# Small Area Estimation for the 2011 Population Census of Hong Kong

John HK LAM
Census and Statistics Department, Hong Kong, China
jhklam@censtatd.gov.hik

**Abstract**

In the 2011 Population Census of Hong Kong, a sample of households were required to provide detailed data on social and economic characteristics in addition to basic demographic information such as age and sex using a long-form questionnaire and the rest of households were only required to provide basic demographic information using a short-form questionnaire. The population count and the key profile of the population were compiled based on the basic demographic data collected from all households. For the more detailed analysis on social and economic characteristics of the population, estimation was performed by means of ratio estimation method using the benchmark data on key characteristics. In order to provide small area estimates on one hand and at the same time the estimates at the aggregate level are consistent with independent administrative data sources, a modified Generalised Regression Estimation method (with calibration estimation process) was used.

Keywords: generalised regression estimation

## 1. Introduction

In Hong Kong, it has been an established practice since 1961 to conduct a population census every ten years and a population by-census in the middle of the intercensal period. In the 2011 Population Census of Hong Kong, only about one-tenth of households were required to provide detailed data on social and economic characteristics using a long-form questionnaire and the rest of households were required to provide basic demographic information such as age and sex using a short-form questionnaire.

The population count and the key profile of the population were compiled based on the data on key characteristics collected from all households through both the short-form and long-form questionnaires. For the other characteristics of the population, estimation was performed by means of ratio estimation method using the benchmark data on key characteristics from both the short form and long form questionnaires. As such, there were three different sets of grossing-up factors, including grossing factors for quarters, grossing-up factors for households and grossing-up factors for persons.

## 2. Problems of Estimation

Using the above method in the small area estimation has the following limitations:

(i) The estimates compiled from both the short-form and long-form data may not agree with that based on long-form data only;

(ii) Different grossing-up factors are obtained when the estimations are undertaken based on different benchmark data (quarters, households or persons);

(iii) As the benchmark data obtained from the administrative systems are related to different dimensions of the population e.g. quarters, household or personal levels, it is difficult to produce the estimates which are consistent with all the benchmark data.

In order to produce small area estimates which, when aggregated, could give results consistent with independent administrative data sources at overall level, a modified Generalised Regression Estimation method (with calibration estimation process incorporated) was adopted to produce one set of grossing-up factors.

## 3. Modified Generalised Regression Estimation

In the 2011 Population Census of Hong Kong, the population count and key population profile were obtained based on the short form variables collected from all households in Hong Kong.

The modified Generalised Regression Estimation method aims to find out a set of calibrated weights (grossing-up factors) which is as close to that of the design weights as possible and, at the same time, the grossed-up figures (estimates) for selected auxiliary variables based on the calibrated weights can reconcile with the known independent estimate of the variables concerned.

The auxiliary variables e.g. number of quarters, age, sex by small geographical areas refer to some key short form variables for which consistency between population count and sample estimate is highly desirable. These variables could be seen as constraints on the calibrated weights.

In the Generalised Regression Estimation method, it is defined that

$Q$ = Control total for occupied quarters

$P_i$ = Control total for auxiliary variable *i*

$f_j$ = Initial grossing-up factor for quarters *j*

$\hat{Q}$ = Estimated total based on $f_j$ 's (= *Q*)

$\hat{P}_i$ = Estimated total for auxiliary variable *i* based on $f_j$ 's

$x_{j(i)}$ = Count of auxiliary variable *i* for quarters *j*

$q_j = \sqrt{\sum_i x_{j(i)} \times x_{j(i)}}$

$$\vec{x}_j = \begin{pmatrix} 1 \\ x_{j(1)} \\ x_{j(2)} \\ \vdots \\ x_{j(k)} \end{pmatrix} \qquad \vec{X} = \begin{pmatrix} Q \\ P_1 \\ P_2 \\ \vdots \\ P_k \end{pmatrix} \qquad \vec{X}_0 = \begin{pmatrix} \hat{Q} \\ \hat{P}_1 \\ \hat{P}_2 \\ \vdots \\ \hat{P}_k \end{pmatrix} = \sum_{j=1}^{n} f_j \vec{x}_j$$

and

*(where i = 1, 2, … k and j = 1, 2, …, n )*

Let Y = $\sum$ $y_k$ be the population total of the target variable to be estimated, and the simple estimator (Horvitz-Thompson estimator) for the population total of Y is defined as:

$$\hat{Y} = \sum f_k \, y_k$$

It can be considered as a problem of finding a set of calibrated weights $w_k$'s which are as close to $f_k$'s as possible and in such a way that the estimates based on $w_k$'s for all the auxiliary variables are equal to the known totals $\vec{X}$. Mathematically, the problem can be recast into a minimization problem as follows:

Minimise $\sum q_k f_k G(\frac{w_k}{f_k})$

subject to the constraints of $\vec{X} = \sum w_k \vec{x}_k$

With the choice of $G(x) = (x - 1)^2 / 2$, $\hat{Y} = \sum w_j y_j$ is the Generalised Regression Estimator for Y and the estimator of the weight of quarters *j* is calculated as follows:

$$w_j = f_j \left[ 1 + (\vec{X} - \vec{X}_0)' \left( \sum_{j=1}^{n} \frac{f_j}{q_j} \vec{x}_j \vec{x}_j' \right)^{-1} \frac{\vec{x}_j}{q_j} \right]$$

To achieve better precision level, the estimation process was performed iteratively at different geographical levels one by one. The grossing-up factors calculated by the ratio estimation were used as the initial values in the Generalised Regression Estimation.

## 4. Conclusion

It was observed that the calibrated weights could converge quickly. Theoretically, all short form variables could be used as auxiliary variables in the modified Generalised Regression Estimation. However, according to the results of some scenario testing, in order not to result in too many extreme calibrated weights, the number of long form records has to be sufficiently large. Moreover, the runtime of the estimation process would be very long if many auxiliary variables are needed to be controlled in the estimation process.

The modified Generalised Regression Estimation method could be used to find a set of calibrated weights to obtain the small-area estimates as well as consistent grossed-up estimates of the selected auxiliary variables at the aggregate levels with the estimates available from the independent data sources so as to facilitate consistent interpretation of the survey results.

**References**

Bankier Michael (2002)
"2001 Canadian Census Weighting"
*Proceedings of Statistics Canada Symposium 2002*

Bell Philip (2002)
"Weighting and Standard Error Estimation for ABS Household Surveys"
*Paper prepared for ABS Methodology Advisory Committee*

Statistics Canada (2009)
*2006 Census Technical Report: Sampling and Weighting*

Census and Statistics Department, Hong Kong SAR Government (2012)
*2011 Population Census Main Report: Volume II*

Ghosh M. and Rao J.N.K (2008)
"Small Area Estimation: An Appraisal"
*Statistical Science*

Münnich Ralf, Burgard Jan Pablo, Vogt Martin (2009)
"Small area estimation for population counts in the German Census 2011"
*Section on Survey Research Methods, JSM*

Rao J.N.K. (2012)
"Small Area Estimation: Methods and Applications"
*Paper presented at the Seminar "Applications of Small Area Estimation Techniques in the Social Sciences" of the Iberoamerican University, Mexico City*

Saei Ayoub and Chambers Ray (2003)
"Small Area Estimation: A Review of Methods Based on the Application of Mixed Models"
*Southampton Statistical Sciences Research Institute, University of Southampton*