Assessing the Quality of Administrative Data in a Census

Select Topics in International Censuses[1]

By Britta Schumacher

Released November 2022

INTRODUCTION

The use of administrative data sources (ADSs) in population and housing censuses (PHCs) continues to increase in many countries. This is motivated by the many benefits ADS integration can bring such as reduced cost and respondent burden, improved timeliness and frequency of results, improvements to quality, and greater flexibility to respond to user needs. In addition, there is increased support for ADS use in national statistical systems reflected through changes in legislation and public and stakeholder acceptability.

ADSs "contain information collected primarily for administrative (not research or statistical) purposes" (United Nations Economic Commission for Europe, 2018), are held by their collecting bodies—government departments, public bodies, and other organizations—and are used for the administration of taxes, benefits, and other services. Over the past few decades, new methodologies to transform ADSs, including population, business, address, and education registers, into statistical registers have been developed allowing for their use in PHCs.

This technical note builds on a previous note, "Using Administrative Data in a Traditional Census,"[2] and presents the various dimensions, tools, and indicators used to assess the quality of administrative data throughout the data life cycle. Throughout, this note provides guidance to National Statistics Offices (NSOs) to assess ADS quality across the census life cycle, ensuring that when ADSs are adopted in population and housing census operations, they do not compromise, but instead enhance census output quality.

QUALITY AND ERROR IN A CENSUS

Errors[3] in statistics produced using ADSs arise via representation and measurement errors (Figure 1). The first relates to errors in the coverage of population units (e.g., individuals or households), and the latter to errors in the measurement of characteristics (e.g., age/gender). Representation errors arise from differences in units supplied and target units, and include over- and under-coverage, identification, and unit errors. Measurement errors result from differences between the characteristics supplied in an ADS and the target characteristics for the PHC, and include relevance, mapping, and comparability errors (Zhang, 2012).

The range of possible errors is greater when an NSO uses ADSs or other big data sources because the NSO itself is not collecting the data, and thus, cannot control for quality in the collection process. Importantly, error in PHC statistics does not result from the administrative sources alone, but also from the integration of ADSs in the census design. Errors accumulate throughout the statistical process, and the resulting impacts on data quality must be weighed against the benefits ADS integration provides such as reduced cost and respondent burden (UNECE, 2021).

[1] This technical note is part of a series on Select Topics in International Censuses (STIC) exploring matters of interest to the international statistical community. The U.S. Census Bureau helps countries improve their national statistical systems by engaging in capacity building to enhance statistical competencies in sustainable ways. Any views expressed are those of the author(s) and not necessarily those of the Census Bureau.

[2] For information regarding the various types of ADSs that may be used to supplement a traditional census, a suggested framework for ADS integration and to understand the challenges and opportunities ADSs present, refer to <www.census.gov/programs-surveys/international-programs/events/training/select-topics-in-international-censuses/using_administrative_data.html>.

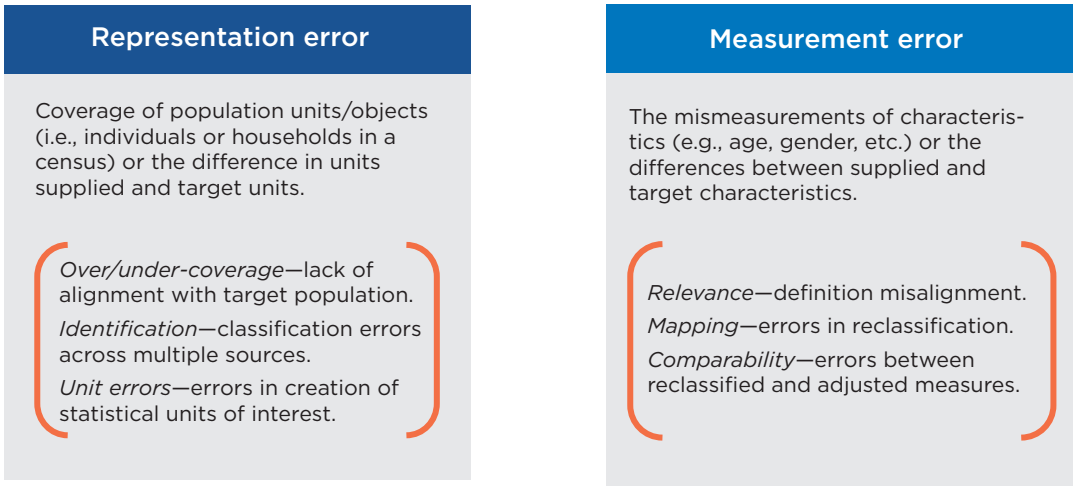
[3] Errors here are the "difference(s) between a final estimate and the true population parameter it represents" (UNECE, 2021 p. 19).

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

United States Agency for International Development
www.usaid.gov

United Nations Population Fund
www.unfpa.org

Figure 1.
Errors Resulting From Administrative Data Use in Census Statistics



Source: United Nations Economic Commission for Europe, 2021.

ADMINISTRATIVE DATA QUALITY FRAMEWORK

The quality of administrative data must be considered in relation to the ways data are collected and processed by the data suppliers (DSs) and by the NSO across the census life cycle—from collection, to processing, to output dissemination. Though statistical design and PHC development are never entirely linear, it is useful to think of quality assessment in a four-stage, sequential framework (Figure 2). The following section describes the four stages identified in Figure 2, with particular focus on stages 1 and 2.

Source Stage

As NSOs establish and maintain connections with administrative DSs, it is important for NSO staff to evaluate the quality of the administrative sources provided. At this stage, data are rarely accessible; instead, NSOs must seek out information about the ADS through exploratory meetings and other channels to determine whether to proceed with the data acquisition initiative or not. This stage is relevant not only to the first acquisition of administrative data (AD), but to every resupply thereafter, as it is entirely possible that the characteristics of resupplied data may differ from the original data acquired in concepts, classifications, collection methods, and processing procedures. The main quality dimensions and the tools/indicators used in the Source Stage assessment are detailed in Table 1.

Data Stage

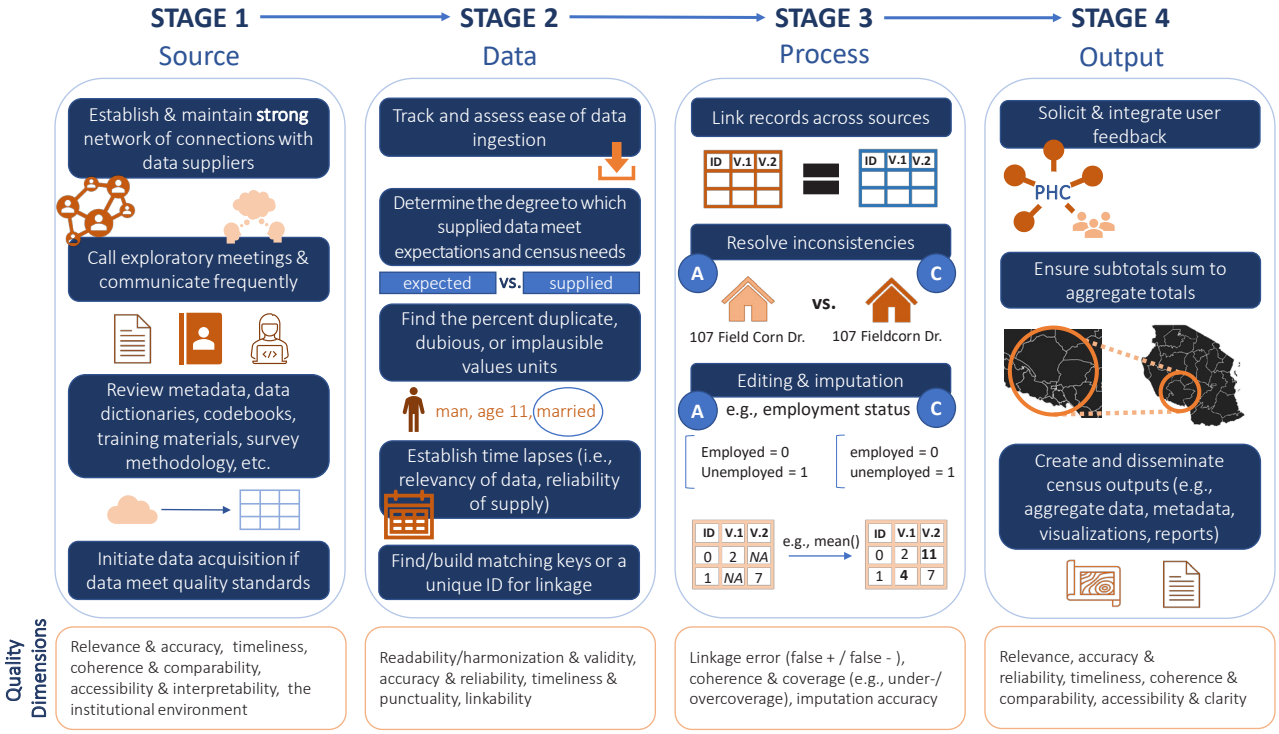
Once NSOs have established a network of connections amongst DSs, assessed the quality of the data each supplier could provide, and contracted with suppliers of quality AD, the data transfer and acquisition process can begin. At this stage, data are acquired and their quality must be assessed based on the expectations and requirements set forth in the Source Stage. Importantly, quality assurance (QA) at the Data Stage reveals the following to the NSO:

- Any changes or corrections that should be implemented in their contracts with DSs.
- All processing of AD needed for their use in the PHC design (e.g., statistical adjustments for error correction).
- The implications of errors in AD on final census outputs, which must be disclosed to data users in PHC metadata.

Together, the Source Stages and Data Stages provide an overall assessment of the input quality of AD. The main quality dimensions and the tools/indicators used in the Data Stage assessment are detailed in Table 2.

Figure 2.

Common Framework for Quality Assessment of Administrative Data Across the Census Lifecycle[1]



[1] This framework details the 5th process, "Quality Measurement and Assurance," identified in Figure 3 of "Using Administrative Data in a Traditional Census," at <www.census.gov/programs-surveys/international-programs/events/training/select-topics-in-international-censuses/using_administrative_data.html>.
Source: U.S. Census Bureau, adapted from the United Nations Economic Commission for Europe, 2011, 2018 and 2021.

Table 1.

Source Stage Quality Dimensions, Definitions, and Tools for Assessment

Quality Dimension	Definition	Tools and Indicators for Assessment
Relevance and accuracy	The degree to which the administrative data sources (ADS) meet the needs of the national statistics office (NSO) with respect to its intended use; the degree to which the data correctly describe the phenomenon they were designed to measure.	Availability of target variables; alignment of concepts, definitions, and classifications; potential for processing error by data suppliers (DSs).
Timeliness	The lapse in time between the period to which the information refers and the date when the data become available to the NSO; this lapse affects the data's relevance.	Frequency of data supply compared to population and housing census requirements.
Coherence and comparability	The degree to which ADSs can be successfully combined with other sources from the NSO; the degree to which data can be compared across geographies, time, and subpopulations.	Comparability over time and across space; presence of unique ID or combination of variables for linkage.
Accessability and interpretability	The ease with which NSOs can obtain and understand ADSs; the availability and clarity of metadata and documentation from DSs.	Restrictions for data access/use; public acceptability; ease of data transfer/receipt; clear and comprehensive metadata.
The institutional environment	The organizational and institutional factors that impact a DSs' capacity to supply quality and timely data.	Strength of relationships with DS; previous experience with DS; formal agreements; DS quality standards.

Source: United Nations Economic Commission for Europe, 2021.

Box 1.

Source Stage and the Memorandum of Understanding

Ensuring the initial and continued provision of administrative data sources (ADSs) requires sustained coordination, cooperation, and collaboration between data suppliers (DSs) and national statistical offices (NSOs). It is useful to work both at the practical level, creating task teams and building relationship networks, as well as at the formal level, establishing agreements such as Memorandum or Memoranda of Understanding (MoUs) with DSs. MoUs lay out the roles and responsibilities of the data provisioning agency and NSO, establish the objectives of the NSO in acquiring the DSs' data, and detail other elements of concern to agencies and data users (e.g., data security, ensuring confidentiality).

MoUs formalize the data exchange between administrative data providers and the national statistics system, and can reduce the myriad challenges (e.g., unpredictable data delivery timeframes, incompatible data formats, staff turnover) facing NSOs in collecting administrative data for statistical purposes. Building and openly publishing MoUs can also: increase public trust, facilitate automation of the data exchange process, streamline data dissemination and reduce response burden, and improve data quality.

One aspect of particular importance is the designation of responsibilities of both parties. Below is an excerpt from a 2017 MoU between Statistics Botswana (SB) and the Nonfinancial Institutions Regulatory Authority (NFIRA):

> 3.1 SB will advise NFIRA on policy, procedures, and regulations relating to the development of official statistics.
>
> 3.2 Parties will assist each other in collection, processing, compilation, analysis, publishing, disseminating, and archiving official statistics.
>
> 3.3 The parties will meet and communicate regularly to discuss matters of common interest.
>
> 3.4 The parties will always promote an environment of complementary and consistent approach, so far as that is consistent with their independent responsibilities.

For more information and guidance regarding MoUs and what elements should be included, refer to the Collaboratives' 2021 draft at <https://unstats.un.org/capacity-development/admin-data/docs/mou-guide-and-template.pdf>.

Source: Collaborative on the Use of Administrative Data, "Memorandum of Understanding" 2021 *draft*.

Process Stage

After NSOs complete a QA of the ADS, they must be processed and transformed by the NSO for use in the PHC. Common processes required for transforming ADSs include:

- Record linkage.
- Assessing coverage error in statistical registers/ADSs.
- Resolving inconsistencies across DSs (e.g., duplicates).
- Editing and imputation.

Each of these processes is informed by the insights gained in the Source Stages and Data Stages. At the Source Stage, for example, NSO staff may learn that there is no unique key/ID for linking data across sources, but do find a suite of variables that, together, allow deterministic linkages to be made. Alternatively, NSO staff may discover that exact matches are not possible and will instead move forward using probabilistic methods that utilize the relative similarity of records to match data across sources.

Though these processes are designed to improve the quality of data, and build outputs fit for and integrated with the PHC design, it is also true that they may introduce uncertainty in the data. This uncertainty may be introduced inadvertently (e.g., false matches, or non-matches) or purposefully (e.g., statistical disclosure control), and must be controlled so as not to impact the quality of census outputs.

Output Stage

While measuring output quality moves beyond the purview of ADS QA, it would be negligent not to include this stage because developing census outputs is the main

Table 2.

Data Stage Quality Dimensions, Definitions, and Tools for Assessment

Quality Dimension	Definition	Tools and Indicators for Assessment
Readability and validity	The degree to which provisioned data can be ingested by the national statistics office (NSO); the difference between data supplied and the metadata collected at the Source Stage, the reference period, etc.	Ease of data ingestion; variables supplied are correctly named/formatted; content and reference period match expectations established in the Source Stage.
Accuracy and reliability	The measurement and representation errors present in the administrative data; the closeness of initially supplied values to the subsequently resupplied values within a dataset.	Percentage of duplicate or dubious (i.e., involved in implausible relations) objects/statistical units; percentage of implausible or out-of-range values; longitudinal comparability.
Timeliness and punctuality	The lapse in time between the period to which the information refers and the date on which the data become available to the NSO; the difference between the expected and actual date of delivery.	Time between date of receipt of NSO/user and the reference period; time between date of receipt of NSO and the contracted supply date.
Linkability	The degree to which data can be linked across sources at the unit/record level; the degree of success affects both the data's accuracy and relevance.	Percentage of unique values (either unique IDs or a combination of variables); prevalence of biased distribution in accuracy indicators.

Source: United Nations Economic Commission for Europe, 2021.

Box 2.

Data Stage and Data Diagnosis in Rwanda

In June 2018, the National Institute of Statistics of Rwanda (NISR) published their "Guidelines for Quality Assessment of Administrative Data." In this document, NISR lays out its own framework for assessing administrative data source quality, including "data diagnosis," the process by which NISR can determine the quality of supplied data.

Data diagnosis involves both data profiling and data integration. The first involves NISR staff inspecting supplied data for errors, discovering inconsistencies, completing partial information, determining the representativeness of the records, assessing outliers, and removing duplicates. The latter is the process of merging and linking data across disparate platforms and data types to align records.

Source: National Institute of Statistics of Rwanda, "Guidelines for Quality Assessment of Administrative Data," 2018.

Box 3.

Process Stage and the National Employment, Unemployment, and Under Employment (ENEMDU) COVID-19 Survey in Ecuador

Officials in Ecuador have been able to improve the quality of identification data in ENEMDU-COVID-19 Survey by linking these data with the "Cedulados" database from the Civil Registry of Ecuador. Instituto Nacional de Estadistica (INE) begins processing the data and assessing the quality once "Cedulados" data transfers from the Civil Registry to the INE via the Government Data Network, a system built to facilitate the exchange of data under information security standards.

Processing occurs under nine macro-activities: 1. Profiling, 2. Correction, 3. Standardization, 4. Accuracy, 5. Identification of changes, 6. Coherence, 7. Uniqueness, 8. Integration, and 9. Pseudonymization, applied according to the data in-hand. Once these steps are complete the "Cedulados" and ENEMDU-COVID-19 data are linked and processed together.

The linkage of the "Cedulados" administrative data sources with the ENEMDU-COVID-19 data increased the quality of identification in the survey from 71 percent to 83 percent and enabled INE to produce timely and accurate information regarding the impact of COVID-19 on Ecuador's labor market and populace.

Source: United Nations Statistics Division and the Global Partnership for Sustainable Development Data, "Collaborative on the Use of Administrative Data for Statistics," 2021.

objective of integrating ADSs in a PHC. Higher quality census outputs result from rigorous assessments of quality at the Source, Data, and Process Stages. In addition to these stage-based impacts, the presentation of census data (e.g., data visualizations and tabulations) on output quality is key. NSOs must follow best practices in data visualization (e.g., color-blind friendly palettes, correct chart/graph for the data type) and output dissemination (e.g., language accessibility, rich metadata) to create and share quality outputs. The main quality dimensions and the tools/indicators used in the Output Stage assessment are detailed in Table 3.

ADS QUALITY RECOMMENDATIONS

Any PHC, regardless of ADS use, must assess the overall quality of census estimates, considering all four stages

Table 3.

Output Stage Quality Dimensions, Definitions, and Tools for Assessment

Quality Dimension	Definition	Tools and Indicators for Assessment
Relevance	The degree to which census outputs meet user needs in terms of both coverage and content.	Integrate user feedback into the census system (e.g., consultation, user needs/ satisfaction surveys).
Accuracy and reliability	The degree to which information captured correctly describes the phenomena measured; the closeness of estimates to subsequent estimates.	Measures of uncertainty introduced via statistical disclosure protection; post-census survey estimation.
Timeliness	The lapse in time between the period to which the census refers and the publication date of statistical products.	Time between collection/acquisition, linkage and publication; frequency of publication compared to user needs.
Coherence and comparability	The degree to which data derived from different sources/methods are similar; the degree to which data can be compared across geographies, time, and subpopulations.	Subtotals (e.g., regional estimates) should correctly sum to national totals; comparability across known population characteristics.
Accessibility and clarity	The ease with which users can access and understand statistical outputs; the availability of metadata for interpretation of outputs.	Clear and comprehensive metadata; extent to which data repository and published reports meet data users' ease-of-use expectations/ needs.

Source: United Nations Economic Commission for Europe, 2021.

Figure 3.
Recommendations for Quality Assessment of Administrative Data Across the Census Life Cycle

Source: United Nations Economic Commission for Europe, 2018 and 2021.

outlined above. The overall quality of estimates requires a balance across all quality dimensions and stages, informed by the needs of census users. A set of recommendations based on the stages and quality dimensions outlined above is outlined in Figure 3.

CONCLUSION

NSOs are increasingly integrating ADSs in PHCs, establishing the need for a common framework through which they can develop their own NSO-specific, QA framework. In this technical note, the four key stages in ADS QA—Source, Data, Process, and Output—are discussed, and key indicators to measure various quality metrics are outlined. ADS quality must be rigorously assessed at all stages to ensure that data, reports, and other PHC products are of the highest quality possible.

REFERENCES

United Nations Economic Commission for Europe, "Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices," New York and Geneva, 2011.

United Nations Economic Commission for Europe, "Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses," New York and Geneva, 2018.

United Nations Economic Commission for Europe, "Guidelines for Assessing the Quality of Administrative Sources for US in Censuses," Geneva, 2021.

Zhang, Li-Chun, "Topics of Statistical Theory for Register-Based Statistics and Data Integration," *Statistica Neerlandica*, Volume 66, Issue 1, pp. 41–63, 2018, <https://doi.org/10.1111/j.1467-9574.2011.00508.x>.

The Select Topics in International Censuses (STIC) series is published by International Programs in the U.S. Census Bureau's Population Division. The United States Agency for International Development sponsors production of the STIC series, as well as the bilateral support to statistical organizations that inform authors' expertise. The United Nations Population Fund collaborates on content and dissemination, ensuring that the STIC series reaches a wider audience.