Incorporating Administrative Data into Population Census 2020

Jeslyn Tan, Ministry of Manpower, Singapore
Jeremy Heng, Ministry of Manpower, Singapore

Abstract

In today's climate of evidence-based policy development, governments around the world demand large amounts of information at short notice. Thus, it is important for national statistical agencies to look beyond traditional survey sources to produce official statistics. For the Population Census 2020, the Manpower Research and Statistics Department (MRSD) of Singapore, the authoritative source for official labor statistics, will be incorporating administrative data to supplement survey data.

There are several stages where administrative data can be incorporated as part of the statistical production process: (1) pre-population of data before the start of survey; (2) data validation after survey fieldwork is completed and (3) utilizing data analytics outside the data collection process. This not only reduces operational costs, but also respondent burden. By further tapping onto technology, the longer-term goals of better operational efficiency and data quality are achieved.

The paper discusses the challenges facing the conduct of the Census, the initiatives taken to tackle the challenges, and how administrative data can be a vital component in the production of official statistics.

Keywords: Administrative Data, Census, Data Analytics

1. Introduction

The Population Census is conducted in Singapore once every ten years, at the end of every decade. Population Census 2020 will be the sixth census carried out in Singapore since Independence and the fifteenth in the series of Census taking in Singapore. The Population Census provides the most comprehensive source of demographic, economic and social information on households and individuals in Singapore. Statistics compiled from the Population Census is used by the government for policy review and development. The information also benefits individuals and businesses by providing insights to the labor market and aids them in decision-making for their careers and businesses.

Singapore's official labor statistics is produced and compiled by the Manpower Research and Statistics Department (MRSD) of the Ministry of Manpower. For Population Census 2020, MRSD advocates that the survey be conducted in a form of a registry database.

With the increasing availability of administrative sources, utilizing these sources can bring about a reduction in operational costs and response burden, and the potential of having microdata at disaggregated levels to enhance the production of detailed statistics. The paper provides an overview on how administrative data will be used in the conduct of Population Census 2020.

2. Pre-population of data before the survey

The Population Census is conducted under the Statistics Act (Chapter 317) which empowers MRSD to collect information from individuals and households. The Act also safeguards the confidentiality of information provided. The Census covers private households in Singapore, but excludes workers living in construction worksites, dormitories and workers' quarters at the workplace and persons commuting from abroad to work in Singapore.

The usefulness of administrative data begins at the pre-survey stage, where data from administrative sources is imported into the survey frame before the launch of the survey. Population Census 2020 will be conducted in the form of a registry database, where respondents only need to verify whether specific information, which is based on administrative records, is accurate. If it is inaccurate or not up-to-date, they can make changes to their information. Similarly, if they have participated in other MRSD surveys and their responses are relevant, the past survey information will be pre-populated into Population Census 2020.

Not all information required for Population Census 2020 is available from administrative sources, especially those related to respondents' sentiments and beliefs. If administrative records are inadequate to answer some questions in the survey questionnaire, respondents will be asked to provide their information. Therefore, both the use of survey and administrative data is important. For example, estimates of the total labor force are derived by compiling administrative records on foreign employment data with survey data on residents.

Figure 1: Population Census 2020 data collection process

The majority of households participate in the survey online or through phone interviews. Households that do not respond through these options are to be enumerated through face-to-face interviews. All the information provided are to be keyed into the secure Integrated Manpower Survey System (IMSS). Should there be missing, invalid or inconsistent entries, respondents are prompted through the system or by interviewers to correct the information provided. The work of interviewers is also subjected to consistency and verification checks to ensure good data quality.

3. Data validation after the survey

3.1 Automated occupation and industry coding

As part of Population Census 2020, information on occupation and industry are key data items that will aid policymakers have an accurate sensing of the labor market. They are also tedious data items to collect, requiring large amounts of time and resources to classify the textual information. In the past, respondents would just provide some details of their occupation and industry and interviewers have to manually classify each of them into one of the thousands of codes of the Singapore Standard Occupational Classification (SSOC) and Singapore Standard Industrial Classification (SSIC).

For Population Census 2020, MRSD seeks to automate the coding of SSOC and SSIC based on various information collected from individuals. This includes age, gender, income, educational qualifications, years of working experience, etc.

Past survey data and administrative records are incorporated for the development of the automatic classification system. For example, administrative data provided by Central Provident Fund (CPF) captures detailed information on employees' income and employment status at an individual level. It also includes information on the industry the individual is working in. This provides insights to a match between certain occupations and industries, and allows the system to make a more accurate prediction for SSOC and SSIC. For example, a driver with a Private Hire Car Driver's Vocational License (PDVL)

and fetching passengers is likely to be an own account worker, rather than an employer or employee. In such a situation, the SSIC and SSOC for the driver would be Passenger land transport (49219) and Private-hire car driver (83226) respectively. Hence, the system provides a consistent and accurate SSIC and SSOC for the driver, as administrative data shows the presence of PDVL.

The automatic classification system helps to alleviate burden on interviewers as the SSIC and SSOC can be coded automatically with up to 90% accuracy. It also ensures consistency amongst interviewers and respondents who may have different interpretations of the same code.

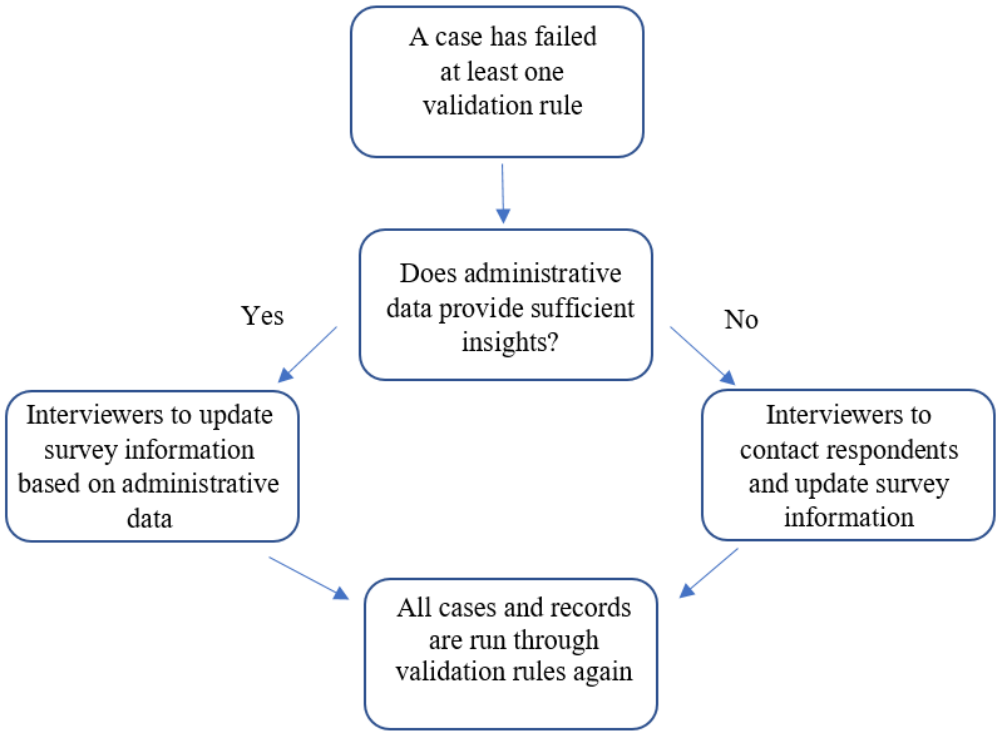
The automatic classification system is also implemented as a chatbot on Telegram, a cloud-based instant message mobile application. As such, field interviewers can access the system on-the-go, thereby increasing the efficiency of conducting phone and face-to-face interviews.

3.2 Validation rules

After data collection, the survey data passes through a set of validation rules in a system called Integrated Manpower Analytics System (IMAS). The validation rules are split into two types of checks, namely "Errors" and "Alerts", and are cumulated from past surveys. Rules for "Errors" check primarily for missing survey data or responses that are illogical. Apart from identifying problems related to routing, IMAS also prompts interviewers to check on information that are provided wrongly. For instance, if a working proprietor is classified as an employee, interviewers are notified that the working proprietor should be an employer or own account worker. Rules for "Alerts" check for possible discrepancies and for scenarios that are possible but improbable. For instance, age difference between spouses of more than 30 years, or an office clerk earning a monthly salary of more than $20,000.

The entire survey dataset for Population Census 2020 will be run through IMAS for validation. Should a case fail one or more validation rules, it will be flagged out to the corresponding interviewer who will cross-check the survey data against administrative data. For example, administrative data from the Ministry of Home Affairs provides information on individuals living in Singapore pertaining to their national identification number, date-of-birth, gender, and other types of personal information. This information can then be matched with other administrative sources to obtain even more information about the respondents.

Figure 2: Incorporating administrative data with validation rules



While there can be inconsistencies across administrative data due to differing definitions and coverage, Population Census 2020 seeks to adopt information from various administrative sources to supplement survey data. MRSD will prioritize the various data sources according to reliability, relevance and timeliness. If there are inconsistencies between survey and administrative data, the cases have to be addressed individually. Interviewers will contact respondents via phone calls or emails to verify certain information where required. They will then review and update the survey information accordingly.

After verification, checks are conducted to retrieve duplicate records of individuals who responded to the survey more than once. This may happen when an individual shifts house during the survey reference period, and both addresses are being selected for the same survey. The duplicate records are deleted to prevent double-counting.

4. Data Analytics

4.1 Predictive and sentiment analysis

MRSD conducts national surveys mainly via three modes: internet, phone interviews and face-to-face interviews. Phone interviews are conducted through Computer-Assisted Telephone Interviews (CATI). Apart from the advantage that CATI provides in enabling interviewers to view pre-populated data in IMSS when conducting interviews, they are also able to key in survey responses directly into the survey system. The system also performs routing or logic checks if survey responses are found to be erroneous or inconsistent. This makes the data collection and validation process more accurate and efficient. Having an online survey system also eliminates the need to have hardcopy survey forms.

As MRSD conducts over 30 surveys each year, there are vast amounts of past survey data readily available. MRSD has tapped on these resources to generate insights and improve the data collection process. The use of predictive analysis is one of them. After each phone or face-to-face interview, interviewers will record down the outcome – whether the interview was completed, respondent was uncontactable or a refusal case. The vast amounts of data available allows us to predict the optimal date and time to contact each household based on respondent demographics matched with administrative data. For example, if a household comprises of just two working adults, the optimal timings to contact the household for an interview would be weekday nights or weekends as they are likely not at home during normal working hours.

Sentiment analysis is also used to provide insights into call center operations and customer service standards. With the use of speech to text analytics, voice conversations can be converted to text data. Text mining of contextual keywords is subsequently applied. If the system detects any disinclination or hostility from respondents, the case will be escalated to supervisors for follow-up action. Similarly, if poor customer service is shown on the part of interviewers, they will be sent for re-training and re-evaluation. As a national statistical agency representing the government, it is important to not only compile quality statistics, but to deliver quality service to the public as well.

4.2 Route optimization

To conduct face-to-face interviews, MRSD hires a large number of field interviewers who perform house visits in order to reach the respondents. Households selected to participate in the survey are spread out across Singapore. Previously, cases were randomly assigned to interviewers which is non-optimal, and a lot of time is spent travelling from household to household. Hence, MRSD has come up with a route optimization program that can allocate cases efficiently to interviewers.

Geographical clustering is first adopted to ensure households are in close proximity to each other. Administrative records on electricity and water usage are also used to determine the likelihood that the household is vacant. The vacant houses are then removed from the clustering.

Within each geographical cluster, the most optimal route is mapped out by utilizing navigation services via Google API. The route optimization program also takes into account obstacles, congestion zones, peak periods, etc. The geographical clusters are then assigned to field interviewers, such that each interviewer only needs to visit the allocated

households in a prescribed order. By following the optimized route, field interviewers are able to minimize travelling time and conduct face-to-face interviews more efficiently.

5. Response rates

Declining response rates is one of the biggest challenges faced by national statistical agencies and MRSD is no exception. One common reason cited by respondents who refused to participate in the survey was that, as a government agency, MRSD should already have all information about them, hence eliminating the need to survey them. Others also feedback that the survey questionnaires are too lengthy and complex, thus increasing their survey fatigue.

By incorporating administrative data into survey operations, respondents only need to verify the information provided is accurate. The registry database helps to streamline the way MRSD collects data from survey respondents. Respondents can update their information seamlessly and no longer need to key in their information into a survey form from scratch. The registry database is designed to be more intuitive and user-friendly than a traditional survey. As such, this will reduce response burden on the respondents, and subsequently reduce operational costs and increase response rates.

6. Conclusion

Today, national statistical agencies have to be transparent in publishing official statistics and compile them in an accurate and timely manner. A Census is often tedious to conduct and requires plenty of time and resources to complete.

With greater availability of administrative data, it is becoming an important component in the statistical production process. Supplementing administrative data to survey data can help to improve data quality and operational efficiency. To ease the data collection process, it is important for administrative records to be timely and relevant as well. Data from various administrative sources have to consistent with each other. Standardization of reporting is crucial to prevent different interpretations of the results.

MRSD seeks to conduct Population Census 2020 in the form of a registry database, where respondents can see their current and past information through the survey system. All amendments can be easily managed by both respondents and interviewers alike to keep the information up-to-date. By improving the operations and methodology of Population Census 2020, it will benefit future large-scale surveys conducted by the government.

References

Department of Statistics Singapore. (2010). Census of Population 2010 Statistical Release 3 Geographic Distribution and Transport. Retrieved from https://www.singstat.gov.sg/-/media/files/publications/cop2010/census_2010_release3/cop2010sr3.pdf

Department of Statistics Singapore. (2018). Singapore Standard Industrial Classification SSIC 2015 (Version 2018). Retrieved from https://www.singstat.gov.sg/standards/standards-and-classifications/ssic

Department of Statistics Singapore. (2018). What is the Census of Population? Retrieved from https://www.singstat.gov.sg/our-services-and-tools/public-sector-surveys/census-of-population-faqs

International Labor Organization. (n.d.). International Standard Classification of Occupations. Retrieved from https://www.ilo.org/public/english/bureau/stat/isco/

Ministry of Manpower, Manpower Research and Statistics Department (2018). Labor Force in Singapore. Retrieved from https://stats.mom.gov.sg/iMAS_PdfLibrary/mrsd_2018LabourForce_preface.pdf

NUS Libraries. (n.d.). Singapore Statistics: Census. Retrieved from https://libguides.nus.edu.sg/c.php?g=145497&p=956004

Singapore Statutes Online. (2019). Statistics Act (Chapter 317). Retrieved from https://sso.agc.gov.sg/Act/SA1973