Applicability of Periodic Censuses Data Quality Assessment Methods in China

Donghua Wan*
National Bureau of Statistics, Beijing, China – wandh@gj.stats.cn

Ran Tao
National Bureau of Statistics, Beijing, China – tomrand@sohu.com

Abstract

From data accuracy check to data error measurement, the data quality assessment research of periodic census continuous improving as people's awareness of the census data production process. A variety of specific assessment methods had developed, include four technical ideas, such as demographic analysis, data consistency check, administrative records check and post enumeration survey. According to the relationship between assessment standard and census system, these assessment methods can be summarized as three assessment ways, consist of internal assessment, external assessment and post survey evaluation. By analyzing the characteristics of each assessment methods and their applicability, this study will help further deepen the theoretical research of periodic census data quality assessment (CDQA) in China.

Key words: periodic census; data quality; assessment method.

1. Introduction

It has been 20 years since 1994 in which China established the periodic census system. The adjustment of census project in 2003 specified that there should be four census every ten years, including Economic Census, Agriculture Census and Demographic Census, in order to strenghthen the position of periodic census in the government statistical survey system. The quality of census data not only affect the credibility of our national government statistics department, but also concerns the scientificity and rationality of many government management decisions. The third Economic Census in China began in late 2013, was the beginning of a new round cycle of periodic census system. The plan of economic census specified that the census should base on the post enumeration survey and correlated historical data and administrative records, in order to evaluate accuracy, consistency and reliability of census data.

This paper reviews the existing research about the data error measurement of periodic census and proposes a scheme to classify the method for data quality assessment, in order to discuss their characteristics and applicability.

2. Reviews Of CDQA

From 1940s, with the probability sampling establish the position and the rise of several study on non-sampling error, people realized that census data was not perfect. Researchers in U.S. Census Bureau tried to check the accuracy of census data and put forward lots of CDQA methods firstly, which were studied and improved by other countries later.

2.1 Demographic analysis

From Whipple's index to Myers' index, Inflow-Outflow to Cohort Survival, the methods of census data error measurement had been developed and improved, according to distribution of gender, race, age, etc. By combining demographic principle and statistical methods, a series of evaluative methods of census data quality was developed, named demographic analysis. The research from U.S. Census Bureau (2004) and National Research Council (2009) indicate that demographic analysis is still adopted as one of the CDQA methods for Population Census in U.S.A. and widely employed around the world.

2.2 Data consistency check

Zarkovich (1963) pointed out that by comparing census data with census respondents' characteristic or their relationships, we can using consistency check to assess the census data quality. The first method is to compare and test the accuracy of historical census data from longitudinal perspective. The second method is based on internal consistency, for example, enterprise's annual spending limit is its sales revenue plus the balance and liabilities in economic census, and the relationship between cultivated area of crops and sown area meets the setting range of multiple cropping index in agricultural census. This method gradually evolved into logistic regulation test which used in census register check nowadays. The third method is to compare the consistency between census data and external independent data. Its premise is that external independent data and census data describes the same population and has relatively similar concepts of comparison and defining range.

Judging from the time-space relationship, the first and second method focus only on the single census data, also known as "internal consistency check". But the third method need external data, named as "external validation" by Biemer & Lyberg (2003). Moreover, only the external data with high accuracy can be used in CDQA.

2.3 Administrative record check

External validation assesses the census population accuracy mainly from macro aspect, with no consideration of the micro matching ability of external data. Countries with perfect administrative records can evaluate the accuracy of census data both from micro and macro aspects by matching check the individual census data and related administrative records. For that the administrative records are collected by other departments of government due to their administrative function and independent from government statistic department, when using it to evaluate census data quality, the evaluating capacity of administrative record check is usually higher than external data verification.

Biemer & Lyberg (2003) proposed Forward Record Check, relative to Eckler & Pritzker (1951) put forward Reverse Record Check, all these methods were named administrative record check, good external administrative data is required for comparing when using them.

2.4 Post enumeration survey

Post enumeration survey which measures census coverage error and content error by extracting one sample from census population and comparing it with the original census records by reinterview. The reinterview technique originated from the study of measurement error in statistical survey in U.S.A. and India. Forsman & Schreiner (1991) summarized two main purposes of reinterview: one is to evaluate the implementation quality of survey; the other is to evaluate the structure of survey's error. The post enumeration survey can measure the original census data error based on the result of reinterview.

With the study of measurement error such as response and interviewer error progressed in 1960s and 1970s, post survey based on sampling design inference was widely used in many countries and gradually developed into design-based post enumeration survey. Marks (1978) brought "capture-recapture model" into post enumeration survey, which was originally used for wild animal total amount estimation, naming it "dual systems estimation". Wolter (1983) discuss the dual systems model in measurement of census coverage error, laid a foundation for model-based methods. Tao (2012, 2014) discuss the design-based and model-based inference in post enumeration survey to census coverage error measurement. U.S.A government employed dual systems model to evaluate the coverage error in agricultural census since 1987. The purpose of post enumeration survey adopted in 2010 population census has changed from correcting the results to process evaluation and quality control, hence could continuous improve census data quality.

3. Approaches and Applicability of Periodic Censuses Data Quality Assessment

3.1 Analysis of Approaches

According to the System Theory, every evolutionary path of realistic system can be considered from the external conditions and internal factors. If we regard census data generating process as a system, basing on the different relationships between evaluation standard and census system, the CDQA methods mentioned above can be categorized into three approaches.

The first one is to obtain assessment standard through internal channels, where the "internal" is relative to the census system to be evaluated. Such standard is included in the original census system, hence can be obtained for sure. but for the reason that the assessment standard is not independent from original census system, even the methods and rules of consistency analysis is correct, this method could not be able to check out the systematic error.

The second approach is to get assessment standard through external channels. It guarantee the independence of assessment standard, and may still confront two problems: the first is that this assessment requires external data sharing the identical survey population, comparable definition and similar concept range with original census, but such data could be un-acquirable; the second is that even the external data with independent generating process is obtained, it's accuracy and effectiveness must be examined in order to ensure a reliable assessment result.

The third approach is post survey, which is applied for CDQA under the assumptions of the relationship between two different survey's data. The assessment standard is partially related to the original census system, hence can be gained for sure. However, the judgment standard obtained from reinterview in this approach is relatively independent, which can avoid systematic error to some extent. The problem for post survey method is to guarantee the effectiveness of assessment method by using necessary technique.

3.2 Classification and Applicability

According to the three approaches summarized above, the CDQA methods could be categorized in table 1, in which their characteristics are also presented.

Table 1 Classification and Characteristics of CDQA Methods

Assessment method	Evaluate approach	Principle	Assessment objective	Assessment basis	Applicable Census Project
Internal Assessment	Demographic Analysis	Natural changes	Population Error	Consistency of Distribution	Population Census
	Internal Consistency Check	Longitudinal Comparison & Logistic Regulation	Individual & Population Error	Longitudinal Consistency & Logic rule	General
External Assessment	External Data Validation	Population Matching	Population Error	External Data	General
	Administrative Record Check	Individual & Population Matching	Individual & Population Error	Administrative Records	General
Post Survey	Design-based Inference	Error Inference	Individual & Population Error	Reinterview Technique	General
	Model-based Inference	Model Estimation	Population Error	Dual Systems Model	General

As shown in table 1, internal assessment approach don't need to take the source and accuracy of external information into consideration and the result is mainly reflected by the internal consistency. Demographic analysis, which estimate the population distribution error by population consistency check, cannot check individual error, was not suitable for census process quality control, the result of demographic analysis is only served as conclusion to post evaluate the accuracy of population census.

If census registration error tendency of individual and population were estimated by logistic regulation, the results of internal consistency check can be used as the basis of process quality control in census. From the longitudinal comparison, internal consistency check can examines the consistency of census historical data from post assessment.

External Assessment approach requires accurate and effective external data. By merely evaluating census data's consistency from statistical population, external data check cannot get the accurate error measurement result. It can only applied as the post evaluate to census from macro aspect. For the precisely measurement of census individual and population error which can be used as the basis of process quality control of census, administrative record check with perfect and reliable administrative records is widely used in various of census projects. Compared with census registration, we can get timely statistic data with lower cost through administrative record method. As big data technology matures, administrative records with high quality are not only served as an important supplementary methods to get census data, but also used in evaluating census data quality.

Post survey is a reinterview based on sampling-survey. Compared with census, sampling survey have lower cost and great timeliness, especially the good applicability when precisely measuring census data error without prefect administrative records. The reinterview data was obtained by design-based inference, could evaluate the measurement errors of survey individual and population. The assessment results can be used in both census process quality control and measurement of aggregate census data error. Model-based inference may effectively measure various type of census population error by making effective connection between census records and post survey records through dual system model. Though couldn't served as the basis of process quality control, the measurement can evaluate aggregate census data error.

4. Conclusions

As the whole observation of all units, the accuracy of census can be represented as completeness for enumerating of every unit and registering of census indexes, in which the coverage error of aggregated census data becomes the research emphasis in theory and practice of census data accuracy measurement. For the consideration of data generating process, the quality assessment should not only based on the accuracy of aggregated results, but also take the census implementation process into account.

Owing to their own characteristics, both internal assessment and external assessment did not take census data generating process into consideration in traditional. With the combination of applicability of internal accuracy check and administrative records check, these two methods can be employed as process assessment in some steps of census data generating process, such as the quality evaluation of survey frame, census objectives' register, etc. The evaluation results can be used as standard of process control, which aiming to correct the quality problems found in census process on time, hence contribute to the success of follow-up work.

Based on reinterview technique, post survey evaluation can be employed in different steps and levels in census data aggregation process, hence has become an indispensable periodic CDQA all over the world. Related literatures both indicated that the post enumeration survey usually carried out after census register work, mainly used to measure the accuracy of census registration data, belong to post assessment. Therefore, the purpose of post enumeration survey around the world is not to correct the census data, but to evaluate the quality of data and related work, in order to provide a reliable standard for continuously improving census work. If the post enumeration survey is employed in one step of the census data generating process, the results can then be served as the standard of quality control for specified step theoretically. Such evaluation would help to correct the data quality problem in this step on time, thus achieve the purpose of necessary process control.

References

Eckler, A. R,, &Pritzker, L. (1951). Measuring the Accuracy of Emumerative Surveys. Bulletion of the International Statistical Institute, 33(4).

National Research Council. (2009). Coverage Measurement in the 2010 Census. Panel on Correlation

Bias and Coverage Measurement in the 2010 Decennial Census, Robert M. Bell and Michael L. Cohen (Eds.) . Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Washington, DC: The National Academies Press.
Zarkovich, S. S. (1963). Quality of Statistical Data. Food and Agricultural Organization of the United Nations, Rome.
Biemer, P. P., &Lyberg, L. E. (2003). Introduction to Survey Quality. Hoboken, New Jersey, John Wiley & Sons.
Forsman, G., &Schreiner, I. (1991). The Design and Analysis of Reinterview: an Overview// Biemer, P. et al. (Eds.). Measurement Errors in Surveys. New York, John Wiley & Sons.
Marks, E. S. (1978). The Role of Dual System Estimation in Census Evaluation//K. Krotki (Ed.), Developments in Dual System Estimation of Population Size and Growth, Edmonton: University of Alberta Press.
Wolter, K. M. (1983). Coverage Error Models for Census and Survey Data. Bulletion of the International Statistical Institute.
Tao, R. (2012). Census Coverage Error and Model Study from the Perspective of Non-sampling Error. Statistical Research, 29(12).
Tao, R. (2014). Expanded Dual System Estimation Model and Its Matching Properties. Journal of Applied of Statistics and Management, 32(2).

Acknowledgments
This research was supported by the National Natural Science Foundation of China (No.71301033).