Data Processing and Dissemination

CHAPTER 4

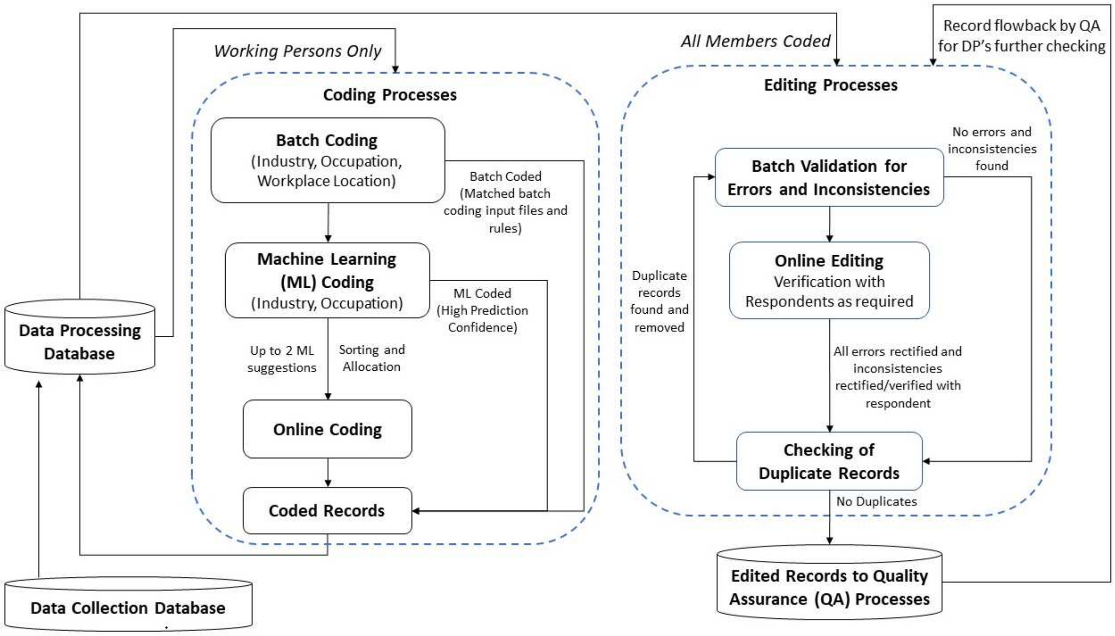
DATA PROCESSING AND DISSEMINATION

4.1 DATA PROCESSING

Census survey returns submitted from the various data collection modes flowed from the collection database to the data processing database for various data verification and processing processes, primarily statistical coding, data validation and editing. The Data Processing team started coding and editing the data after the records were submitted to the system.

The Census 2020 data processing system incorporated enhanced features to streamline operations and reduce manual effort, such as use of Machine Learning (ML) in coding and enhanced batch coding processes (Figure 4.1). After stringent quality checks and the completion of data processing, multiple aggregated cross-tabulations were also generated for broader analysis of data trends and identification of outliers for further checks.

Figure 4.1: Data Processing Workflow


Working Persons Only
All Members Coded
Record flowback by QA for DP's further checking
Coding Processes
Batch Coding (Industry, Occupation, Workplace Location)
Batch Coded (Matched batch coding input files and rules)
Machine Learning (ML) Coding (Industry, Occupation)
ML Coded (High Prediction Confidence)
Up to 2 ML suggestions
Sorting and Allocation
Online Coding
Coded Records
Data Processing Database
Data Collection Database
Editing Processes
No errors and inconsistencies found
Batch Validation for Errors and Inconsistencies
Duplicate records found and removed
Online Editing Verification with Respondents as required
All errors rectified and inconsistencies rectified/verified with respondent
Checking of Duplicate Records
No Duplicates
Edited Records to Quality Assurance (QA) Processes

4.1.1 Coding of Industry, Occupation and Workplace Location

Data collected in the Census 2020 were mostly captured in pre-coded categories except for selected descriptive fields including industry, occupation and workplace location and numerical fields such as income. This allowed majority of the items to be captured and coded at source and reduced the subsequent processing efforts.

Coding was necessary to allow descriptive data to be analysed, summarised and compared. It involved the assignment of codes based on descriptive text information collected on industry, occupation and workplace location according to specific sets of classification codes. Industry was coded using the SSIC 2020, occupation using the SSOC 2020 and workplace location using the 6-digit postal code.

All records underwent batch coding as a first stage. This was an automated process during which appropriate codes were assigned by the application program based on the descriptive information captured and in-built coding rules. Descriptive information captured from Census 2020 survey returns were matched against various input files, prepared using the standardised classifications and past survey data. Where a complete match was found during batch coding, the record was assigned with the specific code.

Enhancements to the batch coding processes included the use of more administrative sources as input files for the coding of industry and workplace location. When the respondent's survey returns were in agreement with administrative records, the records were automatically coded using the corresponding administrative data.

Dropdown lists were also implemented during data collection for the first time in Census 2020 to supplement free-text responses for the collection of industry, occupation and workplace location. Respondents were able to select their company name, occupation and workplace addresses from preloaded lists and these selections were auto-coded accordingly to reduce the proportion of survey returns based on free-text survey returns. This not only reduced respondent's burden but also improved the match rate against input files in batch coding processes by standardising the responses.

ML processes were also introduced and used for the coding of occupation and industry. This was developed in-house and built as a module to be linked seamlessly to the Census 2020 IT system. ML used various algorithms to analyse data to predict a likely statistical code. The machine was trained using historical data from MOM's CLFS. Free-text responses collected for industry and occupation were input into the model for prediction of SSOC/SSIC codes. For the coding of occupation, where the prediction confidence was high, the record was automatically coded by ML. Otherwise, up to 2

ML suggestions at a broader level of classification were displayed to facilitate online coding. For industry coding, only ML suggestions were provided to coders in online coding.

With the enhancements on batch coding processes, batch coding rates for industry and occupation were about 75 per cent and 30 per cent respectively, a significant improvement from about 18 per cent in Census 2010. Batch coding rate for workplace location was about 90 per cent. This significantly reduced the manual effort needed for coding.

Records that were not automatically coded proceeded to online coding, in which coding staff reviewed the entries and assigned suitable codes to the records. Sorting and batch allocation were performed for online coding, to group and assign records with similar characteristics to the same staff to improve the productivity.

During the online coding process, coding staff used a search facility to retrieve relevant information from the input files for determination of the appropriate codes. This search facility used the “key word” search method to find an appropriate match. The results of the search were displayed onscreen. ML suggested codes were also displayed to facilitate industry and occupation coding, where available.

After coding was completed for all working members in the house, the house records flowed to the next phase of processing for editing.

4.1.2 Data Validation and Editing

Data verification and editing were performed to ensure completeness and consistency in data collected. The main processes were batch validation, online editing and checking of duplicated records.

Batch validation was an automated process during which all records went through a series of stringent checks based on rules built in. The “error” and “consistency” rules were consolidated from past experiences with censuses and household surveys. “Error” rules checked primarily for missing key information which should not be left blank and validity of codes. They were also designed to detect records with entries for two or more data items which were logically impossible. “Consistency” rules checked for outlier records and scenarios which were unlikely to occur but were still valid and might exist. Examples of such rules include small age difference between parent and children, and attainment of university qualifications at relatively young age. Records which failed error or consistency rules during batch validation were flagged out and retrieved for online verification and editing.

Editors scrutinised the areas flagged, corrected/accepted inconsistencies highlighted for each record, and contacted respondents for clarifications where necessary. Iterative checks were performed and continued until the records were error-free and had inconsistencies resolved.

Towards the end of data processing, checks were conducted to retrieve records of individuals who were enumerated more than once during data collection. This was the result where individuals or households were enumerated at multiple addresses within the survey period. The duplicated records had to be studied and removed to prevent double counting. The retained records were then passed through the online verification and editing for another round of verification of errors and inconsistencies.

4.2 QUALITY ASSURANCE

Edited records were further reviewed at quality assurance stage. These records were put through a series of additional cross-variables consistency checks. Administrative data were also used to identify possible misreporting. Examples of such checks included inconsistency between occupation and employment status such as persons working as food deliverer, taxi driver or private-hire car driver who were not employed by a company on a fixed salary but misreported as an employee.

Records which failed the quality assurance checks were flowed back to the data processing team for further verification and editing. In addition, preliminary data compilation and analyses which included checking of data against historical trends were performed at regular intervals to identify possible data issues for further review and editing. The quality assurance checks served as an additional data verification process through an alternate perspective separate from data processing, to identify, reduce possible errors and enhance the quality of the data.

4.3 RESPONSE RATE AND MODE OF RESPONSE

Despite the Census being conducted in the midst of the evolving COVID-19 situation in 2020 and the adjustments made in response to the measures put in place, the overall response rate for the Census 2020 remained high at 94.6 per cent. Households who provided partial information, were non-contactable or overseas during the survey period, were categorised as non-responding units.

The proportion of households who submitted their survey returns over the Internet was the largest, at 64 per cent. This was followed by submission of survey returns through CATI at 25 per cent and face-to-face interviews at 11 per cent. More details on the changing modes of response over the years were discussed in Appendix P.

4.4 DATA DISSEMINATION

With the COVID-19 situation and imposition of the CB in April – June 2020, Census 2020 Call Centre operations were scaled back with a small number of staff working from home and field work suspended. Survey returns from some respondents were also delayed till end 2020. The release of the bulk of the Census results was also shifted from February for the previous Census in 2010 to June for Census 2020.

4.4.1 Statistical Releases

From the register-based Census, the merged administrative records provide the basic population count and characteristics such as age, sex, ethnic group, type of dwelling and geographic distribution in Singapore. Basic data on population count and profile from the register-based Census were first released in the Population Trends 2020 report published in September 2020.

Following the completion of data processing for the Census sample enumeration, DOS released a series of topical Statistical Releases (SRs) on detailed Census results during the period of June 2021, as follows:

Publication	Topic	Release Date
SR No. 1	Statistical Release 1 on Demographic Characteristics, Education, Language and Religion	16 Jun 2021
SR No. 2	Statistical Release 2 on Households, Geographic Distribution, Transport and Difficulty in Basic Activities	18 Jun 2021

The SRs present the broad trends and changes between 2010 and 2020. The reports also contain key indicators, comprehensive and detailed tables, charts and descriptions of the concepts and definitions (see Appendix Q – Glossary) which serve to meet the needs of the general public for a wide range of data on Singapore's population and households. They serve as comprehensive reference sources for planners, researchers and other data users requiring in-depth data for their analyses.

The Census 2020 reports are available for free download on DOS's website.

4.5 CONFIDENTIALITY, PRIVACY AND SECURITY

The Census 2020 is conducted under the *Statistics Act*. DOS has the responsibility and obligation to ensure that any personal information provided by respondents for the Census 2020 is kept strictly confidential in accordance with the *Statistics Act*.

Stringent procedures were implemented to ensure that confidentiality was maintained at all times:

- All employees, including casual temporary employees recruited for the Census 2020 project, were required to sign undertakings to safeguard official information and individual information obtained from the Census 2020.
- Access to personal and confidential information was evaluated and granted on a need to know basis.
- Monthly reviews on the access logs were carried out for each module in the Census 2020 system to ensure that there was no unauthorised access.

In the release of statistical information:

- Data are grouped (e.g. information is provided for aggregated age groups).
- Data for specific profiles or geographic areas with populations below a specified threshold are either grouped at broader categories or suppressed.

DOS pro-actively consults with users to balance the need for more information with its requirement to protect confidentiality.