Evolving Statistics, Transforming Decisions

Jeremy Heng

Paper prepared for the 16th Conference of IAOS
OECD Headquarters, Paris, France, 19-21 September 2018

Session 6.A., Day 3, 21/09, 14:00: Central role of national statistical offices

Jeremy Heng
jeremy_heng@mom.gov.sg
Ministry of Manpower

Evolving Statistics, Transforming Decisions

DRAFT VERSION 30/08/2018

Prepared for the 16th Conference of the
International Association of Official Statisticians (IAOS)
OECD Headquarters, Paris, France, 19-21 September 2018

ABSTRACT

The demand for timely and accurate statistics has not been so acutely felt until recently. As governments around the world face increasing pressure to tackle national issues quickly and effectively, they are driven to formulate policies based on objective information, giving rise to the advent of evidence-based policy development. Where official statistics are unable to meet their needs, policymakers are turning to alternative sources of information such as administrative data and big data to strengthen their case for policy decisions. This has led to questions about the role of official statistics in modern government administration.

To ensure the relevance of official statistics, the Singapore Ministry of Manpower seeks to reshape the statistical landscape on two fronts: (i) improve the data collection process through government surveys and (ii) augment survey data with alternative data sources to form a holistic picture for decision-making.

Firstly, government surveys will be conducted through a respondent-centric online system. By leveraging on technology, surveys can be completed and validated on-the-go by respondents, leading to greater convenience, timely participation and better data quality. Survey analytics, including text and speech analytics, are incorporated into the system to gain insights on respondent behaviour and patterns. This allows the Ministry to fine-tune the data collection process dynamically and enhance operational efficiency. Secondly, survey data will be supplemented by data from alternative sources to form an accurate representation of the target population. Where reliable and up-to-date administrative data is available, it will not be collected via surveys. Instead surveys, used in conjunction with focus groups, can be adapted to probe further for qualitative responses, providing greater scope for analysis by policymakers. To operationalise the process, a centralised database is set up that amalgamates data from various sources, and further data processing functions can be performed.

The two-pronged approach leads to the dissemination of more timely and accurate information, while reducing respondent burden and operational costs. Empowered with quality statistics, policymakers can improve the lives of citizens through better policies. The paper discusses the challenges faced by national statistical agencies in data collection and management, processes that mitigate the risks of data gaps and lags, and initiatives that preserve the long-term viability of official statistics.

Keywords: technology, data quality, operational efficiency

INTRODUCTION

Official statistics is facing challenges on many fronts. As modern society evolves, population needs and wants are becoming more sophisticated, just as national issues become globalised and intertwined. As a result, governments around the world face increasing pressure to tackle national problems quickly and effectively, and at the same time cater to the needs and wants of the population. With the size of the task on hand, government officials are less reliant on walking the ground and gathering anecdotal evidence. Instead, they are driven to formulate policies based on objective information, giving rise to the advent of evidence-based policy development.

Government policymakers and researchers demand timely and accurate information in order to have a good understanding of the country's situation. Official statistics has traditionally been used as the main source of information for policy research and analysis. However, in this information age, vast amounts of information such as administrative data and big data are readily available in the public domain. This provides a huge alternative source of information, which has been increasingly popular in recent years as data users recognise its value and potential. Where official statistics are unable to meet their needs, policymakers are turning to alternative sources to strengthen their case for policy decisions. This has led to questions about the role of official statistics, and whether it is still relevant in modern government administration.

The paper focuses on official statistics in Singapore's context, and in particular, labour statistics produced by the Ministry of Manpower (MOM). It presents the challenges facing official statistics, solutions that are designed to tackle these challenges, and initiatives that preserve the long-term viability of official statistics.

BACKGROUND

Singapore's official labour statistics is produced and compiled by the Manpower Research and Statistics Department (MRSD) of the Ministry of Manpower. The department conducts regular national surveys to collect a wide range of labour-related data from households, individuals and businesses. The survey data is cleansed and processed into usable information which is then analysed to provide labour market insights. The statistics and accompanying publications are eventually disseminated to the public and policymakers for their use. The labour market indicators that track the state of the economy include, but is not limited to, employment, unemployment, income, job vacancies, labour turnover, retrenchment, employment conditions and hours worked.

Being a national statistical agency, statistical activities conducted by MRSD are governed by law, namely the Statistics Act. The Act safeguards the confidentiality of information collected from individuals and companies and also makes it mandatory for them to provide MRSD with the necessary information.

CHALLENGES

Rising Expectations

Policymakers and researchers demand a wide range of information for their policy and research needs. Besides quantitative information, they are also looking for qualitative information to paint a fuller picture of the socio-economic situation. For instance, unemployment rates alone may not reflect the exact situation on the ground. Policymakers seek more in-depth information such as the reasons jobseekers are unable to secure work, reasons for leaving their previous job, and their sources of financial support. It is no longer sufficient to produce broad-based statistics as fine, granular data is required for modern-day government administration. Detailed demographics and comprehensive occupation and industry data up to 5-digit classification are needed to have more meaningful data breakdowns for analysis. In addition, timely dissemination of statistics is necessary to ensure policies do not lag current circumstances and become outdated, as policymakers demand for accurate information at short notice.

Data Collection

With greater need for timely, comprehensive and accurate information, more surveys are conducted on a more frequent basis. Sample sizes have expanded to ensure robustness of data even after it is broken down to granular levels. The length of questionnaires has also increased to accommodate the number of data items and indicators needed for research and analysis. An immediate impact of these factors is the rise in operational costs to conduct government surveys. Another implicit cost is the burden placed on households and businesses to participate in surveys. As Singapore has a relatively small population of around 5.6 million, a significant proportion of the population has been selected for government surveys. This results in increasing man-hours spent providing information via surveys, putting a strain on productivity and the economy over the longer term. It has also incurred a social cost, with rising public dissatisfaction over the prevalence of surveys and concerns over data privacy and confidentiality. In recent years, MRSD has faced resistance and refusals from respondents when conducting national surveys. It has resulted in falling response rates which can potentially affect data accuracy and weaken the inferences drawn from the survey results.

Data Validation

Before the data collected is accepted as fit for use, it is validated through a set of validation rules to eliminate any errors and inconsistencies. Sometimes, it involves contacting survey respondents again to verify specific survey responses that are suspicious or inconsistent. As the questions asked in surveys become more complex, there is a higher tendency for respondents to misinterpret and provide incorrect information. As a result, additional effort and time is spent on data validation, putting a strain on resources and increasing the turnaround time for the compilation of statistics. Further burden is also placed on respondents where they are contacted multiple times over each survey.

Data Processing

The impact of conducting numerous surveys is the large amounts of raw data that has to be processed and compiled. The data has to be managed and stored in a systematic manner that allows easy retrieval in future. Being an official source of labour market information, MRSD handles numerous data requests on a daily basis from both internal and external stakeholders. Often, these requests involve data to be cut across different dimensions, and yet should be presented in a simple and intuitive format. With a myriad of data sources available today, having access to complex data in a simplified format is a primary concern for many data users.

Rising Competition

While MRSD produces a significant amount of labour statistics, there are also many administrative data sources that the public and private sector possess. Coupled with the proliferation of social media and big data, this has given rise to alternative sources of information. Data users are becoming increasingly savvy to mine these data to serve their needs. As technology progresses, the field of data science has become more prevalent to facilitate users in their analysis. Although these alternative data sources have their limitations, some users are willing to trade off accuracy for speed in order to have a quick sensing of the socio-economic situation. Data users are relying less on official statistics as statistics compiled from traditional survey sources are generally deemed to be "too slow" and not responsive enough to fast-changing economic conditions.

SOLUTIONS

The challenges faced by MRSD are synonymous with that faced by many national statistical agencies around the world. MRSD has implemented several measures and initiatives to tackle these challenges, with some of them in the pipeline in upcoming years.

Administrative Data Sources

Administrative data can be a useful supplement to survey data if utilized correctly. As a national statistical agency, MRSD is able to make use of the powers of law, namely the Statistics Act, to requisite data from individuals and businesses for statistical purposes. In recent years, MRSD has been acquiring administrative databases and tapping on them in order to build up a consolidated database. As administrative data is collected mainly for administrative purposes, it may not always be appropriate to use it to draw conclusions about the target population. Nevertheless, there are administrative sources that are useful in validating survey data. For example, data pertaining to the Central Provident Fund (CPF), which is part of Singapore's social security system, is able to offer insights on the work income of residents. As collecting income data from surveys is often problematic due to possible under-reporting, having an alternate source of information is valuable as it can detect possible errors and discrepancies in survey data. Further checks with affected survey respondents can be conducted to correct these discrepancies. Databases of certain occupations are another

example of administrative sources that supplement survey data. These databases are compiled by their respective regulatory bodies which provide details of workers such as real estate agents, insurance agents, taxi drivers, maids, etc. This provides a profiling of a segment of the workforce and a basis of verifying individual occupational information.

The downside of using administrative data is that the database can be outdated and does not accurately reflect current economic conditions. The time lag, which can be months or even years, limits the potential of using administrative data to replace survey data totally. Instead, it can be used as a tool to enhance data quality. Nevertheless, there are types of data which are minimally impacted by time lags. Information on individual demographics such as race, gender, date-of-birth and education level, do not change often, if at all. Hence, most questions related to demographics need not be specially collected through government surveys.

Focus Groups

Besides the use of administrative data, focus groups are sometimes conducted to obtain qualitative information that are difficult to collect via surveys. MRSD conducts focus groups for two purposes. Firstly, it seeks to gain insights on labour market issues to aid ground sensing. For example, with recent trends in graduate unemployment and underemployment, focus groups are conducted to find out more about their experiences and motivations towards "graduate" work. Secondly, focus groups help to obtain feedback on respondents' attitudes and perceptions towards government surveys. With two-way communication, MRSD is able to better tweak their outreach efforts to maximize survey satisfaction and participation. One such feedback is that many respondents prefer to complete surveys on-the-go during their commute to and from work. Some also demand greater clarity on the objectives and results of the survey, thus emphasizing the need for effective public communications.

Survey System

MRSD seeks to develop a "respondent-centric system of the future" to facilitate participation of surveys by the public. With the proliferation of smartphones and tablets around the world, people are spending large amounts of time on their mobile devices wherever they go. With increasingly hectic lifestyles, it is crucial that respondents are able to complete surveys on-the-go at their own convenience. A decade ago, MRSD was among the first national statistical agencies in the world to develop an integrated online survey system where respondents and interviewers could submit, access and manage survey information in a single platform. This lowered operational costs and provided convenience to respondents to participate in surveys online. In recent times, respondent behaviourial patterns have changed such that it is no longer adequate just to have an online survey platform. It has to be mobile responsive as well to cater to the mobile crowd. MRSD is currently developing a mobile responsive version of its survey system to further boost online participation. Other than mobile capabilities, MRSD is tapping on various data analytics tools to enhance data quality and operational efficiency. The next section explains the initiatives to be incorporated into the new survey system that will help drive official statistics into the future.

NEW INITIATIVES

The new features and initiatives not only improve data collection but also downstream processes such as data processing and analysis. As surveys are mainly conducted via three modes – internet, phone and face-to-face, each mode is made simpler and more efficient, translating to greater participation and higher response rates.

Gamification

While having a mobile responsive system is a first step towards increasing online participation, gamification is incorporated to ensure continuous participation and loyalty. Game mechanics such as points and levelling up are added in the online survey system and applied to both respondents and interviewers.

For respondents, participation in each survey earns them survey points which enables them to reach certain milestones. Respondents can earn shopping vouchers for every milestone achieved, with increasing voucher value for higher milestones. In addition, prompt participation within a timeframe will earn respondents bonus points, translating to better rewards. However, if respondents do not participate online, interviewers are deployed to reach out to them instead. For interviewers, their remuneration is pegged to the number of cases they complete. Cases are mainly completed by interviewers via phone or face-to-face interviews, who then input the survey data into the online system. As interviewers complete more cases and reach higher milestones, their remuneration per case will increase, leading to higher exponential earnings.

Data Pre-Population

With the prevalence of administrative data, certain data items can be pre-populated in the survey system. When participating in government surveys, respondents just have to enter their unique identification number and all relevant administrative information will be populated. They can then verify the information and update any changes through the survey system where necessary. This reduces the number of survey questions that need to be answered from scratch, thus alleviating respondent burden.

Auto-benchmarking

When individuals and companies participate in surveys, they often would like to know where they stand relative to others. For instance, individuals may wish to know where their salary lies relative to the median in similar occupations. Companies may want to know the amount of wage increments they offer relative to industry norms. Therefore, the new survey system will include an auto-benchmarking tool that allows individuals and companies to benchmark their own information against others. After the relevant statistics are compiled, respondents will receive an auto-benchmark link via email. This contains not only the survey results, but also a tailored benchmarking report based on the information they submit through the survey. , Consequently, respondents are reciprocated for their participation in the survey, which will help increase response rates in the longer term.

General public users can also access the auto-benchmarking tool made available online. By keying in their own information, they can also benchmark themselves against the various industry and national norms. This helps educate the public on important labour market information and improves the statistical literacy of the population.

Speech Analytics

Part of MRSD's survey operations involve a contact center environment where respondents can call in to participate in surveys via phone interviews. Additionally, interviewers in the contact center make outbound calls to reach out to selected respondents who have not completed the survey. With a few hundred calls handled daily on average, there is a large amount of information, including survey responses, that reside within these calls. As manually listening to call exchanges is tedious and time-consuming, integrating speech analytics into the survey system is more efficient and can bring out useful insights from the conversations. For example, by transcribing speech into text, MRSD can assess the quality of phone interviews conducted by interviewers. Weaker interviewers can then be put through further training or mentoring. By filtering specific keywords and phrases, speech analytics can also help in data validation by drawing out hidden insights from phone conversations. In situations where interviewers misinterpret or overlook certain responses in their conversations with respondents, speech analytics adds an additional layer of checks to improve the quality of data collected.

Text Analytics

Text Analytics is another useful feature that draws meaningful insights from unstructured text data. Text data can come from survey feedback or responses. Like speech analytics, it can pick out constructive feedback or complaints by respondents. By applying it on text messages, email and other feedback channels, MRSD is able to resolve interviewer lapses promptly and raise service standards.

When survey data is collected, it is sometimes distorted or erroneous. While validation checks can help rectify these errors, text analytics can go a step further and uncover exceptional response patterns that may suggest a misinterpretation of survey concepts or a poorly designed survey questionnaire. This enables MRSD to tackle any systemic issues in interviewer training or questionnaire design.

Auto-coding

When collecting occupation and industry data via surveys, interviewers have to assign a standard classification code to it, which can be challenging as it relies heavily on the individual's discretion to produce an accurate code. The survey system will include an auto-coder to automate the conversion of raw occupation and industry data into standard occupation and industry codes. The auto-coder analyses the common keywords used in occupation and industry descriptions and recommends the most appropriate code to use.

This ensures consistency among interviewers and respondents who may have their own understanding of occupation and industry definitions, and eliminates any subjectivity based on individual interpretations.

Route Optimization

Field interviewers often to have to conduct house visits to households that are uncontactable by telephone. It is more effective to visit households that are within close proximity to each other in any given day. When assigning to interviewers, cases would be grouped according to their geographical location to maximize the use of time and effort. Therefore, each interviewer would focus on a specific geographical area to conduct house visits. Currently, grouping and assigning cases is largely a manual process and can be tedious and time-consuming. Hence, MRSD is developing an automated case assignment feature in the survey system that can analyze household addresses and assign cases to interviewers according to location. This eliminates the manual process and any subjectivity due to individual interpretations, leading to greater efficiency.

In addition, by making use of Global Positioning System (GPS) and route optimization algorithms, the survey system will compute the optimal travel route based on the interviewer's location in real time. The computation is done relative to a cluster of homes that the interviewer chooses to visit within the day. This enables interviewers to cover the greatest number of cases within the shortest possible time. The system can also alert interviewers of any nearby cases at any given location. For added flexibility, interviewers are able to set the detection radius. Overall, the benefit of having an automated and smart assignment tool is a more effective fieldwork process.

FUTURE WORK

Technology will continue to play an increasingly important role in official statistics. While still not widespread currently, we foresee data science, and in particular machine learning, to be an integral part of the statistical production process. To kick-start the process, MRSD will be collaborating with various data science firms to explore the feasibility of using machine learning algorithms to predict individual and aggregated data. For example, these algorithms can be run on past datasets to impute missing data points. Similarly, it can also improve operational efficiency by determining the probability of contacting a respondent successfully on any given day and time. It can also recommend the best time to contact a respondent based on its profile. With greater attention placed on machine learning, it can potentially change the way official statistics is produced in future. As statistics evolve over time, it will transform policy decision-making and implementation by the government, all of which culminates in better lives for its people.