BC Municipal Population Estimates – An Error Evaluation

(UNPUBLISHED)

BC Stats

Introduction:

BC Stats has been producing annual population estimates for the municipalities and the unorganized areas in British Columbia for many years. Municipal population estimates for the current year are usually published in the month of December. The population estimates are used by many different public and private agencies. Most importantly, the estimated population figures are use by the Ministry of Community Services as a basis for allocating grants from the provincial government to each municipality. The estimates are also used internally by BC Stats as the basis for its annual small area population estimates and projections.

The core of the estimation system is a computerized routine called the Generalized Estimation System (GES) derived from a cross-sectional regression model. In short, the model estimates annual population change for each area by correlating changes in indicators that have been influenced by, and correlated with, population increase or decline. These indicators are termed symptomatic indicators. The annual population change for each area from the base year (a census year) combined with the population in the base year will form the population estimate for each area. The individual estimates will then be adjusted to conform to a total B.C. population obtained each year from Statistics Canada.

The accuracy of the population estimates depends largely on the quality, consistency and relevancy of the symptomatic indicators used in the model. Indicator data are obtained from different administrative files. Over the last fifteen years, different indicators have been added and deleted from the model. Data from the federal programs, namely Family Allowance and Old Age Security were once used as indicators. Family Allowance was deleted due to its transformation to a non-universal program. Old Age Security data has not been used since 2003 when it was replaced by telephone data. Electrical meter records from BC Hydro have always been used as the major indicator since it was proven to be the most relevant indicator so far. The 2006 municipal population estimates on which this paper focuses used electrical meter and telephone line hook up as indicators in the model.

Error Evaluation:

The evaluation of the accuracy and understanding the nature of the error of any population estimation methodology are necessary in order to better utilize the estimates as well as to improve on future estimations. Error measurements are only possible when the actual figures are available to compare with the estimates. With the availability of the 2006 Census figures, evaluation of the 2006 municipal population estimates becomes possible. In this paper, all the errors discussed were calculated from the differences between the estimates and the preliminary 2006 Census population figures. At the time of this study, final 2006 estimates - census plus adjusted net census undercount - were not yet available. However, the preliminary censal estimate figures used in this study were adjusted by including estimates of the 2006 net census undercounts that were based on the final estimates of 2001 net census undercount.

The 2006 estimates provide population count (as of July 1st of 2006) for 185 areas in the province. Of those, 158 were incorporated municipalities with the remaining 27 being unorganized areas. This paper will include all 185 areas for error evaluation. Population estimates for 28 regional districts were derived from aggregating the 185 sub-provincial areas and errors in the regional district level are also discussed. Also, the municipal population estimates were also converted geographically into population estimates for the Local Health Areas (LHAs) used by the Ministry of Health. Error structure of the 2006 LHA population estimates will also be presented.

Evaluation Measures:

The accuracy of the 2006 population estimates was assessed through a number of summary measures:

- average absolute percent error (AAPE)
- median absolute percent error (MAPE)
- standard deviation about the average of the absolute percent errors
- frequency distributions of percent errors
- Index of Misallocation (often referred to as Index of Dissimilarity)

Of the above listed measures, Index of Misallocation (IM) warrants some further explanations:

When population estimates are used as a basis for the allocation of government funding such as grants to municipalities, the extent to which a given total population is misallocated among the component parts will directly affect the extent to which there exists misallocation of the population-based funding. Where misallocation occurs, one or more areas may obtain more than the intended entitlements at the expense of one or more other areas. It is therefore desirable to have some measure of the misallocation that could result. Such a measure is found in the Index of Misallocation (IM).

The Index of Misallocation gives the percentage of all funds allocated on a per capita basis that would have to be reallocated in order to achieve zero misallocation. For example, an IM of 1.0 implies that a shift of one percent of the per capita grants budget is required in order to change the allocation of funds from that obtained using the estimated populations to that obtained using the census figures. Hence, IM measures the inequity associated with the set of population estimates when used for population-based grants.

Algebraically, the Index of Misallocation is expressed as:

$$IM = \left[\frac{\sum_{i=1}^{n} \left| C_i - E_i \right|}{\sum_{i-1}^{n} C_i} \right] \times 100 \times \tfrac{1}{2}$$

Where:

C_i is the census population for area *i*.
E_i is the estimated population for area *i* adjusted such that $\sum C_i = \sum E_i$.

Evaluation Results:

The overall error structure of the 2006 estimation is presented in Table 1 below. At the level of municipalities and unorganized areas, the preliminary average absolute error of the 2006 estimates was 6.9 per cent with a median of 4.6 per cent. It is shown in Table 1 that the estimated population figures were higher than a preliminary estimate of the actual population most of the time since 83.3 per cent of the areas had a positive error. Also, 55 per cent (or 101) of the areas had an AAPE less then 5 per cent whereas only about 11 per cent (or 20) of the areas had an AAPE greater than 15 per cent. The error structure was much better at the regional district level. There are a total of 28 regional districts in the province and the 2006 population estimates had an overall 5.6 per cent error, significantly lower than the 6.9 per cent seen in the estimates for the sub-regional level. Also, a relatively lower number of the regional districts had an average error higher than 10 per cent.

Table 1. Overall Errors of the 2006 Estimates vs. 2006 Census

	Municipalities and Unorganized Areas	Regional Districts
AAPE	6.9%	5.6%
MAPE	4.6%	4.1%
Standard Deviation of AAPE	7.2%	4.6%
Index of Misallocation	1.5%	1.1%
No. of Areas Estimated	185	28
% of Areas with Positive Errors (Over Estimation)	83.3%	92.9%
No. of Areas with Error less than 5%	101	16
No. of Areas with Error Between 5 % & 10%	42	7
No. of Areas with Error Between 10% & 15%	22	3
No. of Area with Error greater than 15%	20	2

Grouping the 185 areas by their population sizes yields another error analysis of the estimation model. Table 2 below shows the errors calculated from grouping the municipalities and unorganized areas into seven groups of ascending population sizes. Similar to that observed in Table 1 where errors for the regional districts and sub-region are compared, Table 2 shows that the estimation system tend to be more accurate for areas with higher population than for areas with smaller population size.

Table 2. Errors by Population Size – 2006 Estimates vs. 2006 Census

Population Size	AAPE	MAPE	No. Of Areas
0-500	17.1%	15.5%	10
501-1,000	10.8%	5.3%	17
1,001-2,500	9.1%	5.5%	31
2,501-5,000	7.5%	6.7%	30
5,001-10,000	5.6%	4.7%	24
10,001-25,000	4.8%	3.8%	39
25,000 or More	3.0%	2.1%	34
Overall	6.9%	4.6%	185

The group with the largest population size (25,000+ people) includes 35 areas that are mostly the larger municipalities ranging from Vancouver, Surrey, and Burnaby to Campbell River (see Table 4 at the end of report). The AAPE in this group is 3.0% which is the lowest among all seven groups and considerably lower than the overall average of 6.9 per cent seen in Table 1. Also, the total population in this group represented the majority (77.1%) of the B.C. population in 2006. On the other hand, the smaller population group with 500 or fewer people had the largest AAPE at 17.1 per cent although they only accounted for 10 of a total of 185 areas. This group includes areas such as Port Clements, Slocan, Lytton and Silverton. This group only accounted for less than one per cent of the total 2006 B.C. population. As can be seen in Table 2, about half of the 185 areas had a population of more than 5,000 persons in 2006 and these areas together had an average error that was less than the overall average provincial error. The three groups of areas having 5000+ population had an AAPE less than six per cent, which was lower than the other four smaller population groups that show an AAPE of more than seven per cent. This indicates that the estimation model used in 2006 was more accurate for estimating population for areas that had a population size of 5000 persons or more.

Since 2003, population estimates for B.C. Local Health Areas (LHAs) were obtained from converting the municipal estimates through geographic translation. The geographic relationships between LHAs and municipalities were established based on the 2001 Census geography. For administrative purposes. there were a total of 98 LHAs in the province in 2006. Table 3 below shows the errors of the 2006 LHA population estimates when compared to the 2006 adjusted Census. Similar to the error structure observed above for the regional districts, the errors for LHAs were lower than that for the municipalities due to the fact that

LHAs are, on average, larger than municipalities and that the model tends to be more accurate for larger areas. Another reason is also that the errors in each of the sub areas tend to cancel out each other when they were aggregated to form the estimate for a larger area. Table 4 at the end of this paper show the errors for individual LHAs.

Table 3. Errors of 2006 LHA Population Estimates

AAPE*	5.2%
MAPE!	4.1%
Standard Deviation of AAPE	4.5%
Index of Misallocation	1.4%
No. of Areas Estimated	89
% of Areas with Positive Errors (Over Estimation)	83.1%
No. of Areas with Error less than 5%	52
No. of Areas with Error Between 5 % & 10%	25
No. of Areas with Error Between 10% & 15%	9
No. of Area with Error greater than 15%	3

Discussion of Results:

Overall the model employed for the 2006 population estimates for municipalities and LHAs yielded satisfactory results with acceptable errors. The average absolute median error for the municipal estimates was 4.6% indicating that half of the 185 areas had an estimation error less than or equal to 4.6%. The model has a higher degree of deficiency in estimating population for areas of small population with 5,000 or fewer people. This was expected as the correlation between the population and the symptomatic indicators used for estimating the average changes in the population tends to have a higher variance for smaller samples of population. On the aggregated geographic level, the model performed better with lower overall errors across all population sizes due partly to the effect that errors of opposite signs cancelling each other during the aggregations.

The accuracy of any population estimates utilizing a regression model lies heavily on the quality of the independent variables (the symptomatic indicators) used in the model, as well as the accuracy of the censal estimate itself. The accuracy and relevancy of the indicators are important determining factors of the success of the regression model. In the 2006 estimation, residential electrical meter and telephone hook-up were used as the indicators to estimate the annual changes in population for each area from the base population obtained from 2001 Census. The changes in these hook-up between each year from 2001 to 2006 were used to approximate population changes from the base year to arrive at the 2006 estimates. The quality of telephone data (obtained from Telus) as a population indicator has deteriorated over the past several years due mainly to the increasing number of households substituting their house telephone lines with cellular phones. This phenomenon was more prominent in large urban areas such as Vancouver, Burnaby, Surrey, and Richmond with relatively higher proportions of younger population.

Data from electrical meter hook-ups, which tends to be a more stable indicator, also has it shortcomings. One of the observations from electrical meter data is that it may not account for population living in the secondary suites that do not have a separate electrical meter. Any living arrangement of this kind that was formed or occupied after the base year would not have been reflected by the annual meter changes. Again, this problem seems to be more prevalent in major urban centers where most secondary suites exist. Another inherent problem with electrical meter data is that it may not be effective in showing population decline. Almost all electrical meter remains in place even when a residence is not occupied, and in many cases most meters would show electricity consumptions even when the residences were not occupied in order to maintain the function of some appliances. Hence, part of the process of the population estimates during the data preparation phase is to filter out electrical meter records that were believed to be unoccupied based on available information. This process is itself an estimation and could introduce error in the final estimates. These drawbacks in the indicators provide some explanations for the tendency of under estimation by the model for areas with larger population.

Estimation errors also tend to be greater when the population estimation was made for a year that is farther away from the base year. The 2006 municipal population estimate was the last one produced before the new 2006 Census figures became available. Hence, it is a good choice for error evaluation as it should potentially show greater errors when compared to those produced in other years during the inter-censual period of 2001-2006.

Another challenge in using any administrative data as indicators for population estimation is the accurate geo-coding of the administrative records. Electrical and telephone record data must first be coded to municipalities and unorganized areas before they could be used by the model. In the case of telephone data, the postal code in each record is used as a reference for converting into its corresponding municipality or unorganized area through a translation table. Problems arise from the fact that a small percentage of the records had missing postal codes and that the translation cannot be perfect due to that fact that some postal codes cross municipal boundaries, a problem that is more significant for small municipalities. The geo-coding of electrical records is somewhat simpler and more accurate because each records is associated with a geographic code, namely Capital District Code, and Capital District are intended to align perfectly with each municipal boundary.

Telephone data was dropped from the model used in the recently released 2007 municipal population estimates due the problems discussed above. Instead, data from the health registry of B.C. was used as the indicator along with electrical records. The records from the health registry theoretically include everybody in the province and therefore should be more representative of the province's population. The results from the 2007 model that used electrical and health data appeared to be satisfactory and consistent. The approach to using the health records will be further refined in future releases. At the same time, BC Stats is also investigating other administrative databases for their appropriateness for use in the estimation model in order to achieve higher degree of accuracy in its population estimates.