Survey Methodology, December 1985
Vol. 11, No. 2, pp. 219–222
Statistics Canada

Experience with Small Area Population Estimates[1]

ROSEMARY K. BENDER[2]

ABSTRACT

Statistics Canada's current methodologies forestimating the population of census divisions and census metropolitan areas are the regression-nested and component methods. This paper presents the experience with these estimates for the period 1981 to 1985, focusing on problems encountered with the input data on family allowance recipients.

KEY WORDS: Regression-nested estimates; Component estimates; Family allowance recipients; Postal code files.

1. INTRODUCTION

Statistics Canada's current methodologies for estimating the population of census divisions (CDs) and census metropolitan areas (CMAs) are the regression-nested and component methods. The regression estimates for 1982, 1983 and 1985 were published in Catalogue No. 91-211 on schedule. Those for 1984 were only made available in March of 1985. There was a delay in obtaining the input data on family allowance. Furthermore, as explained below, we encountered problems with the quality of these data. In particular, the resulting population estimates for CMAs were not acceptable and an alternate methodology had to be used.

Component estimates of the population for CDs and CMAs have been published in Catalogue No. 91-212 on schedule for 1982 and 1983. We should release the 1984 estimates by April 1986. An evaluation of the component estimates produced thus far has shown the data to be of good quality.

2. ADJUSTMENTS

Since introducing the regression estimates for CDs and CMAs in 1982, some adjustments to the data and the methodology have been necessary. They are summarized below:

- For the 1983 estimates for the CD Chicoutimi and the CMA Chicoutimi- Jonquière in the province of Quebec, the family allowance data was adjusted based on the growth pattern of the previous year. The problem was traced to postal codes used to obtain the family allowance data.
- In 1984, 17 census divisions estimates were imputed with preliminary component estimates.
- In 1984, we decided to publish for the CMA of Calgary, estimates based on the annual census conducted by the city. This will be done for the entire 1981-1986 period.
- In 1984, we developed a new methodology for all CMAs other than Calgary, which aggregates census division regression estimates. This will be used for the entire 1981-1986 period.

The following sections explain the problems encountered in more detail.

---

[1] Abridged version of the paper presented at the meetings of the Federal-Provincial Committee on Demography held on November 28-29, 1985, Ottawa, Canada.

[2] Rosemary K. Bender, Demography Division, Census and Demographic Statistics Branch, Statistics Canada, 4th floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

3. PROBLEMS WITH INPUT DATA FOR REGRESSION ESTIMATES

There was a delay in producing 1984 estimates due to problems encountered in obtaining data on family allowance recipients from Health and Welfare Canada, and the appropriate postal code translation files necessary to process these data.

i) *Family Allowance Data*

The numbers of Family Allowance recipients as of June 1, is generally available by mid September of each year. The 1984 data from Health and Welfare Canada however, were delayed as a result of decentralization of the regional operations of the program in Ontario. Problems were also encoutered in the files of all provinces with respect to information on effective dates of transfer and reason codes for inter-area transfers. The 1984 data were released to Statistics Canada in an unedited form in November. Corrective actions were taken by Health and Welfare Canada, and Family Allowance data as of June 1, 1985 was on schedule.

ii) *Postal Code Files*

The data on family allowance recipients from Health and Welfare Canada is coded by postal code. Therefore, to identify the children receiving family allowance in each CD and CMA, a file must be created that groups the postal codes by CD and CMA. This is done using a master file that contains all the postal codes in Canada, with detailed geographic codes that are used to assign the postal codes to any level of geographic disaggregation.

Problems have arisen that were unexpected and in some cases had serious consequences. For our estimates, it is important that the postal code files used each year by Health and Welfare Canada be consistent with the one that was used to develop the regression model. The only change in the file should be the addition of new postal codes. Any shifting of postal codes from one region to another can result in changes to the population that do not actually occur.

The problems we encountered stem from the fact that since we developed our regression model, different divisions and departments have produced the postal code files. In 1982 and 1983, it was done by the Administrative Data Development Division of Statistics Canada. In 1984, the Standards Division of Statistics Canada took over the responsability and in 1985 it was done by Health and Welfare Canada. Each had its own approach resulting in family allowance data that was not consistent from year to year. Two different types of problems arose. We have resolved the first. However, the second will persist throughout the 1981-1986 postcensal period.

The first source of difficulty was the shifting of postal codes from one area to another. The master file is created by the Standards Division of Statistics Canada. However, in some cases, the CD or CMA geographic code is blank or wrong. For CDs this occurs mostly with rural codes, where postal codes often refer to post offices covering large territories across CD boundaries. The inclusion of the CMA geographic codes is fairly recent, and the quality improves each year. Thus, our initial assumption that the postal code file would be consistent from year to year was not quite true. There are changes made each year.

Our files were initially created by the Administrative Data Development Division (ADDD) of Statistics Canada. They made changes in their copy of the master file before proceeding to group the data. In 1984 the Standards Division took over producing our file. When we became aware of the consequences this would have, we developed with ADDD a way to match the original master file with the latest master file from Standards Division, adding only the new postal codes. Any changes to the CD or CMA codes were ignored. We realise that by doing this we do not have the most accurate postal code file available. However, for our

purposes, we are interested in the changes to the proportions of children receiving family allowance. The effect of using some erroneous, but consistent postal codes is that we include or exclude some children from another area in the calculation of proportions. The proportions would not be significantly different from those using correct postal codes, but would change if these children were suddenly excluded or included.

This process of adding only new codes to our postal code file improved significantly the quality of the 1984 family allowance data for census divisions. Only 17 of the 231 regression estimates of CDs (excluding those of British Columbia, as they produce their own regression estimates) needed to be imputed. Because of the delay in obtaining the data, we were able to use preliminary estimates from the component method. For census metropolitan areas, there were still inconsistencies, which we beleive are due to a different type of problem.

When the postal codes are grouped by CDs and CMAs, they are also converted into ranges of postal codes. For example, if the postal codes A1A1A1, A1A1A2, A1A1A3 and A1A1A4 all have the same CMA code, then they will be combined into the range A1A1A1-A1A1A4. However, in processing the over 600,000 postal codes, certain assumptions are made, depending on the software. If, in the above example A1A1A2 was not there, the program may still create the same range, assuming that if A1A1A2 did exist, it would have the same CMA code as the others in the range. This type of assumption could alter the family allowance data processed for each region. Furthermore, if different softwares are used each year, serious inconsistencies can arise.

We believe this is the major cause for the poor quality in the family allowance data for CMAs. The softwares used by the ADDD and Standards Divisions were different. What complicated matters even more was that as of 1985, the entire operation is now done by Health and Welfare Canada, again using a different software. We therefore had to disregard the data and develop an alternative methodology for CMAs.

4. METHODOLOGICAL CHANGE FOR CMAs

The CMA estimates previously released for 1982 and 1983 were based on the same regression-nested procedures as for census divisions. In the evaluation of the 1984 estimates, however, estimates for many census metropolitan areas were found to be inconsistent with alternate sources and past growth trends. As described above, the problems seem more related to the quality of the input files rather than to methodology.

Taking into account these inconsistencies as well as comments from the provincial focal points, it was decided to use an alternate methodology. This new methodology was previously developed for estimating various CMA components of population change. It consists of aggregating census divisions regression estimates, using the ratio of the population of the CMA to that of overlapping CDs, as observed the previous year by the component method. In comparing estimates for 1981, obtained through this methodology, with the 1981 Census counts for census metropolitan areas, an average absolute error of 1.3% as observed, as compared to 2.3% for the previous methodology.

To maintain consistency in methodology for the entire 1981-1986 period, the alternate method has been used to derive the CMA estimates for 1982 to 1985, and will be used for 1986. That is, estimates of population for CMA's other than Calgary are obtained by aggregating the census division regression-nested estimates, and those for Calgary as described below, are based on the annual census conducted by the city.

In 1984, it was found that the regression-nested estimates for Calgary CMA for 1982 and 1983 were too high in comparison with the census counts conducted annually by the city of Calgary. The component estimates also supported the idea of adjusting the regression-nested estimates for Calgary. It was decided to publish estimates based on the city of Calgary

census count extrapolating the April data to June 1. This is in line with Statistics Canada policy where, when there is a complete enumeration, this should be considered over an estimate prepared by an indirect procedure, unless there is evidence that the enumerated count is suspect.

5. COMPARISON WITH OTHER DATA SOURCES

The regression and component estimates are compared with alternative data sources whenever possible. We receive from the Saskatchewan and Alberta governments the number of people registered in their respective health care programs. These data are used in the regression model. However, they are also evaluated for consistency with the family allowance data and past growth trends. In most cases they were consistent, and differences were traced to the problems encountered with family allowance data.

The Quebec Bureau of Statistics produces annual population estimates of their administrative regions which are subdivisions of the Quebec CDs. Their data are comparable to ours except for the CD of Nouveau Québec. This census division, located in northern Quebec, is largely comprised of unorganized territories, and it is difficult to estimate the population. The BSQ generally adopts our estimates, though for 1984 it imputed its own estimate for Nouveau Québec.

We also appreciate feedback from users who may have access to specific local area data.

6. CONCLUSION

The methods used to produce population estimates for census divisions and census metropolitan areas have in general functionned very well. However, in the case of the regression estimates, problems with input data made it necessary to impute estimates for certain CDs with alternate data, and to revise the methodology for CMAs.

The problems encountered were mostly related to the family allowance data and the postal code files that are necessary to process these data. Most of the problems have been resolved. However, as Health and Welfare are now taking over the responsibility of creating the postal code files, the 1986 data may still have problems of consistency and will have to be carefully evaluated.

Despite these problems, the regression methodology with certain adaptations will be used to produce estimates for 1986. If, however, we decide to continue with the methodology for the 1986-1991 period, we must first ensure that consistent postal code files be processed by the same department throughout the period.