operational and statistical quality, hence ensuring that the quality of the data used in the census, and the estimates they produce, continuously improve.

332. To support improvements in quality, the NSO can also work with the supplier to develop suitable tools, systems and standards (e.g. online interfaces, clear definitions, agreed areas of best practice, etc) to improve the collection, processing and transmission of data.

7.2.4 Independent review of methods

333. Independent review of census design and methods will encourage the continued improvement of quality, i.e. achieving the best balance between quality dimensions to meet the needs of users. Such reviews should be carried out by population and methodology experts.

334. In August 2018, Stats NZ established a panel of experts to provide advice and guidance to Stats NZ on the methods used in creating the 2018 Census dataset, as well as to users on the quality of the resulting dataset. The panel endorsed the statistical approaches used for

including administrative enumerations in the dataset, and concluded that the inclusion of those records improved the coverage and accuracy of population counts for the core demographics: age, sex, place of usual residence, and ethnicity (2018 Census External Data Quality Panel, 2019).

335. Similarly in the UK, the external Methodological Assurance Panel has three aims: 1) to provide external, independent assurance and guidance on the statistical methodology underpinning 2021 census estimates and those based on administrative sources, 2) identify significant gaps and risks in methods and make suggestions for mitigation and 3) review administrative data methods and contribute to their continuous improvement (UKSA 2018). Panel review will take place between 2018 and 2023.

7.2.5 Sensitivity analysis

336. As well as having population experts reviewing the overall method, quality will be improved by engaging experts in an analysis of particularly concerning topic areas or quality decisions throughout all quality stages, which we will call sensitivity analysis. Sensitivity seeks to establish the extent to which the method used is able to "count a population within a geographic region or demographic group", which "can be used to understand bias in census data, and plan for the next census by identifying the groups most difficult to count" (Stats NZ 2019e, p.5).

337. Statistics New Zealand engaged external providers to assess both the methods used to add people to the 2018 Census dataset and the fitness of the dataset for three important use cases, including determination of electoral boundaries. A sensitivity analysis of the methods used to add people to the 2018 Census file found that the threshold for inclusion in meshblocks had the most impact on who was included in the census file and that the threshold used was a sensible balance. Further sensitivity analysis determined that 2018 Census data was robust for the purpose of determining electoral boundaries and the electorate boundaries drawn using census counts were not likely to be impacted by the choice of threshold for adding administrative enumerations at the meshblock level (Stats NZ, 2019d, Stats NZ, 2019e). This was an important finding in support of the quality of the census dataset.

7.2.6 Quality reports and metadata

338. Within the last QA stage, a report should be produced to document the results of quality assessment and assurance throughout the census production. This report should include information against each QA stage as well as communicate to users where and how each quality dimension was considered. In order to enable the producers and users of statistics to appraise and feedback on quality decisions and determine whether the right balance has been achieved across the quality dimensions, sufficient metadata around quality assessment is necessary.

Box 12: Quality metadata in Spain

In the work leading up to the first administrative-data--based census in Spain, an extra categorical variable providing an assessment of data quality based on the origin of each value is being developed, to provide users with a variable-specific quality indicator (Pérez Julián, Casaseca and Argüeso Jiménez 2018). As previously noted, in Spain a population statistical register is created by linking the population administrative register (Padrón) with multiple administrative sources. This can be visualised as a huge matrix in which the census variables are considered columns and each person is presented by a row, so the matrix cells would contain particular values for each individual per variable. In order to help users understand the quality of census data, for each census variable another categorical one will be added to inform of the quality of each cell value. As explained below, this categorical variable is intended to inform users on quality either directly or indirectly.

The initial proposal to develop this quality measure for each cell is based on the type of methodology or source used to fill each cell value (see Table 7). Typically, a cell value derived from an up-to-date administrative source has the highest quality and one derived through deterministic imputation the lowest. In this way the quality of each cell value can be understood by users in an indirect way.

Additionally, the quality measure for each cell value depends not only on the nature of the underlying source and methodology, but also on the rest of characteristics of each individual. For instance, where a 20 year old person is missing values for the variable legal marital status and industry of his/her main and these are deterministically imputed to 'single' and 'Accommodation and food services' respectively, the chances are that the first imputed value is much more reliable than the second one. The relation between age and legal marital status is likely to produce good deterministic imputation estimates while is not the case when imputing a value for industry. Several such rules have been developed to inform the quality of imputations based on known individual characteristics.

Therefore, another proposal is a more direct way would be to provide a quality punctuation variable for example in a scale from 1 to 4 where 1 would be the highest quality and 4 the lowest one in order to help users in the understanding of how 'good' or 'bad' an imputation can be considered.

By both mechanisms, the indirect one or direct one, offer enormous potential in the assessment of output quality in two dimensions: by variable and by unit or subpopulations. It is proposed that all users should have free access to these quality variables in the census microdata release for 2021 (approximately 10 per cent of the whole census product) and would have specific methodological notes with explanations.

*Table 7: Initial proposal of categories indicating source quality by type**

DATA TYPE	DESCRIPTION	QUALITY
DS	Information provided by direct sources up-to-date.	Highest
DSN	Information provided by direct sources but not up-to-date.	
CS	Past census information.	
PI	Probabilistic imputation.	
DI	Deterministic imputation.	Lowest

**adapted from Pérez Julián, Casaseca and Argüeso Jiménez 2018, p.4)*

7.3 Case studies

7.3.1 Portugal: quality assessing the population register

7.3.1.1 Background

339. The *Census Admin* project (short for *Census with Administrative Data*) is part of the framework for the development of a National Data Infrastructure which includes Statistics Portugal's (SP) strategy of data integration, from several sources, to respond to an increasingly complex society with new expectations towards statistics.

340. Central to the project is the creation of a Resident Population Dataset (object type statistical population dataset, SPD), covering a set of characteristics – geographical, demographic and socio-economic – of the resident population in Portugal. SP´s goal is to report population statistics from the SPD from the 2021 Census onwards.

341. The SPD prototype was built in 2015 with reference to the 2011 population. Meanwhile, four new annual editions were created, with annual reference dates from 2015 to 2018.

342. For each annual edition, the consistency of the SPD results is evaluated by systematically comparing it against population estimates and known population characteristics. Additionally, comparisons with census' test results have been considered to measure SPD´s results quality.

7.3.1.2 SPD population counts by geographical level

7.3.1.2.1 Evaluate 2018 SPD results from national to regional geographical level

343. The resident population in Portugal, estimated through administrative data by the SPD, for 2018, is 10 300 502 persons, representing a relative deviation of +0.2 per cent when compared to the 2018 Population Estimates (PE) released by SP. The PE provide the official figures of the annual resident population in Portugal, using cohort components and the population census concept. Its calculations are based on the natural and migratory demographics, with information from: live births, deaths, emigration and immigration estimates.

344. The national level results obtained in the *Census Admin* project are very optimistic, considering the different assumptions, methodologies and distinct sources of these two types of statistical production: SPD and PE. Consistently, across all the annual editions of the SPD, relative deviation between these two sources is less than 0.5 per cent (under or over coverage).

345. At a regional level (NUTS II), the 2018 SPD-PE relative deviation oscillates from -0.4 (Centro) to 3.5 per cent (Algarve); Lisbon Region with -0.1 per cent.

346. The results of the Portuguese SPD are also promising at the municipality level: for 2018, more than 76 per cent of the 308 municipalities present levels of under or over coverage within 5 per cent, when comparing to the PE; it should be noted that in 64 municipalities of the country, the relative deviation SPD-PE is under 1 per cent (under or over coverage). Only a small number of municipalities (15), mainly low populated, show relative differences greater than 10 per cent (higher or lower).

347. Combined with the geographical distribution, the SPD captures part of the demographic and socio-economic dimensions. For example, the SPD-PE relative differences in the age structures of PE are very small for most age groups and across all SPD editions (main differences occur in elderly people).

7.3.1.2.2 Evaluate 2015 SPD results at a local geographical level

348. Comparisons have also been carried out at a lower geographical level, the parish or Local Administrative Units – level 2 (LAU2). As detailed below, the 2016 Census Test (2016 CT), on September, 26th, contributed to evaluate the 2015 SPD results (reference date 31 December) at the LAU2 level.

349. The analysis of the results of the 2016 CT showed that in 4 of the 5 parishes of the sample, where it was possible to guarantee exhaustiveness on data collection, 2015 SPD estimated more population than that which was actually enumerated. The relative deviations varied from -14.1 per cent to -5.7 per cent. Overall, the population counts in the 2016 CT, when compared to the estimated 2015 SPD, has a deviation of -8.8 per cent.

350. In order to evaluate how the 2015 SPD estimates are close to the reality observed in the field, microdata from 2016 CT was linked to 2015 SPD results and for those who matched (about 80 per cent), their characteristics were compared. For place of usual residence, for e.g., 90 per cent of respondents lived administratively in the same LAU2, i.e., the LAU2 where they were enumerated was the same as that registered in 2015 SPD (quite satisfactory considering the 9 months' time lap between the CT and SPD reference date).

351. If we take the place of usual residence at the municipality level as a basis for comparison, the equality rates are globally around 93 per cent, since 3.2 per cent of the individuals in the 2016 CT matched to 2015 SPD lived administratively in another parish of the same municipality.

7.3.1.3 Final observations

352. The focus of this work is to assess the quality of the Portuguese SPD to estimate the resident population.

353. We explored the consistency quality dimension and for that purpose we showed results of several comparisons: with the population estimates, disaggregating by geographical level (national to regional total) and with the census tests (finer geographical level).

354. The set of administrative information currently integrated in the SPD has a high potential for the transition to a registered-based or combined census model. At a national and regional level, the consistency of the SPD results are huge, however, at a lower geographical level, census test showed that there is still room to improve SPD quality estimates. SP so investing in more robust estimation methods and review 'signs of life' rules. Nevertheless, although the population counts at parish level present some differences, the structure and characterization of the parish population given by the 2015 SPD is very consistent with that collected in 2016 CT.

Chapter 8. Conclusions and recommendations

355. Administrative data can be used across the different census methodologies and to support all stages of the census process, including constructing an address frame, supporting field operations, enumerating the population, collecting census variables, quality assurance, editing and imputation, modelling and estimation. Their use can provide more frequent and timely statistics about the population; improvements in accuracy and reliability; and significant reductions in costs and respondent burden.

356. However, there are significant quality challenges to assess and overcome before an administrative source can be used in a census. Most significant among these is that administrative data have, in general, not been collected for the purpose of a census. As such, the NSO may have little control over the concepts and definitions used; the target population; the collection, processing and quality assurance procedures; and the data methods, structures and systems used.

357. The Guidelines have set out Stages of quality assessment, set against a number of quality dimensions, with associated tools and indicators to guide the user through the process of assessment. The application of the Guidelines should help readers to make decisions about the use of administrative data in the census, whilst supporting a process of continuous assessment and improvement. Throughout the Guidelines, a number of proposals and recommendations have been made, which are summarized below.

8.1 Recommendations

i. The NSO should **identify the administrative sources** that may be relevant to their census, set against different use cases. It is important to **set out what the expected or required outcomes of using the source would be**, against which an assessment of relevance can be made. This could include improvements to the efficiency of the census operation in terms of reductions in cost and respondent burden; improvements to the quality of the census; or the delivery of new or enhanced census outputs. Central to such assessment is setting out what the administrative source needs to deliver in terms of the target population and the required measurements from this population for the census use case. Chapter 2 of the Guidelines and the case studies across the other chapters provide examples of how administrative data have been used in several different countries.

ii. The **relationship between the NSO and the administrative data supplier** is of critical importance (Chapter 4). This should be supported by robust mechanisms of communication, written agreements and an excellent understanding of the needs of both parties. There must also be an agreed legal basis for the supply and use of the data. To help build the relationship and secure a data supply, the NSO should identify areas of benefit to the supplier. This could be with respect to feedback mechanisms to help the supplier better understand their data, through collaborations on areas of common interest, or by helping the supplier (through the use of their data in the census) to support

the wider good. Of course, feeding back on possible quality concerns with the data has the added benefit of facilitating ongoing quality improvements.

iii. **The NSO should engage with the supplier to gain an in-depth understanding of the data source**. This should translate into the creation of clear and comprehensive metadata about the administrative source. The metadata will provide a useful reference both for the census and for any other statistics that might benefit from use of the source. Chapter 4 provides details of the metadata that should be collected, along with various references to the relevant literature.

iv. Since administrative data are generally not collected for the needs of the census, it is important for the NSO to **understand and assess differences between the required populations, concepts, definitions and time-related dimensions**. More generally, a thorough assessment of the coherence and comparability of the administrative source, along with its limitations across the various quality dimensions, is essential. This includes the linkability of the source if this is a requirement for use in the census. This assessment will inform the processing stages, including mapping and derivation, editing and imputation, and the linkage and integration of sources (where decisions are made between and across sources based on their quality) (Chapter 6).

v. The NSO must **understand any restrictions and challenges to acquiring and integrating an administrative source into the census** (Chapter 4). This could include resources and costs; risks associated with the Supplier's ability to deliver on time to the required quality; and whether the use of a source is acceptable to the public and users of census data. In this respect there are important trade-offs that the NSO must consider. Specifically**, the value of the administrative source must be assessed against its usefulness for the census, set against the effort and risks of acquiring and using the data**.

vi. The NSO has limited control over the collection and processing of the administrative data, which can be subject to changes in population coverage and the measurements from the population over time. This can be due, for example, to legal, policy, procedural or system changes affecting the data and/or their delivery (Chapter 4). **The NSO must therefore assess and manage a level of risk**. The risk should be managed by working with the Data Supplier on potential or planned changes; by being flexible and responsive to change; and by reducing reliance on any single data source or item where possible, whether through the use of other data sources or by adapting processes/methodologies (Chapter 6).

vii. **It is important that the public and data users understand how and why administrative data are being used in the census** (Chapter 4). The NSO should, therefore, be transparent about the use; providing a clear justification of the benefits set against any risks and costs (i.e. a strong proportionality case exists). This can be achieved through **good communication, including the publication of the procedures and policies in place that support the effective use and protection of data**.

viii. The inclusion of administrative data sources in census production should be preceded by adequately resourced feasibility research which provides a 'proof of concept' for the planned integration of administrative data into the census production. It is advisable to

carry out a number of test runs (using real data) well in advance of the main census to ensure any unforeseen issues are identified, allowing enough time to correct or adjust the methods, processes or systems (as described in Chapter 5 and Chapter 6).

ix. **Expert review** (working with data suppliers and subject experts) and **comparisons between sources and over time are important to identify any quality concerns** with a source or register. The use of well-designed surveys (linked to administrative data or registers) can be particularly important in identifying and adjusting for coverage and measurement errors (Chapter 5 to Chapter 7).

x. **The NSO should record and publish the results of the quality assessment and assurance throughout the census production**, including the Data, Process and Output Stages. This will enable producers and users of the census to appraise and provide feedback, supporting an ongoing dialogue. This is important to ensure that users understand the strengths and limitations, and can help determine whether the right balance has been achieved across the dimensions of quality (Chapter 7).

xi. The NSO should **develop its own quality assurance framework and strategy**, supported by clear and comprehensive documentation and training procedures. These Guidelines provide a useful basis to support this, along with the reference material and case studies within the Guidelines. The strategy should build the continuous assessment and improvement of administrative data into the plans and procedures for the census. This should include the communication links between the NSO, users and the data suppliers.

8.2 Areas for further development

358. The Guidelines have focused on the assessment of the quality of administrative sources for use in censuses, while providing some information about the processes used to integrate and transform data to improve quality. The quality of census outputs that use administrative data is also covered briefly. However, the Guidelines do not provide a **wider total error framework** or a model for how the error from each source translates into the error in the final census estimates, taking account of the changes in quality due to processing (which can reduce or increase error).

359. The development of such a model that takes account of all sources of error is partly addressed by the total error framework adopted by Statistics New Zealand (Reid et al., 2017). The framework builds on Li-Chun Zhang's (Zhang 2012) extension of the Total Survey Error (TSE) paradigm (see Groves and Lyberg 2010; Biemer 2010). It has three phases covering: 1) an assessment of the single sources, 2) an integrated data set assessment; and 3) an estimation and output assessment. The work of the ESSnet KUMUSO on the quality of multisource statistics (WP1 Quality, 2019) also provides a useful framework for assessing the quality of statistical outputs based on multiple sources (survey and administrative data).

360. This could be an area for further development and international collaboration with **a specific focus on how such a framework can be applied to censuses**. This could include examining how a total error framework or model can be developed and used to assess the

quality of census outputs based on multiple sources. It could also include work to understand how the impact (and compounding impact) of various errors across the stages of the census can inform decisions about the best overall statistical design for the census.

361. Finally, the Guidelines have focused on the assessment of administrative data, but there are **other sources of commercial data that present opportunities for use to improve or enhance census statistics** (e.g. geospatial data, mobile phone data). The quality Stages, dimensions, tools and indicators within these Guidelines are to a great extent applicable to sources beyond administrative data. This too could be an area requiring further international work, with a specific focus on whether and how the tools and techniques for assessing the quality of such sources for use in the census differ from those identified here.

References

Abbott, O. (2009). 2011 UK Census Coverage Assessment and Adjustment Methodology. Population Trends 137, Autumn 2009. https://www.ons.gov.uk/file?uri=/census/2011census/howourcensusworks/howwetookthe2011census/howweprocessedtheinformation/coverageassessmentandadjustmentprocesses/censuscoverageassessment_tcm77-189757.pdf

Abbott, O., B. Tinsley, S. Milner, A. C. Taylor and R. Archer (2020). Population statistics without a Census or register. Statistical Journal of the IAOS, 36(1), 97-105.

Argüeso, A. (2019). Population and Housing Census in Spain will be fully register-based. https://ec.europa.eu/eurostat/cros/system/files/mr-antonio-argueso_census-will-be-fully-register-based_es.pdf

Asamer E.-M., F. Aztleithner, P. Ćetković, S. Humer, M. Lenk, M. Moser and H. Rechta (2016). Quality Assessment for Register-based Statistics - Results for the Austrian Census 2011. Austrian Journal of Statistics Vol. 45, No. 2, pp. 3-14

Australian Bureau of Statistics (2009). ABS Data Quality Framework, May 2009, cat. no. 1520.0, ABS, Canberra, https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Quality:+The+ABS+Data+Quality+Framework

Brown, J., C. Bycroft, D. Di Cecco, J. Elleouet, G. Powell, V. Račinskij, P. Smith, S.-M. Tam, T. Tuoto, and L.-C. Zhang (2020). Exploring developments in population size estimation. Survey Statistician 82, 27-39.

Cerroni, F., G. Di Bella and L. Galiè (2014). Evaluating administrative data quality as input of the statistical production process. In: Rivista Di Statistica Ufficiale. ISAT Available from: https://www.istat.it/it/files/2014/10/Articolo-7_Evaluating-adfministrative....pdf

Chambers, R. (2001). “Evaluation Criteria for Statistical Editing and Imputation.” National Statistics Methodological Series 28: 1–41. https://statswiki.unece.org/pages/viewpageattachments.action?pageId=263229091&preview=/263229091/278038163/3.5_Austria_NL_Moldova%20draft_3_5.docx

Chieppa, A., G. Gallo, V. Tomeo, F. Borrelli and S. Di Domenico (2018). "Knowledge discovery for inferring the usually resident population from administrative registers" in Mathematical Population Studies, Pages 92-106,. https://www.tandfonline.com/doi/abs/10.1080/08898480.2017.1418114

Choi, H. (2019). Adjusting for linkage errors to analyse coverage of the administrative population. Statistical Journal of the IAOS, 35(2), 253-259.

Cornell University Research Data Management Service Group (2020). Guide to writing "readme" style metadata. https://data.research.cornell.edu/content/readme

Crescenzi, F., G. Sindoni and D. Zindato (2014). Lessons learned from the 2011 Italian census and innovations leading towards a continuous census. Note by the National Institute of Statistics of Italy, presented at the UNECE/Eurostat Group of Experts on Population and Housing Censuses, Sixteenth Meeting, Geneva, 23-26 September 2014, https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2014/mtg1/15_E_Italy.pdf

Daan Zult, P., P de Wolf, B. Bakker and P. van der Heijden (2019). A linkage error correction model for population size estimation with multiple sources. http://isi2019.org/proceeding/2.STS/STS%20VOL%203/#p=335

Daas, P.J.H., J. Arends-Tóth, B. Schouten and L. Kuijvenhoven (2008). Quality Framework for the Evaluation of Administrative Data. Paper presented at the European Conference on Quality in Official Statistics. http://www.pietdaas.nl/beta/pubs/pubs/21Daas.pdf

Daas P., S. Ossen, R. Vis-Visschers and J. Arends- Tóth (2009). Checklist for the Quality evaluation of Administrative Data Sources. The Hague: Statistics Netherlands. https://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluation-administrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683-4477be82ee60

Daas, P., S. Ossen, M. Tennekes, J. Burger and F. Cobben (2012). Input Quality of administrative data (BLUE-ETS WP4). Presented at Quality 2012. Available at: http://www.pietdaas.nl/beta/pubs/pubs/Q2012_Session23_presentation.pdf

Eurostat BLUE-ETS (2011). List of quality groups and indicators identified for administrative data sources. Retrieved April 23, 2020, from https://mtennekes.github.io/downloads/publications/BLUE-ETS_WP4_Del1.pdf

Eurostat (2013). Use of administrative and accounts data in business statistics https://ec.europa.eu/Eurostat/cros/system/files/SGA%202011_Deliverable_6.5.pdf

— (2014). ESS Handbook for quality reports. Available at: http://ec.europa.eu/Eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf

— (2017). European Statistics Code of Practice, revised edition 2017 https://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice

— (2019) Quality Assurance Framework of the European Statistical System. Version 2.0. https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646

— (2020). European Statistical System Handbook for Quality and Metadata Reports, https://ec.europa.eu/Eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf

Eurostat ESSnet KOMUSO (2016).Checklist for Evaluating the Quality of Input Data https://ec.europa.eu/Eurostat/cros/system/files/essnet_wp1_report_final_version4.pdf

— (2019). Quality Guidelines for Multisource Statistics https://ec.europa.eu/Eurostat/cros/system/files/qgmss-v1.1_1.pdf

Eurostat ESSnet MIAD (2014). MIAD deliverable B2, B3 - Quality check list for the Source phase [Data]. Available from: https://ec.europa.eu/Eurostat/cros/content/miad-deliverable-b2_en

Falorsi, S. (2017). Census and Social Surveys Integrated System. Note by the National Institute of Statistics of Italy, presented at the UNECE/Eurostat Group of Experts on Population and Housing Censuses, Nineteenth Meeting, Geneva, Switzerland, 4–6 October 2017, https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2017/Meeting-Geneva-Oct/WP23_ENG.pdf

Gallo, G., A. Chieppa, V. Tomeo and S. Falorsi (2016). The integration of administrative data sources in Italy to increase Population Census data availability. Note by the National Institute of Statistics of Italy, presented at the UNECE/Eurostat Group of Experts on Population and Housing Censuses, Eighteenth Meeting, Geneva, Switzerland, 28-30 September 2016, https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2016/mtg1/CES_GE.41_2016_3E_Italy_rev_2.pdf

Gath, M. and C. Bycroft (2018). The potential for linked administrative data to provide household and family information. Retrieved from www.stats.govt.nz.

Gibb, S., C. Bycroft and N. Matheson-Dunning (2016). Identifying the New Zealand resident population in the Integrated Data Infrastructure (IDI). Retrieved from www.stats.govt.nz

Harron, K., H. Goldstein, and C. Dibben (2015). Introduction. In K. Harron, H. Goldstein, & C. Dibben (Eds.), Methodological Developments in Data Linkage. New York: John Wiley & Sons.

Iwig, B., M. Berning, P. Marck and M. Prell (2013). Data Quality Assessment Tool for Administrative Data

International Organization for Standardization (2015). Quality management systems — Fundamentals and vocabulary. ISO 9000:2015(en) https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:en

Lavigne, M. and C. Nadeau (2014). A Framework for the Evaluation of Administrative Data. In Proceedings of Statistics Canada Symposium.

Lothian, J., A. Holmberg and A. Seyb (2019). An evolutionary schema for using "it-is-what-it-is" data in official statistics. Journal of Official Statistics 35, 137-165.

Oberski (2018).

O'Hare, W.P. (2019). Methodology Used to Measure Census Coverage. In: Differential Undercounts in the U.S. Census. Briefs in Population Studies. Springer, Cham

Office for National Statistics (ONS) (2016). Methodology of Statistical Population Dataset V2.0 https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/methodology/methodologyofstatisticalpopulationdatasetv20

— (2019a). High Level Statistical Design for the Transformed Population and Social Statistics System. https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP132-Statistical-Design-of-Future-Population-and-Social-Statistics-System.docx

— (2019b). Developing our approach for producing admin-based population estimates, England and Wales: 2011 and 2016. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/developingourapproachforproducingadminbasedpopulationestimatesenglandandwales2011and2016/2019-06-21#summary-and-next-steps

— (2020). Joined up data in government: the future of data linking methods. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/joined-up-data-in-government-the-future-of-data-linkage-methods#acknowledgements

Pérez Julián, M. P., C. Casaseca and A. A. Argüeso Jiménez (2018). Assessing quality in a register-based census. Paper presented at the European Conference on Quality in Official Statistics, Krakow

Reid, G., F. Zabala, and A. Holmberg (2017). Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ, Journal of Official Statistics, 33(2), 477-511. doi: https://doi.org/10.1515/jos-2017-0023

Rogers, N. and L. Blackwell (2020). A statistical quality framework for longitudinally linked administrative data on international migration. Available from: https://www.ons.gov.uk/releases/astatisticalqualityframeworkforlongitudinallylinkedadministrativedataoninternationalmigration

Schnetzer M., F. Astleithner, P. Cetkovic, S. Humer, M. Lenk and M. Moser (2015). Quality Assessment of Imputations in Administrative Data, Journal of Official Statistics, Vol. 31, No. 2, pp. 231–247, http://dx.doi.org/10.1515/JOS-2015-0015

Scholtus and Bakker (2013). ?

Shafer, G. (1992). Dempster-Shafer theory. Encyclopedia of artificial intelligence. http://fitelson.org/topics/shafer.pdf

Shipsey, R. and J. Plachta (2020). Linking with anonymised data- how not to make a hash of it. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/linking-with-anonymised-data-how-not-to-make-a-hash-of-it

Statistics Austria (2019). Quality assessment of administrative data - Documentation of Methods, http://www.statistik.at/wcm/idc/idcplg?IdcService=GET_PDF_FILE&dDocName=122178

Statistics Canada (2017). Statistics Canada's Quality Assurance Framework https://www150.statcan.gc.ca/n1/pub/12-586-x/12-586-x2017001-eng.htm

Statistics New Zealand (2017). Experimental population estimates from linked admin data: 2017.

— (2019a). Overview of statistical methods for adding admin records to the 2018 Census dataset
— (2019b). Linking 2018 Census responses to the Integrated Data Infrastructure
— (2019c). Dual system estimation combining census responses and an admin population
— (2019d). Electoral boundaries sensitivity analysis of 2018 Census data
— (2019e). Predicting the quality of admin location information for use in the 2018 Census
— (2019f). Population counts sensitivity analysis of 2018 Census data
— (2020). Guide to reporting on administrative data quality. https://www.stats.govt.nz/methods/guide-to-reporting-on-administrative-data-quality
UK Statistics Authority (UKSA) (2015a). Quality Assurance of Administrative Data - Setting the Standard.
— (2015b). Administrative Data Quality Assurance Toolkit. Version 1 January 2015. https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2015/12/Quality-Assurance-Toolkit.pdf
— (2018a). Code of Practice for Official Statistics (Edition 2.0).
— (2018b). Methodological Assurance Review panel – Census. Retrieved from: https://www.statisticsauthority.gov.uk/about-the-authority/committees/methodological-assurance-review-panel-census/
— (2019). Quality Assurance of Administrative Data (QAAD) toolkit.
— (2020). Ethics Self Assessment Tool. Available from https://uksa.statisticsauthority.gov.uk/about-the-authority/committees/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/
UNECE (1992). Fundamental Principles of Official Statistics. Available from https://unece.org/statistics/fundamental-principles-official-statistics
— (2014). A Suggested Framework for the Quality of Big Data. Available from: https://statswiki.unece.org/x/H4mZB.
— (2015). Conference of European Statisticians Recommendations for the 2020 Censuses of Population and Housing.
— (2017). Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys and Administrative Data Sources. Version 2.0. https://statswiki.unece.org/download/attachments/185794796/Quality%20Indicators%20for%20the%20GSBPM%20-%20For%20Statistics%20derived%20from%20Surveys%20and%20Administrative%20Data%20Sources_Final.pdf?api=v2
— (2018a). Guidelines on the use of registers and administrative data for population and housing censuses. http://www.unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20184.pdf
— (2018b). Annex F – Portugal Case Study in Guidelines on The Use of Registers and Administrative Data for Population and Housing Censuses. p.64-67.
United Nations (2009). Handbook on Geospatial Infrastructure in Support of Census Activities Studies in Methods: Series F No. 103, ST/ESA/STAT/SER.F/103 New York: United Nations Department of Economic and Social Affairs, Statistics Division https://unstats.un.org/unsd/publication/seriesf/Seriesf_103e.pdf
United States Census Bureau (2009). History: 2000 Census of Population and Housing. https://www.census.gov/history/pdf/Census2000v1.pdf
Vega Valle, J. L., A. Argüeso Jiménez and M. Pérez Julián (2020). Moving towards a register based census in Spain. Statistical Journal of the IAOS. 36(1): 187-192.
Yucel and Zaslavsky (2015).

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 66: 41–63, DOI: 10.1111/j.1467-9574.2011.00508.x.

Glossary of Terms

Accessibility: The dimension of quality that is defined generally as the ease with which users are able to access the data.

Accuracy: The dimension of quality that refers to the degree to which the information correctly describes the phenomena it was designed to measure. More simply put, accuracy is the proximity between an estimate and the unknown true value.

Address register: A register of residential addresses, often used for the purposes of creating enumeration areas comprising comparable numbers of dwellings. In cases of multi-occupied dwellings there can be more than one dwelling under a given residential address.

Administrative enumeration (New Zealand): The process of collecting data taken from an administrative source for the purpose of supplementing data recorded on questionnaires collected in a field enumeration.

Administrative data: Data held on registers and other administrative sources relating to information collected by government and/or other organizations primarily for administrative (not research or statistical) purposes, such as registration, transaction and record keeping, usually for the provision of public services.

Administrative (data) source: A data holding that contains information collected primarily for administrative (not research or statistical) purposes. Such sources include administrative registers (with a unique identifier) and possibly other administrative data without a unique identifier.

Administrative population: The population set of objects or units (e.g. people, dwellings, businesses) that is captured by the administrative source or register.

Administrative register: A systematic collection of unit-level data organized in such a way that updating is possible (where 'updating' is the processing of identifiable information with the purpose of establishing, bringing up to date, correcting or extending the register. Such registers are primarily used in an administrative information system in which the data are used in the production of goods and services in public or private institutions or companies. Administrative registers used for statistical purposes are normally operated by the state or jointly by local authorities, but some registers operated by private/commercial organizations may also be used.

Administrative unit: The units for which administrative data are recorded. These may or may not be the same as those required for the statistical output (which are referred to as statistical units).

Attribute: A socio-demographic or economic characteristic relating to an administrative or statistical unit for which information is required for the purpose of the census.

Benchmarking: Comparing data, metadata or processes against a recognised standard.

Big data: Large, often unstructured data sets that are available, potentially in real time, but which are difficult both to process efficiently and quality assure using traditional methods and technologies. The amount and variety of data available is growing rapidly, and such data sets are available in many formats, including audio, video, computer logs, purchase transactions, sensors and social networking sites. Some of these data are freely available on the web, whereas others are held by the private sector to which there may be no free access.

Census day: The date of the reference period for the census as a whole, irrespective of when the data are collected.

Census estimates: A term used by some countries to describe the census output data to reflect the fact that the published figures do not purport to be true counts and that there must always be some degree of uncertainty (however small) in the accuracy of the numbers.

Clarity: The dimension of quality that relates to the availability of any supplementary information or metadata that may be necessary of help the user to interpret and understand the accompanying data.

Classifications: Statistical classifications provide a set of related categories in a meaningful, systematic and standard format e.g. the NSO's standard for classifying occupations. Classifications are generally developed to support policy making and because of that, to organize and present statistics.

Coherence: The dimension of quality that refers to the degree to which data that are derived from different sources or methods, but refer to the same topic, are similar.

Combined census: A census based on a combination of data taken from administrative registers and collected on questionnaires.

Comparability: The dimension of quality that refers to the degree to which data can be compared over time and domain.

Daas hyperdimensions: High-level dimensions or 'views' of quality of an administrative source to be used for statistical purposes. The three key dimensions refer to: the source; the metadata; and the data itself.

Data controller: See 'Register owner'.

Data editing: The process by which data that exhibit errors, logical inconsistencies and spurious values are detected and corrected.

Data journey: the totality of the processes raw data is subject to from collection to their use in the production of statistics, much like the Generic Statistical Business Process Model (GSBPM).

Dempster-Shafer theory: A generalization of the Bayesian theory of subjective probability.

Derived variable: A new variable formed by using the data from other variables**.**

Dual System Estimation: A statistical method, based on a capture-recapture technique, applied to estimate the size of a population.

Estimates: The term is used in these Guidelines to refer to the statistics produced in census outputs, and reflects the processes undertaken by NSOs to adjust the input data to take account of under- or over-coverage, errors, missing counts and measures to control statistical disclosure.

Field enumeration: The process of collecting information on individual persons, households and/or housing unit covering the whole population (or a sample of it) using questionnaires

Frame: Any list, material or device that delimits, identifies, and allows access to the elements of the target population. A statistical register is a specific example.

Imputation: The process by which missing input data items are replaced with plausible and consistent values.

Input data: The data (sometimes referred to as 'raw data') derived from an administrative source, before any processing or validation by the NSO.

Input quality: The quality of the raw administrative data as it is supplied to the NSO by the administrative authority

Linkability: The ability to link data from several different administrative data sources to the same unit, usually by means of a unique identification number or code.

Measurement error: error in the measurement of variables or characteristics (e.g. age, gender etc). They include several types of error within variables including relevance (definition misalignment), mapping (errors in the re-classified measures due to poor equivalence between supplied and target classifications which may therefore require adjustments, e.g. through imputation) and comparability errors (errors between the re-classified and adjusted measures).

Meshblock: : The smallest geographic unit for which statistical data is collected and processed by Statistics New Zealand.

Metadata: Data that describe or define other data. This broadly refer to anything that users need to know to make proper and correct use of the real data, in terms of accessing, processing, interpreting, analyzing and presenting the information. Metadata include, for example, file descriptions, codebooks, processing details, sample designs and fieldwork reports. Metadata should be distinguished from 'Paradata' which generally refer to the details that describe the process by which the census data are collected, either from administrative sources or a field enumeration/survey.

Objects: In some of the literature (e.g. Zhang 2012), the term 'object' is used to refer to the units within an administrative dataset. The term is used to distinguish between units in the administrative data and the statistical units after this data has been transformed in some way. This is particularly relevant in cases where the unit (or 'object') in the administrative register differs from the target statistical unit. For example, where a tax register, where the units of a yearly tax returns (i.e. the same person may make several returns in one or multiple years), is converted into individual 'people'.

Output data: The processed data as it is used in statistical outputs.

Output quality: The quality of the processed data as it is used in statistical outputs.

Padrón (Spain): The Spanish population register, usually compiled for each Municipality.

Paradata: See 'Metadata'.

Periodicity: Within the context of the supply of administrative data, this is the time period between reference dates for consecutive input datasets. For the census more generally, it is the time between the dates of consecutive censuses (census days).

Population register: A statistical register and a frame of persons usually resident (however defined) in a given country. Additionally, it often provides some demographic characteristics of individuals.

Privacy Impact Assessment: A process which assists organizations in identifying and managing the risks to privacy arising from new projects, initiatives, systems, processes, strategies, policies and business relationships.

Process quality: The effect of changes to the quality of data being used for the purpose of the census during the processing of the raw data by the NSO.

Punctuality: The dimension of quality that relates, when referring to data, to the time lag between the planned (and often pre-announced) publication dates and actual publication dates. In the context of the administrative source, it relates to the time lag between the expected (or contracted) date of the delivery of the data to the NSO and the actual date of delivery.

Raw data: See 'Input data'.

Register: A systematic collection of unit-level data organized in such a way that updating is possible. Updating is the processing of identifiable information with the purpose of establishing, updating, correcting or extending the register.

Register-based census: A census where all data is collected from administrative registers. A census based on combination of data taken from registers and questionnaires is called a 'combined census'.

Register keeper: See 'Register owner'.

Register owner: The authority responsible for keeping and maintaining an administrative register (also referred to as the 'Register keeper' or 'Data controller'.

Relevance: The dimension of quality which, when referring to data, refers to the degree to which they meet the needs of users in terms of coverage and content. When referring specifically to data sources, the dimension refers to the degree to which such sources contain data that meets the needs of the NSO with respect to their intended use.

Reliability: The dimension of quality that refers to the degree of closeness of data values to earlier or subsequent data.

Representation error: error in the representation of the intended population units or objects (e.g. individuals or households in a census). They include errors relating to over and under-coverage (lack of alignment with target population), identification (errors in classifying a unit based on inconsistencies across multiple sources) and unit errors (errors in the statistical creation of statistical units of interest where they do not exist in any available data source).

Rolling census: An alternative approach to the traditional model of census taking by means of a cumulative continuous survey, covering the whole country over a period of time, rather than on a particular day. There are two main parameters to consider in a rolling census: (a) the length of the periodicity, which itself is linked to the frequency of updating required; and (b) the sample size, which depends on the budget and the level of geographical analysis required for dissemination.

'Signs of life': An indicator used to minimize the over-coverage of persons recorded on different administrative registers derived by applying strict criteria or 'activity rules' to ensure that only living individuals who are usually resident are included in the census estimates.

Source quality: The quality of administrative sources from which data is supplied to NSO for the purpose of the census.

Statistical disclosure control: The process(es) by which the raw data taken from an administrative source or collected in the field is modified during data processing in order to avoid the disclosure of information about identifiable individual persons or households.

Statistical register: A register processed for statistical purposes. A statistical register could be based on one or several administrative registers. Statistical registers are also referred to as 'secondary registers'.

Target population: The universe for which information is required. The target population is the set of the statistical units.

Test data: smaller supplies of data from an administrative source/register shared with NSOs for the purposes of feasibility research and the testing of systems.

Timeliness: The dimension of data quality that refers to the lapse of time between the period to which the data refer (in the case of census data this is usually Census Day) and the date of publication of the data. In the use of administrative data, timeliness also refers to the length of time between the date of the event recorded in the data source and the date when the data are delivered to the NSO.

Unit: The smallest entity to which any administrative data item refers. For the purpose of the census, units may refer to individual persons, households, buildings or dwellings.