مـركـز الإحـصـاء
STATISTICS CENTRE

Register Based Statistics

Methodology and Quality Guidelines – Guide No. (19)

July 2023

Contents

1.Introduction

Within the framework of the international recommendations and trends confirming the adoption of administrative data as the main source to produce statistics, and to achieve the objectives of the Statistics Centre – Abu Dhabi (SCAD). The objectives emphasize the adoption of the administrative data in producing statistics, SCAD carries out the necessary procedures to build and develop the statistical registers of all types. In this context, it relies on administrative data from government entities. Therefore, the integration process of the concerned data into comprehensive registers at the level of Abu Dhabi shall become much easier.
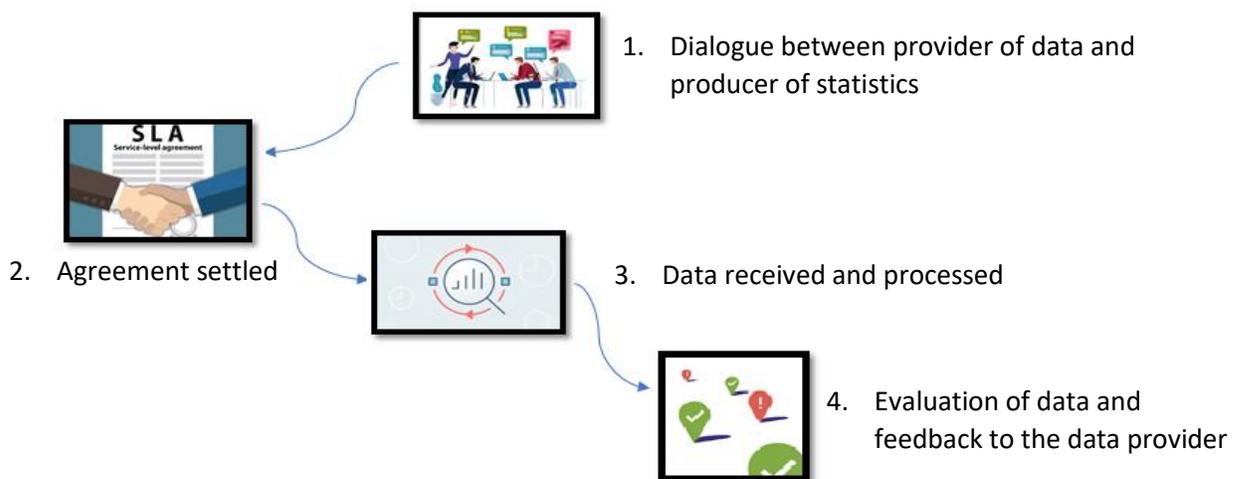
In addition, within the work frame concerned with building statistical registers covering multiple aspects in the emirate of Abu Dhabi, this guide provides general guidance on the process of building and preparing statistical register from administrative data.

2.Register based statistics

Administrative data are collected by governments or other organizations for non-statistical purposes to provide overviews on registration, transactions, and record keeping. These types of data are used to produce management information.

Administrative registers are one type of administrative data source and become statistical registers after passing through statistical processing to make it fit for statistical purposes. The steps of receiving admin data from entities:

Figure 1: Steps of receiving admin data

1. Dialogue between provider of data and producer of statistics
2. Agreement settled
3. Data received and processed
4. Evaluation of data and feedback to the data provider

1. *Talk to each other: must make relationships between entity and understand the needs and requirements for each entity.*

2. *Service Level Agreement (SLA): Preparing a service level agreement between entities enhances to providing administrative data, Meet the needs, and supporting the existing strategic partnership between the entities. In each SLA there is a list of variables. Also, documentation is needed for definitions, classifications, and time reference periods, etc.*

3. *Receive and process the data from the entity.*

4. *Check and do quality assessment (for example check number of variables based on SLA).*

2.1. Advantages and challenges of administrative data

Administrative data offers advantages in quality, cost, and time. It is sometimes considered more accurate than self-reported survey data. However, the purpose of collecting data is most often other than the purpose of producing statistics.

Nonetheless, administrative data does not come without a few potential challenges: access, merging, and quality. Accessing administrative data requires strong relationships with national and/or local authorities. In some cases, authorities may not be inclined to share the information. Once accessed, consolidating administrative data with other data often entails merging different databases together: this can be an extensive task when no common unique identifiers exist across the databases. Finally, while in some cases administrative data can provide high accuracy, in others, it may be badly reported, not exhaustive, or not at all existents. Not all governments have the same capacity to collect this information.

2.2. Advantages and challenges of statistical registers

The statistical register assumes a key role in statistical production. The importance of statistical registers is reflected through the following:

- Saving the time spent collecting data in comparison with other methods.
- Cutting off the costs related to the provision of statistics and data.
- Limiting the burdens imposed on community members and respondents.
- Including new additional topics and variables that may be covered by the data registers.

The challenges of statistical registers are:

- The concepts and definitions used as required in terms of laws and regulations may be unsuitable for statistical purposes and not conform to international standards. The records may contain insufficient data to provide the detailed information required to accurately classify the variables.

- Some of the data may not be accurately or carefully recorded and could be incomplete or inaccurate since those items of information will be of secondary interest or value to the primary administrative purpose for which they were designed.
- There could be legal restrictions or confidentiality provisions that restrict access to the records. This is usually the case with taxation records.

2.3. Vocabulary

The following table provides the main concepts and definitions used in the register-based statistics:

Table 1: Main concepts and definitions

Term	Definition
Administrative data	Administrative data means data collected by, or on behalf of, national and local authorities other than a producer of official statistics, for administrative purposes in conformity with legal bases other than statistical legislation. E.g., educational data, health data, tax data, pension data etc.
Administrative register	Administrative register is data collected for administrative purposes. Register is collected and maintained outside of the National Statistical System.
Statistical register	A statistical register is a register created for statistical purposes normally by statisticians. They are typically created by transforming data from registers and/or other administrative data sources. See Figure 2, page 6.
Statistical classification	A statistical classification or nomenclature is an exhaustive and structured set of mutually exclusive and well-described categories, often presented in a hierarchy that is reflected by the numeric or alphabetical codes assigned to them, used to standardize concepts, and compile statistical data.
Data processing	Processing covers a wide range of operations performed on data, including by manual or automated means. It includes the collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure, or destruction of personal data.

The statistical register shall be built based on the following foundations:

- Users' needs include the need to create a statistical register.
- The legislation: the availability of a legal reference for using the administrative registers data to achieve statistical purposes.

- The design: a plan shall be elaborated on building statistical registers by the concerned parties in compliance with the implementation stages according to a timeframe for each of the stages of the process.
- Expertise: technical expertise is required in the field of preparing and processing administrative data to achieve the statistical register purposes.
- The infrastructure: it provides the following tools and electronic systems:
  - Building an information network through which administrative registers are linked to facilitate the data flow process to obtain accurate and updated statistical registers.
  - Preparing suitable applications for data retrieval.
  - Providing an effective system for data security.
  - Electronically linking entities or units providing data to obtain the same in real-time.

- Building partnerships with entities providing administrative data.
- Preparing a clear methodology related to data linkage.
- Preparing a statistical register data processing methodology.
- Working on building capacity and training team members.
- Unifying the statistical classifications used between the different entities.
- Unifying the concepts and definitions used in the various administrative registers.
- Providing the main technology infrastructure that meets the requirements of statistical register building.

3. Statistical process based on administrative data

Principally, data is transformed and moved referring to three stages. This process was designed by SCAD to improve the flow from administrative data to statistical registers. The first stage is the design stage which determines objects, variables, concepts, and data sources. The second stage is the preparation stage. The third stage is the processing stage (edit and process data).

Figure 2: Stages to improve administrative data to statistical registers
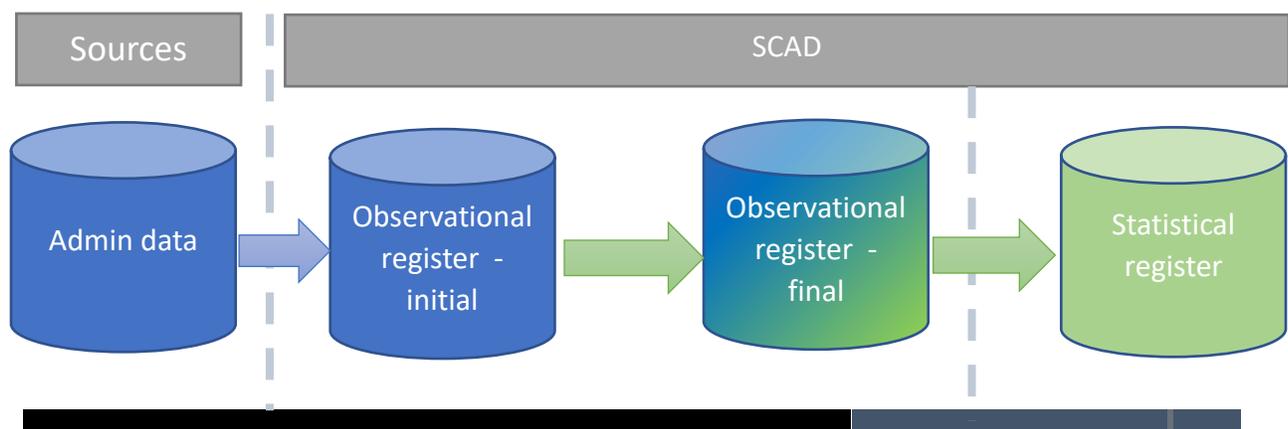
3.1. Designing the process

Determining the purpose of the statistical register

This is one of the most significant stages in the process of preparing the administrative registers being processed for statistical purposes. In fact, the clear determination of the purposes related to registers building facilitates planning for the next stages, while considering the implementation of the following upon ensuring compliance of the final results with the main objectives. Note that a statistical register has multiple purposes, e.g., being a frame for surveys and basis for estimating statistical values. For example, a business statistical register serves many different products within the economic statistical field.

The register building objectives emanate identifying the nature of the data available in administrative registers which can be used to produce new statistics that are consistent with the statistical entity and partners' needs.

The objectives of the statistical register shall be determined as follows:

- Determining the needed information from the required data.
- Determining the statistical variables covered by the data.
- Determining the ways and methods adopted in building the register and production of statistics.
- Determining how the statistical register can be used as a frame.

Determining the variables of the statistical register

In order to obtain data that meet the requirements for building statistics, the necessary variables for the statistical register building shall be determined as follows:

- Determining the types and characteristics of the main variables that will be collected from different sources.
- Determining and defining variables derived from the main variables of the register and listing them under the statistical register database. Make sure the definitions are unique. Also, check that time references periods are documented for each variable.
- Determining the statistical classifications that will be used in building the register.

Determining the concepts and data sources

It is necessary to identify the terms and concepts used in the statistical register, in order to measure the statistical process from the user's point of view. In addition, international terms and definitions shall be used to facilitate the comparison and consistency-related processes in the register. Whenever a new definition or term is used, whereby it is not internationally defined, an internal related classification shall be developed and adopted.

Within the framework of building statistical registers, it is necessary to obtain the data through the registers of several local, federal, private, governmental, and semi-governmental entities. Therefore, the main sources of data included in the statistical register to be built shall be determined.

3.2. Preparation Stage

This stage involves communicating with the entities providing data and concluding Service Level Agreements (SLA) with them in order to facilitate the data collection process. It also relates to agreeing on common formulas outlining the electronic linkage mechanism foundations and data transmission from the concerned entities, their periodicity, and the identification of the data to be obtained.

This stage also involves preparing and designing a unified data collection form covering all the available variables at the entities providing data, as well as providing the concerned entities with the relevant form to be adopted and filled in with their data.

Preparing different technical rules

This stage involves preparing the requirements of various tools and technical rules, as follows:

- Providing the data engineering specialist team with the following:

  o Names of the variables received from the data of each source.
  o List of the classifications of variables listed under the administrative register data from the source along with the corresponding symbols for the variables in the statistical register; and,
  o preparation of the automated validation rules related to the register variables.

- Preparing and designing an electronic database for the register, based on the unified data collection form. This database includes the main variables of the register, in addition to the variables that are derived from the register data.
- Preparing an electronic link system to collect data from different sources, by designing a system to link with each of the administrative data providers individually, in order to obtain data easily, quickly, and accurately, Therefore, data will be listed under the databases of each entity, which will facilitate the relevant review, editing, and follow-up processes. This shall constitute a preliminary stage preceding the data collection stage from different sources within a single database, whereby they shall be prepared to include all the required data and thus, to be transferred to the main register.

3.3. Processing stage

The data editing stage comprises procedures that include detection of erroneous, inconsistent and outliers, to improve the data quality, related to the purpose of the final statistical register. Editing is done through the whole process, from collection to processing at micro-level (data), to processing at macro-level (statistics).

Types of editing:

- Structural editing: this type is used to verify the variables' data coverage and completeness.
- Consistency test: this test is designed to verify the logical consistency between the variables within different databases, such as testing the consistency of the relationship between age and marital status, between age and education levels, etc.
- Identification of missing values: this step is conducted by reviewing the database and verifying that no cell is empty, which means ensuring that all cells are filled with data through referring to the data source or statistical processes.
- Identification of duplicates: this type of editing aims at verifying that the database does not contain any duplicates for the registered values of some variables.
- Detection of extreme values: this step is conducted through examining data distribution and relying on data series for previous periods or on different data sources, such as comparing the values of the variables between registers derived from different entities. This shall facilitate the detection of values that may be considered extreme values. This expertise is provided by subject matter.

Processing missing and extreme values

This step tackles processing missing or extreme values of some variables through imputation using alternative values for each of the missing or extreme values. The most important methods used in this field are the following:

- Imputation using specific values: these values are determined using the logic of linkage between variables. For instance, the gender of an individual is linked to his name.
- Stochastic imputation: this method is applied through adopting the Cold-Deck Imputation method which applies the imputation formula of the missing values in the register for the current reference period and replaces them with similar values under the same prevailing conditions prevailing of a previous reference period. The Hot-Deck Imputation method may also be used, where it depends on selecting values that are not listed in the concerned register or from other registers with the same conditions related to the missing values.
- The regression method: this method depends on the availability of auxiliary variables with complete data. Based on these variables, a regression model shall be built under which the variable with missing data is referred to as dependent variables, compared to the auxiliary variables referred to

as independent variables. The accuracy of the imputation results in this context mainly depends on the convergence and correlation extent between the independent variables with the dependent variable. After building the regression model, the dependent variable missing value may be predicted through imputing independent values in the regression model.

4. Monitoring the quality of register-based statistics

The quality of the outcome of register-based statistics is assessed in exactly the same way as the outcome of survey-based statistics. The five internationally recognized quality components are used to describe and explain the quality for the end users of the statistics. However, when it comes to input data, the quality aspects are different. The reasons for that are mainly two:

- administrative data are collected for other purposes than statistical, and
- there is no random sampling from a population when using administrative data.

These two factors affect how we process the data to statistics. Some elementary monitoring activities are.

- Check the completeness of the admin data.
- Check identification number and key variables.
- Check data sets received with needs and defined in SLA.
- Check variables received with needs and defined in SLA. (Must be readable format.)
- Check that values reported are within the expected range.
  - o Compared with external sources.
  - o Compared with previous month, quarter, year data.
- Contact data suppliers if something is unclear or there are big changes.
- Check missing value: reasons and extent.

5. References

- United Nations, Guidance on Building Statistical Business Registers, link: https://unstats.un.org/unsd/business-stat/SBR/Documents/UN_Guidelines_on_SBR.pdf. (Downloaded 20th June 2023)

- ESCWA, Statistical Business Registers Quality, link: http://www.aitrs.org/sites/default/files/ESCWA%20SBR%20quality%20ECE.pdf. (Downloaded 20th June 2023)

مركز الإحصاء
STATISTICS CENTRE

الرؤية: ببياناتنا نمضي نحو غدٍ أفضل
Vision: Driven by data for a better tomorrow

www.scad.gov.ae

adstatistics